UNIONE EUROPEA
Fondo Sociale Europeo

Ministero dell'Università e della Ricerca

PON RICERCA E INNOVAZIONE 2014 - 2020

REACT EU

Università Ca'Foscari Venezia

Corso di Dottorato di ricerca

in Lingue, Culture e Società Moderne e Scienze del Linguaggio

ciclo XXXVII

Tesi di Ricerca

Multi-modality For All

Tecniche di sottotitolazione e trascrizione automatica in approccio human-centered

SSD: L-LIN/11

Coordinatrice del Dottorato
ch. prof. Anna Cardinaletti

Supervisore
ch. prof. Giulia Bencini

Dottoranda
Martina Pucci
Matricola 989044

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020, risorse FSE REACT-EU
Azione IV.4 "Dottorati e contratti di ricerca su tematiche dell'innovazione"
e Azione IV.5 "Dottorati su tematiche Green"

SUMMARY

Multi-modality For All

Tecniche di sottotitolazione e trascrizione automatica in approccio *human-centered*

This doctoral work is composed of three studies. The main aims were 1) to investigate the habits of use of audiovisual translation products (captions, intralingual and interlingual subtitles) in L2 speakers of English to support speech processing and content comprehension, 2) to investigate how captions generated by automatic speech recognition (ASR) systems affect speech processing and comprehension of lectures delivered in English to university students (L2 speakers of English), 3) to evaluate the quality and reliability of automatic transcriptions in real-life settings, and 4) create and evaluate a series of graphical features to be implemented in the display formats of automatic captions to signal users the level of confidence the ASR system has in its transcription. We asked 42 L2 speakers of English to state their habits of using audiovisual translation products and their opinions on using live captions in class through a questionnaire. The findings indicate that these students regularly use audiovisual translation products when watching audiovisual content in English to support speech processing (specifically, speech perception and segmentation) and content comprehension. They would also likely use this technology if the transcription accuracy were sufficiently high; otherwise, they may find it confusing and distracting. We then assessed the performance of an ASR system by qualitatively analyzing a set of transcriptions from various topics (politics, education) and different spokespersons (English, Italian), each affected by particular acoustic-, and environment-, and speaker-related features. The corpus analysis revealed an overall good performance of the ASR system. Still, it showed a trend where some of the factors considered

influenced the number of errors in the transcription and the confidence level of the system. In the same analysis, we evaluated the reliability of confidence scores as a metric to create a series of color-coded markups, aiming to signal to users the confidence levels of the ASR system in its transcription and improve reliability. This analysis revealed an overall positive medium correlation between the type of transcription (correct, incorrect) and the confidence scores. Based on this analysis, we created two sets of graphical features that were employed in a between-subject experiment to assess if 1) the different display formats affected students' attention, 2) transcription errors affected students' content comprehension, and 3) the reliability of the automatic captions improved with the use of different display formats. Sixteen participants were asked to watch a seminar-style video lecture delivered in English with captions in one out of four display formats. Results showed that errors in the automatic captions did not impede comprehension, but they rendered automatic captions unreliable for students who used the text to aid speech perception and content comprehension, at the same time reducing listening effort. One display format (V2) was attested to be the most informative to students. Moreover, a statistical comparison between the four experimental display formats revealed a trend where the V2 display format helped participants the most with content comprehension, at the same time increasing the reliability of the text. Participants also expressed interest in testing the different markups in class. In conclusion, these findings deepen our understanding as of why and when L2 speakers use captions to assist speech processing and content comprehension in diverse contexts, provide insights into if and how this technology can aid information accessibility, and what engineers and (educational) institutions can do to make sure that this tool offers adequate access to information for all.

Preface

Martina Pucci was supported by a doctoral student Research and Innovation scholarship (*PON Ricerca & Innovazione 2014-2020* - CUP: H75F21002110005) funded by the Italian Ministry of Universities and Research (MUR) and the European Union (FSE REACT EU).

The research reported in this thesis was conducted in partnership with the company *Cedat85*.

Portions of the text included in Chapters 1, 2 and 6 of this dissertation have been published in Publication n° 1 (2023, see list below).

Disclaimer

In compliance with Ca' Foscari University of Venice guidelines on the responsible use of AI for research[1], I state that I utilized the online tool *Grammarly*[2] to identify spelling and grammatical errors and rephrase some unclear sentences. Adjustments to the text were made under strict personal review to ensure that the modified text aligned with the original version I provided for rephrasing.

[1] *Intelligenza artificiale (AI)*. (2025). Università Ca' Foscari Venezia. Retrieved March 12, 2025, from https://www.unive.it/pag/49804.
[2] *Grammarly: Free AI Writing Assistance*. Retrieved March 12, 2025, from https://www.grammarly.com/.

List of publications by year

2023

1. Pucci, M. (2023). Towards Universally Designed Communication: Opportunities and Challenges in the Use of Automatic Speech Recognition Systems to Support Access, Understanding and Use of Information in Communicative Settings. In *Design for Inclusion* (pp. 18–25). IOS Press. https://doi.org/10.3233/SHTI230394

2022

1. Venturini, S., Vann, M. M., Pucci, M., & Bencini, G. M. (2022). Towards a More Inclusive Learning Environment: The Importance of Providing Captions That Are Suited to Learners' Language Proficiency in the UDL Classroom. In *Transforming our World through Universal Design for Human Development* (pp. 533–540). IOS Press. https://doi.org/10.3233/SHTI220884

List of presentations by year

2025

- [Poster presentation] Pucci, M. & Bencini, G. M. L. (2025). Assessing the usefulness of ASR-generated captions in an educational context with L2 speakers of English, *Psycholinguistics in Flanders 2025* (PiF 2025), University of Lille, 19-20 May 2025, Lille (France).
- [Oral presentation] Pucci, M. & Bencini, G. M. L. (2025). Using ASR-generated captions in higher education: the importance of providing accurate and reliable transcriptions to L2 speakers of English to support speech processing and access to information, *16th International Conference on Live Subtitling and Speech-to-Text Interpreting* (ICoLS 2025), University of Leeds, 4 July 2025, Leeds (United Kingdom).

2022

- [Poster presentation] Venturini, S., Vann, M. M., Pucci, M., Bencini, G. M. L. (2022). Do captions benefit everyone? Comparing the effects of automatic versus human captions on learner comprehension*, IMPRS Conference in Interdisciplinary Approaches in the Language Sciences 2022*, Max Planck Institute for Psycholinguistics Nijmegen, 01-03 June 2022, Nijmegen (The Netherlands).
- [Poster presentation] Venturini, S., Vann, M. M., Pucci, M., Bencini, G. M. L. (2022). Do captions benefit everyone? Comparing the effects of automatic versus human captions on learner comprehension. *Linguistics and speech technology: from theory to tools to theory* Summer School (*LingTech 2022*), 29 August – 02 September 2022, Università del Salento, Lecce (Italy).
- [Oral presentation] Venturini, S., Vann, M. M., Pucci, M., Bencini, G. M. L. (2022). Towards a more inclusive learning environment: the importance of providing captions that are suited to learners' language proficiency in the UDL classroom. *Sixth International Conference on Universal Design: Transforming our World through Universal Design for Human Development*, 07-09 September 2022, Brescia (Italy).
- [Oral presentation] Pucci, M. (2022). Towards Universally Designed Communication: Opportunities and challenges in the use of technology to support access, use, and understanding of audiovisual information. *Sixth International Conference on Universal Design: Transforming our World through Universal Design for Human Development*, 07-09 September 2022, Brescia (Italy).

List of Figures

List of Tables

Table of contents

Introduction

The use of audiovisual translation products (captions, interlingual and intralingual subtitles) to aid content comprehension and improve access to information has a long-standing history (Kuo, 2004; Juang & Rabiner, 2005). However, in recent years, this practice has seen significant growth, mainly due to advancements in technology, including the rise of streaming services, portable devices, and the overall expansion of the Internet. Two surveys conducted in 2023[3] and 2024[4] by the web platform *Preply* have highlighted that 51% of the Americans interviewed "watch [audiovisual] content with subtitles most of the time." The younger generations - viewers from the Gen Z[5] (68%) and Millennials[6] (56%) demographic cohorts - are the ones who use subtitles the most while watching videos on streaming platforms. The surveys also summarize the reasons behind the popularity of subtitles, attributed to several factors. According to American viewers, subtitles not only help with comprehension of the content (26%), but also assist speech processing (especially perception and segmentation - e.g., when native speakers encounter unfamiliar accents or dialects, or when the dialogues are muffled by background music or noise on the screen, respectively 25% and 15% of the answers).

Over the past 25 years, daily access to the English language has significantly increased worldwide, offering valuable opportunities also for individuals to improve their language skills in real-world contexts through exposure to authentic linguistic input (Montero Perez, 2022; Jia & Hew, 2023). In particular, video streaming platforms provide various forms of audiovisual translation products to improve information accessibility for all viewers (3PlayMedia, 2024). These tools also assist

[3] Preply (2023). Why America is obsessed with subtitles. https://preply.com/en/blog/americas-subtitles-use/.
[4] Preply (2024). Why America is (still) obsessed with subtitles. https://preply.com/en/blog/america-still-obsessed-subtitles/.
[5] Individuals born between 1997 and 2012. Source: Wikipedia. Generation Z. https://en.wikipedia.org/wiki/Generation_Z.
[6] Individuals born between 1981 and 1996. Source: Wikipedia. Millennials. https://en.wikipedia.org/wiki/Millennials.

individuals while learning new languages (Dizon, 2016; Dizon & Thanyawatpokin, 2021). The benefits of using audiovisual translation products to improve language knowledge are well-established by the literature on the topic (see Montero Perez *et al*., 2013 for a meta-analysis; Gernsbacher, 2015; Montero Perez, 2022 for reviews). The positive effects related to the simultaneous presence of two types of input (audio + text, also called *bimodal input*) are currently related to the *Dual Coding Theory* (Paivio, 1991), which suggests that when the "information is repeated by means of different channels, readers and learners retain better information" (Chan *et al*., 2019: p. 243).

English has also become a widespread language of instruction due to the internationalization of academic institutions and a multilingual and multicultural student population (van Gauwbergen *et al*., 2024). In this context (like in audiovisual products), both intrinsic speech variability in instructors (e.g., accent, speech rate), individual differences in students (e.g., language proficiency), as well as the acoustic characteristics of lecture halls, may challenge L2 speakers when processing speech, increasing the load on cognitive processes (Goh, 2000; Mattys & Wiget, 2011; Mattys *et al*., 2012 for a review; van den Heuij *et al*., 2018). These factors may negatively affect content comprehension and learning in university courses taught in English as an L2/*lingua franca* (Wald & Bain, 2008). Offering *live subtitles* during lectures may aid students' ability to process speech and improve access to information (van Gauwbergen *et al*., 2024; Robert *et al*., 2021). Specifically, live subtitles could assist L2 speakers in aiding speech perception and segmentation (Bird & Williams, 2002; Mitterer & McQueen, 2009; Charles & Trenkic, 2015) and improving content comprehension (Chan *et al*., 2019). However, while live subtitles have the potential to enhance speech processing, adding this type of audiovisual translation products to an already rich environment (a lecture) may increase students' cognitive load (Sweller, 2004), potentially hindering learning.

Previous research has explained this potentially detrimental effect through the means of the *redundancy effect* (Sweller, 2005; Kalyuga & Sweller, 2014), a theoretical concept developed within the frameworks of *multimedia learning* outlined by Richard E. Mayer in 2002 (based on the limited capacity assumptions of cognition, and - specifically - working memory) and *cognitive load theory* proposed by Sweller (1988) (defined as "the load imposed on the learner's cognitive system while performing a particular task" – Chan *et al*., 2019: p. 241). In short, the presence of the textual input both in the slides and in the subtitles could impose an additional load on students' cognitive resources, making it difficult to allocate their attention and process information effectively. In summary, processing multiple input types - such as the professor's speech, the text on the slides, and, optionally, live subtitles, all presented in a second language (L2) - could be challenging for students: for this reason, educators should consider eliminating redundant information that could potentially overload

students' cognitive system and interfere with their learning process. In this context, educators should also keep in mind the natural variability between speakers, not only in terms of language proficiency, but also of individual cognitive characteristics and personal learning strategies (Kruger, 2013; To, 2024). Therefore, while using live subtitles in an educational setting may benefit some students, it could be detrimental to others. However, research findings on the benefits of (live) subtitles in classrooms are still scarce and mixed (e.g., Kruger, 2013; van Gauwbergen *et al*., 2024), indicating that further investigation is needed.

In recent years, technological advancements have also led to the increased use of live captions generated by *automatic speech recognition* (ASR) systems. However, transcriptions generated by ASR systems are often far from perfect, requiring human professionals to manually correct the text for reliable access to information (Romero-Fresco & Fresno, 2023). Transcription accuracy is critical in supporting content comprehension (Chan *et al*., 2019) and communication in general (Shimogori *et al*., 2010) because errors may alter the content of the original spoken message and, therefore, halt communication (Cao *et al*., 2018; Orellana *et al*., 2024). Therefore, providing automatic live captions with high accuracy is essential to guarantee access to knowledge and improve learning (Ryba *et al*., 2006; Butler *et al*., 2019), and previous research suggests that users find ASR captions and transcripts useful when they are highly accurate (Wald & Bain, 2008; Shimogori *et al*., 2010; Butler *et al*., 2019). On the one hand, the effectiveness of an ASR system depends on its architecture, mainly the quality and amount of training data contained within its acoustic model and lexicon (Kuhn *et al*., 2024; O'Shaughnessy, 2024). On the other hand, previous research has shown that providing the ASR system with high-quality audio recordings is equally important for achieving accurate, reliable transcriptions. The acoustic characteristics of the environment and the quality of the devices with which speech is collected significantly affect the outcome of transcriptions and the degree of confidence an ASR system has of its transcription (Berke, 2017; Alharbi *et al*., 2021). In the same way, factors related to speech variability between speakers (e.g., physical traits that affect voice structure, regional or foreign accents, speech rate) may further affect the number of words correctly/erroneously transcribed (Benzeghiba *et al*., 2007) and the degree of confidence the system has in transcribing the words from the original speech stream (Berke, 2017). For this reason, it may be helpful to show users how confident the system is with its transcription by manipulating how words are displayed on the screen (for instance, by showing words in different colors). According to Wald and Bain (2008), graphical indications of confidence levels would help users detect errors in live captions. However, they also stated that this concept should be investigated further before being implemented in ASR systems (p. 443). Indeed, some studies have explored the possibility of implementing specific markups or display formats to show errors or display the confidence of the

system to users both in automatically generated transcriptions (e.g., Vertanen & Kristensson, 2008) and live captions (Piquard-Kipffer *et al*., 2015; Shiver & Wolfe, 2015; Berke, 2017; Berke *et al*., 2017). However, these investigations have not specifically focused on L2 speakers of English.

To the best of our knowledge, this is the first project that aims to investigate the habits of use of audiovisual translation products (captions, intralingual, and interlingual subtitles) and the role of automatic captions in aiding speech processing and content comprehension in L2 speakers of English in diverse settings (at home, at university), at the same time seeking to create a set of graphical features (i.e., color-coded markups) to be implemented in the text of automatic captions based on the reliability of an ASR metric called *confidence score*. The usefulness and impact both on cognitive processes (speech processing, attention) and content comprehension of these markups are then tested. In this project, we also aim to assess the robustness of a speaker-independent ASR system in real-world applications by analyzing a corpus of automatically-generated transcriptions.

The studies developed in the project used a user-centered approach (Vredenburg *et al*., 2002). The research project was divided into three phases. In each phase, we carried out the following research activities:

1. *Phase 1*

- Investigation of the habits of L2 English speakers in watching audiovisual content in English and utilizing audiovisual translation products (captions, intralingual, and interlingual subtitles) to support speech processing and content comprehension in everyday life.
- Questionnaire aimed to investigate the potential use of automatic captions in class during lectures delivered in English.

2. *Phase 2*

- Analysis of the performance of a traditional ASR system in real-world applications.
- Investigation on the potential correlation between confidence scores and transcription type (correct, incorrect transcriptions).
- Definition of the characteristics of the graphical features (color-coded markups) to display the confidence scores of the ASR system in the text of automatic captions.

3. *Phase 3*

- Study on the effects of automatic captions and different color-coded markups on the comprehension of seminar-style lectures in L2 speakers of English.
- Questionnaire aimed to collect insights and opinions of L2 speakers of English on the color-coded markups shown in the automatic captions.

Chapters 1 and 2 summarize the relevant literature for the studies conducted in the three phases. In Chapter 1, we focus on the use of technology to aid access to information for diverse populations. In particular, we give an overview of automatic speech recognition (ASR) systems, their basic traditional architecture, and the factors affecting their performance. In Chapter 2, we describe the psycholinguistic processes involved in speech processing and content comprehension, focusing on L2 speakers of English. We also summarize the relevant literature on the benefits of providing subtitles to aid content comprehension and learning in the target population. Lastly, we discuss the existing literature on using live subtitles and automatic captions in educational settings. Chapters 3, 4, and 5 present the findings from the three studies conducted for this doctoral project. Finally, Chapter 6 provides a general discussion on the use of ASR systems and automatic captions in real-world settings (particularly in academic contexts) and broader considerations about aiding speech processing and content comprehension in L2 speakers of English with the help of this technology based on the results of the three studies conducted.

Chapters 1, 2, and 6 include the text of a paper I published in 2023 titled "Towards Universally Designed Communication: Opportunities and Challenges in the Use of Automatic Speech Recognition Systems to Support Access, Understanding, and Use of Information in Communicative Settings"[7] (see the complete reference in the footnote below and in the list of references at the end of this document).

[7] Pucci, M. (2023). Towards Universally Designed Communication: Opportunities and Challenges in the Use of Automatic Speech Recognition Systems to Support Access, Understanding and Use of Information in Communicative Settings. In *Design for Inclusion* (pp. 18–25). IOS Press. https://doi.org/10.3233/SHTI230394

1 Multimodality For All[8]

The use of automatic speech recognition systems to support access, understanding and use of information in communicative settings

1.1 Introduction

Access to information in critical domains for citizenship and well-being such as, for example, health, emergency information, individual rights and the law by any means of communication (e.g., television, the internet) are fundamental rights for all citizens. However, there are instances in which individual circumstances or characteristics hamper access to information or communication. Unlike physical barriers, communication barriers are difficult to overcome due to the heterogeneity of populations: factors such as literacy, proficiency, disability, can interfere with access to information if this is presented in ways that are unavailable or inaccessible to the end user (Wald & Bain, 2008). The scientific community has contributed to the matter by proposing strategies to overcome communication barriers depending on different contexts and speakers. For example, some guidelines[9,10,11] suggest the employment of *bi/multimodality*, that is, the presentation of information in more than one modality (e.g., spoken + written modalities), also with the help of technological tools[12]. One of the devices that can help with the bi/multimodal presentation of information is

[8] **Disclaimer** - This chapter includes excerpts from the following publication: Pucci, M. (2023). Towards Universally Designed Communication: Opportunities and Challenges in the Use of Automatic Speech Recognition Systems to Support Access, Understanding and Use of Information in Communicative Settings. In *Design for Inclusion* (pp. 18–25). IOS Press. https://doi.org/10.3233/SHTI230394

[9] UDL: The UDL Guidelines [Internet]. 2018 [cited 2022 Dec]. Available from: https://udlguidelines.cast.org/.

[10] Definition and overview | Centre for Excellence in Universal Design. 2020 [cited 2022 Dec]. universaldesign.ie. Available from: https://universaldesign.ie/what-is-universal-design/definition-andoverview/.

[11] The 7 Principles | Centre for Excellence in Universal Design [Internet]. 2020 [cited 2022 Dec]. Available from: https://universaldesign.ie/what-is-universal-design/the-7-principles/.

[12] UDL: Offer alternatives for auditory information [Internet]. [cited 2022 Dec]. Available from:

automatic speech recognition (ASR), a system that turns speech signals into written transcriptions (Karpagavalli & Chandra, 2016; Jurafsky & Martin, 2025).

In this chapter, I will focus on the use of ASR systems in facilitating access to information for diverse populations. I will provide a brief overview of traditional ASR systems and the current limitations of the technology, along with the factors affecting the robustness of these systems. Finally, I will discuss the potential exploitation of confidence scores to improve reliability of the output of ASR systems.

1.2 The use of technology to aid access to information

Experts globally have created guidelines for designing products, environments, and communication systems (e.g., *Universal Design* – UD and *Universal Design for Learning* - UDL principles[13]). The ultimate goal is to ensure that users benefit from final designs without having to be further adapted to their needs. However, while this approach can contribute to eliminating architectural barriers in public or private buildings, there is no easy solution to remove *communication barriers*. Different native languages between two or more speakers or various proficiency levels in a foreign language can hinder access and dissemination of information in communicative settings. In the same way, the modality with which information is conveyed is another frequent obstacle. One modality may be available to a large segment of the population, but not to others: for example, audio input for deaf and hard-of-hearing people or written input for blind people. If information is delivered in only one modality (e.g., spoken or written only), it could preclude access to crucial information for these individuals, potentially excluding them from communication. Luckily, technological advancements in the last decades have contributed to removing some of these barriers. Today, these tools can assist people in presenting information in more than one modality simultaneously (*bi/multimodality*), such as combined spoken and written input. For these reasons, UD principles and UDL guidelines encourage designers and instructors to combine the use of various technological tools to deliver information multimodally, supporting users and communication.

Researchers have been studying the cognitive mechanisms underlying the processing of bi/multimodal input for years (see Adesope & Nesbit, 2012 for a meta-analysis). Specifically, research has highlighted how the presentation of information in multiple modalities (e.g., audio + written

https://udlguidelines.cast.org/representation/perception/alternatives-auditory.
[13] See *footnotes* 7-9, page 6.

input) benefits language comprehension, vocabulary learning, and memory for content (Gernsbacher, 2015; Montero Perez, 2020). For example, the simultaneous presentation of auditory input and written transcription helps diverse students recover missing or incomplete information (Butler *et al*., 2019). Another example concerns low-proficient speakers of a foreign language. In this case, written input can help these speakers segment their interlocutor's speech stream while following a lecture (Venturini *et al*., 2022).

With this in mind, governments and supranational organizations have developed policies and projects aimed at improving the use of technology and enhancing the inclusion of diverse users in various settings, especially in the educational one[14,15,16]. The higher education sector has begun to equip its buildings with more advanced technological tools and adopt UD and UDL guidelines to promote inclusion, but this process is still ongoing and requires time, since infrastructures also need to be upgraded to ensure access to education to all (Bencini *et al*., 2018; Van Den Heuij *et al*., 2018; Bencini *et al*., 2021). Specifically, UDL guidelines recommend leveraging technology to enhance individual autonomy and encourage the use of alternative learning strategies. These guidelines also prompt instructors to explore different methods of presenting the content of their lectures to provide easier access to information and improve communication, combining multimodality and technology use.

In the last few decades, developers have focused on building a system that institutions are starting to employ to present information multimodally and improve communication, that is, *automatic speech recognition* (ASR).

1.3 A brief introduction to Automatic Speech Recognition (ASR) and its basic architecture

In the automatic speech recognition (ASR) task, speech signals (in the form of waveforms) are mapped to the appropriate sequence of words by means of an algorithm and converted to written text (Karpagavalli & Chandra, 2016; Jurafsky & Martin, 2025). This technology has a long-standing history behind its development. However, the most important breakthroughs happened in the 1970s-

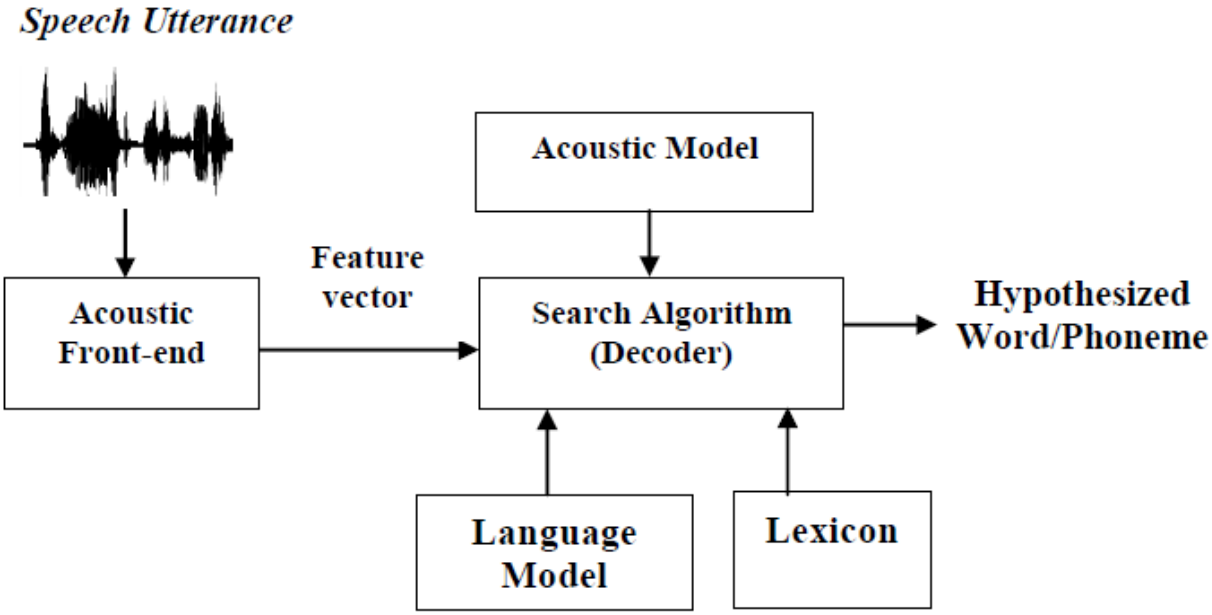
[14] ICT for Inclusion [Internet]. European Agency for Special Needs and Inclusive Education. [cited 2022 Dec]. Available from: https://www.european-agency.org/activities/ict4i.
[15] Digital Education Action Plan (2021-2027) | European Education Area [Internet]. [cited 2023 Jan]. Available from: https://education.ec.europa.eu/node/1518.
[16] Piano Nazionale Scuola Digitale – Scuoladigitale [Internet]. [cited 2022 Dec]. Available from: https://scuoladigitale.istruzione.it/pnsd/.

1990s (with the development of the classic architecture of ASR systems) and, more recently, with the advent of *artificial neural networks* (ANNs) and *deep neural networks* (DNNs) starting from the beginning of the twenty-first century (Juang & Rabiner, 2005; O'Shaughnessy, 2024). ANNs and DNNs are currently employed in many technological applications, such as voice assistants, chatbots, search engines, and other Artificial Intelligence (AI) applications (Dhanjal & Singh, 2024; Feng *et al*., 2024). Most importantly, ASR systems are utilized to generate *automatic captions*, both during live events and offline on streaming platforms (e.g., Chan *et al*., 2019; Romero-Fresco & Fresno, 2023). Since the partner company utilizes an ASR system with traditional architecture, we will focus on describing it in the following paragraph.

The basic architecture of a traditional ASR system reveals that it replicates in a simplified way some of the processes involved in human speech processing (especially word recognition) (Scharenborg, 2007; Karpagavalli & Chandra, 2016; O'Shaughnessy, 2024). ASR systems are composed of five components (Juang & Rabiner, 2005; Karpagavalli & Chandra, 2016; Dua *et al*., 2023; Jurafsky & Martin, 2025) (*Figure 1*):

1. The *acoustic front-end* is devoted to speech signal analysis and feature (or parameter) extraction;
2. The *acoustic model* is a list of statistical representations of the sounds (phones) of words;
3. The *language model* is a module devoted to word identification. It clusters phones into words, helping the acoustic model disambiguate the phones in a chain also on the basis of *grammar* (i.e., statistical syntactical rules). It also contains the *n-gram model*, which groups words based on the statistical probability of appearing together in an ordered sequence;
4. The *lexicon* is a list of words (with their phonological description) that interacts with the acoustic and language models. As part of the building process of the system, developers create and define not only this list of words but also the data contained in the acoustic and language models;
5. The *decoder* is an algorithm that searches for "the most likely word sequence *w* given the observation sequence *o*, and the acoustic-phonetic-language model" based on the target language.

Figure 1. *Basic architecture of an Automatic Speech Recognition (ASR) system.*



Note. Credit for the picture: Karpagavalli & Chandra (2016), p. 395.

While building ASR systems, developers provide sounds, words, and rules of one (or more) target language(s) to the two models and the lexicon. This process (alongside the training stage) aims to maximize the decoding processing of the speech signal and increase the accuracy of transcriptions from speech to text. When given an input, the a*coustic front-end* analyzes the waveforms and extracts the relevant feature vectors (i.e., a portion of information extracted from the waveform in a small-time window). This first component of the ASR system converts the signal from an analog, physical representation to a digital signal by applying a series of algorithms. These feature vectors are then passed to the *decoder*, where words are identified based on the extracted features and the hypothesized correspondence with the data contained in the acoustic (AM) and language (LM) models and the lexicon (measured with the *acoustic model* and *language model scores*) (*pattern-matching approach*, Karpagavalli & Chandra, 2016). The output of this decoding process is the written transcript of the hypothesized words: each word in the transcription is assigned a value which is the highest combined scores of AM and LM and indicates "the reliability or probability of correctness" (Jiang, 2005: p. 456) (Yu & Deng, 2015; Karpagavalli & Chandra, 2016; Jurafsky & Martin, 2025). Typically, outputs are generated within a few seconds; however, the time taken to produce the output may depend on several factors, such as the computational power of the systems or the quality of the speech signal that affects the duration of the process at the acoustic front-end level (Saraclar *et al*., 2010).

1.4 Factors affecting the performance of ASR systems

ASR is rapidly becoming one of the most relevant tools for multimodal access to information. This technology improves human-human interaction in settings where language barriers impede communication (for example, in the absence of interpreters or when they provide interpreting services only for one language) (Butler *et al*., 2019), providing multimodal access to information. These systems can be employed in different settings, such as work meetings, conferences, national and supranational institution sessions (e.g., national parliaments), and legal processes. Users can read whole transcripts after the end of such meetings, but they can also access information on a PC with the aid of real-time captions due to the fast speed of conversion of the speech signal into written transcripts, facilitating communication (e.g., Shimogori *et al*., 2010; Berke *et al*., 2017). The advantage of this method lies in the fact that machines transcribe speech inputs faster than humans. In this last case, the transcription process is a costly and time-consuming task that usually requires days to be carried out (Chan *et al*., 2019): ASR systems thus help users accelerate this process by automating it. However, even if ASR systems have specific algorithms implemented in their structure that clean the input signal (specifically, in the acoustic front-end), these systems have trouble with dealing with the processing of speech signals from many real-world applications. *Speech variability* is one of these factors that affect ASR systems' performance (Benzeghiba *et al*., 2007; Nakamura *et al*., 2008; Karpagavalli & Chandra, 2016; Alharbi *et al*., 2021; Feng *et al*., 2021; Jurafsky & Martin, 2025). Speech variability is linked to:

- Characteristics of speakers (e.g., age, gender, physical traits that affect voice structure);
- Sociolinguistic factors (e.g., regional or foreign accents, regional or ethnic dialects);
- Spontaneous speech (speech rate, connected speech, disfluencies such as false starts and hesitations, repairs);
- Speech type (i.e., read speech – where only a human reads a text to a machine - versus conversational speech – where two or more humans talk to each other);
- Emotions.

Each of these factors define each of us as unique speakers and challenge the ability of ASR to accurately decode speech signals. They alter the spectrum of the speech signal, affecting its features and preventing the correct decoding of sounds, lowering the accuracy of transcriptions and the

confidence of the ASR system in its transcription[17] (Benzeghiba *et al*., 2007; Alharbi *et al*., 2021). For example, if an ASR system is trained only with speech from speakers of standard dialects of a language, it will have difficulty when dealing with the mapping of sounds and recognition of words pronounced by speakers with under-represented varieties of a language or L2 speakers (Emara & Shaker, 2024; Feng *et al*., 2024).

Additionally, signal degradation may be caused by external factors, including the structure of the ASR system itself (Laurent *et al*., 2014; Alharbi *et al*., 2021; Emara & Shaker, 2024; Feng *et al*., 2024), environment, ambient noise and the quality of the hardware that collects the speech signals (Gillespie & Atlas, 2002; Benzeghiba *et al*., 2007; Karpagavalli & Chandra, 2016; Jurafsky & Martin, 2025). It is clear that, if these systems are not accurate enough at transcribing speech signals in different real-life contexts, there is a risk that these could not be a reliable aid to access information for a wide range of users, as errors in the text may hinder speech processing and content comprehension, potentially altering in a drastic way the message contained in the speech (Butler *et al*., 2019; Chan *et al*., 2019; Romero-Fresco & Fresno, 2023).

1.5 Evaluating the performance of ASR systems to ensure quality and reliability of their output

The improvement of the architecture of ASR systems has always posed challenges for engineers (O'Shaughnessy, 2024; Jurafsky & Martin, 2025). Ideally, ASR systems should function efficiently in a variety of challenging conditions. However, many of these systems still struggle with transcribing speech signals from real-world settings. This is often due to factors related to the unique characteristics of speakers, environment, and acoustics, or a combination of these elements that alter the structure of speech signals, making it challenging for ASR systems to analyze and decode the signals effectively (e.g., Dua *et al*., 2023). These challenges are evident not only in the number of correctly transcribed words and the overall accuracy of transcription, but also in the confidence with which systems produce their output (Li, 2018; Jurafsky & Martin, 2025). To assess the robustness and performance of ASR systems, engineers must assess the quality of the automatic transcriptions.

[17] See the next paragraph for a discussion on the importance of the two measures - accuracy and confidence level - when evaluating the performance of ASR systems.

This evaluation allows them to make necessary adjustments and provides users with a reliable system that ensures full access to information.

The following two paragraphs will introduce the two metrics that engineers use to evaluate the performance of ASR systems - namely, accuracy and probability of correctness of transcriptions (also known as *confidence scores*).

1.5.1 Assessing accuracy of transcriptions

It is fundamental to provide high-quality transcriptions to ensure access to information in all real-world contexts, to all individuals (Bain *et al*., 2002; Kuhn *et al*., 2024). Therefore, engineers need to evaluate the performance of ASR systems and assess their robustness. Professionals utilize various metrics to complete this task, with the *word error rate* (WER) serving as the standard evaluation metric (Goldwater *et al*., 2010). The WER compares the ASR-generated transcribed output (the hypothesis text, or *HYP*) with a *reference text* (*REF*) created by a human transcriber (Jurafsky & Martin, 2025). The REF text is prepared and pre-processed by professional transcribers to facilitate comparison with the HYP transcription. These transcribers utilize various linguistic strategies to minimize errors, particularly those arising from discrepancies between the REF and HYP texts, a process known as *text normalization* (Kuhn *et al*., 2024). For instance, such discrepancies may occur due to differences in spelling systems used by the human transcriber and the ASR system, leading to the detection of false errors[18]. These two texts are then aligned and compared using *sclite*, an open-source tool contained in the *Speech Recognition Scoring Toolkit* package (SRCT)[19] developed by the *National Institute of Standards and Technologies* (NIST)[20] (Jurafsky & Martin, 2025). *sclite* calculates the minimum edit distance in words between the two texts and lists the substitution, insertion, and deletion errors committed by the ASR system in the HYP text. The WER score is calculated using the following formula:

$$\text{WER} = 100 \times \frac{SUB+DEL+INS}{N}$$

[18] E.g., the presence of the word 'color' (American English spelling) in the REF could lead to the detection of a substitution error in the HYP by sclite if the latter contains the British English spelling ('colour'). The most common classes of words that undergo text normalization are numbers, dates, and monetary amounts (Jurafsky & Martin, 2025: Chapter 16, pp. 19-20).

[19] *Usnistgov/SCTK*. (2025). National Institute of Standards and Technology. https://github.com/usnistgov/SCTK (Original work published 2016).

[20] *National Institute of Standards and Technology*. (2025, February 5). NIST. https://www.nist.gov/

The sum of substitutions (SUB), deletions (DEL), and insertions (INS) divided by the total number of words in the REF (N) gives back a number between zero and one: the lower the result, the higher the accuracy of the transcription generated by the ASR system.

The WER score is not the only metric used for evaluating the performance of an ASR system. Over the years, researchers have criticized this metric since it fails to assess how the errors in the transcription affect the overall message (see Kuhn and colleagues, 2024, for a brief overview of some of the other metrics developed in recent years). Among those, the NER model (Romero-Fresco & Martinez, 2015) is the most innovative: it aims to assess the performance of the ASR systems "by analyzing the extent to which errors affect the coherence of the subtitled text or modify its content" (Romero-Fresco & Martinez, 2015: p. 1). The model was initially developed to evaluate the quality of live subtitles produced by *respeakers*, who are professional transcribers that repeat speakers' words into an automatic speech recognition (ASR) system. This system then generates automatic transcriptions that can be immediately corrected if errors occur. Unlike WER, professionals using the NER model must manually evaluate the type (editions and recognition errors) and severity (serious, standard, and minor) of errors, or any changes in the output (correct editions) (Romero-Fresco & Martinez, 2015). To calculate the overall score, they use the following formula:

$$\text{Accuracy} = \frac{N - E - R}{N} \times 100$$

The subtraction of the total number of words and commands (i.e., punctuation marks, speaker identification), the edition, and the recognition errors are divided by the total number of words and commands. The higher the outcome of the formula (range: 0-100%), the higher the accuracy, with 98% standing as the threshold for acceptable accuracy. The NER model has also been employed to assess the quality of automatic captions. In a recent study, Romero-Fresco and Fresno (2023) conducted a study where they compared the quality of human live captions (created by professional respeakers and stenographers) and automatic captions (generated by several different ASR systems) in English. Using the NER model, they assessed the quality of transcriptions from a data set of 798 minutes of live captions (2018-2022). The analysis revealed that automatic captions were less accurate than human live captions and fell below the recommended threshold for acceptable accuracy. However, the performance of ASR systems has been improving over time.

Like WER, the NER model has some limitations: for instance, human evaluators need extensive training, carrying out the scoring process is time-consuming, and the measures to assess the errors are subjective (Kuhn *et al.*, 2024). However, some of these drawbacks have been addressed recently, especially the partial automatization process of the scoring. The release of the *NER Buddy* - an AI-

based tool that helps evaluators in the scoring and assessment processes - appears to be imminent. This tool promises to partially accelerate and enhance the evaluation process of the quality of both human and automatic live captions using the NER model (Romero-Fresco *et al*., 2024).

1.5.2 Estimating reliability of ASR-generated transcriptions through *confidence score* values

As we previously mentioned in section 1.3, the final output of ASR systems (in this case, written text) is provided with a value that indicates the reliability of the recognition or probability of correctness of the transcription via a pattern-matching approach of the elements in the acoustic and language models and the acoustic observations in the speech signal (Williams, 1998). This value is called the *confidence score* (Li, 2018). It is a value comprised between 0 and 1, and it can be computed at the word, utterance, or speaker level (Jiang, 2005; Vertanen & Kristensson, 2008). The higher the confidence score, the higher the probability that the transcribed token was correctly recognized. Overall confidence scores are affected by the single values printed out by the acoustic and language models. If feature vectors are extracted and combined effectively (i.e., if the speech signal is clean enough and the features in the speech signal match those in the models), the confidence scores should be highly reliable (Li, 2018). However, ASR systems are not immune to errors, even when they output their transcriptions with the highest score (Jiang, 2005).

Engineers typically use this metric to assess the performance of ASR systems, but it can also be used to detect out-of-vocabulary tokens, eliminate transcriptions generated by disfluencies or noise, perform keyword spotting, and run dialogue systems (Williams, 1998; Li, 2018; Swarup *et al.*, 2019). However, previous research on the development and use of automatic captions and transcriptions emphasized the potential usefulness of visualizing the confidence levels with which the ASR system transcribed words through graphical features in the text.

1.6 Using color-coded markups to display confidence

This research topic began to circulate in the context of the *Liberated Learning Project* (Bain *et al.*, 2002). The project was launched in 1998 in Canada in association with IBM and a network of universities across Canada and the United States. The primary objective was to evaluate the

advantages of utilizing automatic speech recognition to provide ASR-generated captions and transcriptions for individuals with disabilities, enhancing their access to information and learning in higher education (Ryba *et al*., 2006). In one of their works, Wald and Bain (2008) proposed that users could be made aware of the level of confidence of the ASR system in its transcription by using color-coded markups (i.e., change of colors) or displaying the sounds using the International Phonetic Alphabet (IPA) notation in order to improve the reliability of the written output. Researchers then conducted several projects with different populations, developing different prototypes of display formats and investigating the usefulness of these graphical features by testing their impact on content comprehension. All the studies were carried out using user-centered designs, collecting opinions, feedback, and insights from users to optimize the development of these display formats (Vredenburg *et al*., 2002). For instance, Piquard-Kipffer and colleagues (2015) developed three display formats (orthographic, IPA, and pseudo-phonetic) to improve the reliability of the text output generated by an ASR system for deaf and hard-of-hearing persons. They subsequently tested the efficiency of these formats in displaying the degree of confidence of the recognized items through a two-phase study where participants were also asked to provide their opinions and insights on the different display formats. Results highlighted that users preferred to read the correctly recognized words in a bold font, while incorrectly recognized words were preferred to be displayed in the pseudo-phonetic format. However, users lamented the difficulty of remembering that the confidence scores computed by the ASR system were not entirely reliable since - for instance - some words were transcribed with high confidence but were incorrect. As a result, they found it difficult to trust the written output. An earlier study by Vertanen and Kristensson (2008) found similar results. They explored the effectiveness of providing visual feedback using a color-coded markup system (words were printed out in shades of red based on the degree of the system's confidence) to highlight low-confidence transcriptions during a dictation task. The researchers also employed questionnaires to collect users' opinions and thoughts on the various display formats. The findings of this study indicated that the color-coded markup was effective in helping users identify errors. However, it did not significantly enhance the speed and accuracy of word recognition, particularly for incorrect words with high confidence scores. The authors thus suggested that confidence scores should be used with caution when developing color-coded markups. Shiver and Wolve (2016) tested a similar color-coded markup to improve accessibility to information on the Internet for deaf and hard-of-hearing individuals. They asked their participants to watch four captioned videos with different display formats (automatic captions with a color-coded markup, automatic captions without a color-coded markup, no captions, and human-created captions) and answer some questions on the content of the video. Additionally, they were asked to complete a questionnaire to give their opinions on their viewing preferences. Results from

the comprehension task highlighted that users benefited from the presence of ASR-generated captions. In fact, the highest comprehension scores were observed in the group assigned to the experimental condition that included the color-coded markup. Additionally, the follow-up questionnaire revealed two key preferences among participants: 1) they preferred captions without the markup, but 2) they would appreciate the option to use ASR-generated captions with a color-coded markup that highlights potential errors in the transcription.

The most recent project that aimed to determine the usefulness of displaying confidence scores in automatic captions was led by Berke (2017). In this project, he investigated whether providing information about the confidence of ASR using graphical features in the text of captions would help deaf and hard-of-hearing persons in one-on-one meetings. Additionally, it aimed to investigate which markup was the best to provide this information visually. In their 2017 paper, Berke and colleagues reported the results of a study where they asked participants to express their opinions on the best markup to display confidence. Initially, a group of deaf and hard-of-hearing individuals was asked to evaluate a series of color-coded markups and select their preferred formats. In a second study, they asked another group of participants from the same population to watch four videos captioned with four display styles (no change, italics, underline, and yellow), answer a series of multiple-choice questions on the content of the videos, and fill out a questionnaire on their preferences and insights on the different markups. Results indicated that although participants initially showed interest in implementing such display formats in captions, they ultimately preferred not to have any markup indicating the confidence of the ASR system. Berke and colleagues proposed that these results might be related to the perceived uselessness of such markups. Additionally, participants indicated that the markups were distracting and confusing, and this conclusion was attributed to the extra layer of information that users needed to process within an already complex multimodal context (which included video, text, and the interlocutor). However, results from the comprehension task revealed a non-statistically significant difference between comprehension scores across conditions, suggesting that markups did not affect participants' performance.

1.7 Preliminary conclusions

Over the last few decades, information and communication technology advances have widened access to information and communication for a wide range of users. Electronic devices, websites, and mobile apps have been developed following guidelines on how to eliminate communication barriers

via *bi-* and *multimodality*, that is the simultaneous presentation of spoken and written modalities[21]. At the educational level, instructors have been encouraged to follow UDL guidelines to create multimodal content with the same aim[22]. Automatic speech recognition (ASR) is one of the tools that can be used to present information multimodally, supporting equity in accessing information for diverse learners and improving communication (e.g., Ryba *et al*., 2006; Wald, 2006a; Wald, 2006b; Wald, 2007; Wald & Bain, 2008; Butler *et al*., 2019). However, due to its weaknesses, this technology needs to be refined before being safely implemented in many real-world settings (e.g., O'Shaughnessy, 2024).

In this chapter, we presented ASR technology and its fundamental architecture (sections 1.3 and 1.5) while also briefly addressing the opportunities it offers (section 1.2) and the current challenges it faces (section 1.4). Specifically, we highlighted how the potential development and implementation of specific graphical features such as a color-coded markup could help users trust or not the ASR transcriptions in automatic captions, especially in the context of aiding L2 speech processing and content comprehension in educational settings (section 1.6). Since previous research on this topic with different populations yielded mixed results or is still lacking (e.g., Vertanen & Kristensson, 2008; Berke *et al*., 2017), we deem it important to continue researching the benefits or drawbacks this solution may bring to L2 speakers of English, our target population.

[21] See footnotes 7-9, page 6.
[22] See footnote 10, page 6.

2 Helping speakers process and learn an L2 through the use of AVT products[23]

A psycholinguistic overview of speech processing, comprehension and learning and the effects of bi- and multimodal input on these processes in L2 speakers

2.1 Introduction

This chapter will provide an overview of the psycholinguistic mechanisms involved in speech processing, comprehension, and learning with a focus on L2 speakers. It will discuss the challenges both populations encounter when dealing with these processes, as well as the benefits that the simultaneous presentation of written and auditory (or bimodal) inputs can bring to them. Lastly, the chapter will briefly examine the use of different audiovisual translation products - particularly intralingual (or same-language) subtitles and captions - to aid speech processing, content comprehension, and learning.

[23] **Disclaimer** - This chapter includes excerpts from the following publication: Pucci, M. (2023). Towards Universally Designed Communication: Opportunities and Challenges in the Use of Automatic Speech Recognition Systems to Support Access, Understanding and Use of Information in Communicative Settings. In *Design for Inclusion* (pp. 18–25). IOS Press. https://doi.org/10.3233/SHTI230394

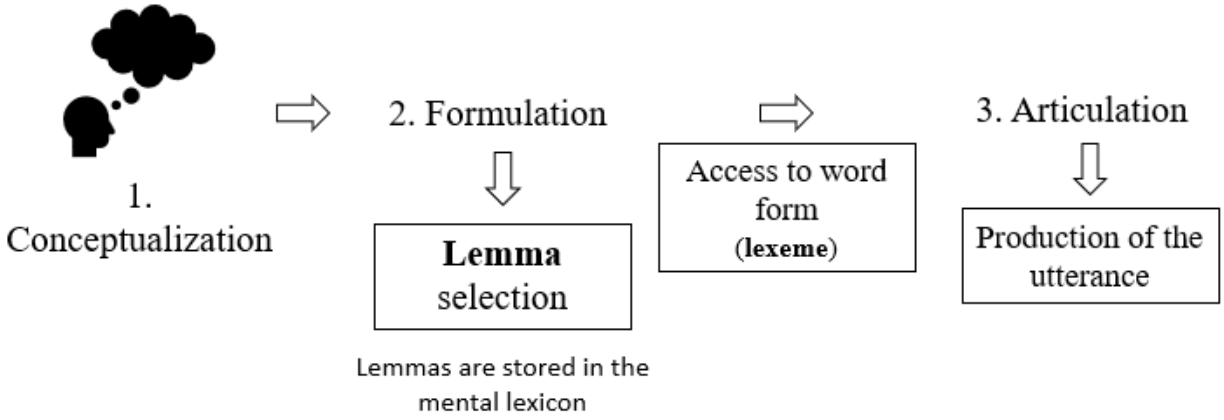
2.2 Psycholinguistic mechanisms underlying speech processing and comprehension

2.2.1 A brief overview of the main psycholinguistics mechanisms behind speech processing and comprehension

Language use is a dynamic process in which speech production and comprehension interact with one another (Pickering & Garrod, 2004).

From the language production standpoint, speakers intend to convey a message (*Figure 2*). Speakers first generate the message by building it on their general and situational knowledge (*conceptualization*, 1). Second, they must select those elements of language (words, or *lemmas* - the abstract concept of a word) that best express the intended message and arrange them in a structure by following the grammatical and syntactic rules of the language in which they want to express their message (*formulation*, 2). They then access the sounds (*lexemes*) related to the words they selected to convey their message, plan the articulatory movements, and finally articulate the utterance (*articulation*, 3) (Warren, 2013: pp. 15-16).

Figure 2. *Overview of the processes involved in language production.*



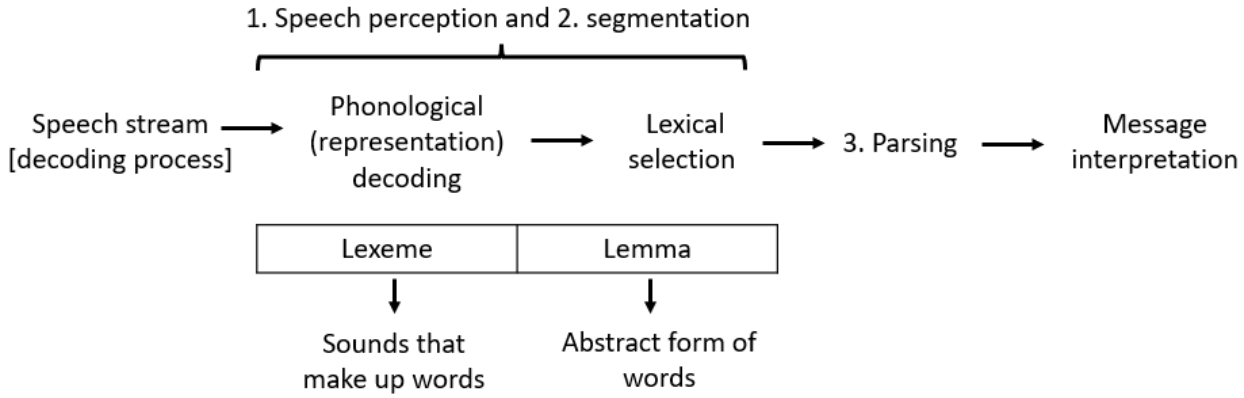
Note. The figure above was inspired by *Figure 2.1*, p.16 in Warren, 2013 (full reference in bibliography).

On the other hand, from the *language comprehension* standpoint (*Figure 3*), the receiver/hearer aims to process the utterance to extract its meaning and interpret the contained message (Garcia Lecumberri *et al*., 2010). The hearer needs to deal with a fast and continuous speech stream and

perform some tasks on it to extract the relevant features before reaching their goal of interpreting the message (Byrd & Mintz, 2010). The decoding process starts with the recognition of the sounds in the speech stream (*speech perception*, 1) (*Figure 3*). During this process, with the knowledge of the phonetic properties and phonemic repertoires of a language, the receiver segments the stream into the meaningful linguist building blocks, identifying the boundaries between the words and matching the sounds to their abstract form (*speech segmentation*, 2) (Cutler, 2012; Charles & Trenkic, 2015). The relationships between words and the internal structure of the utterance are recombined during *parsing* (3), ultimately leading to the interpretation of the message (Goh, 2000; Warren, 2013).

Although these processes may appear linear, they are highly interactive (Warren, 2013). In fact, in language comprehension (and production as well), bottom-up (perception, word segmentation, and word recognition) and top-down (i.e., contextual and situational information) processes continuously interact with one another and between various levels or representations (Cutler & Clifton, 1999; Warren, 2013).

Figure 3. *Overview of the processes involved in language comprehension.*



Note. The figure above was inspired by *Figure 1.2*, p.5 in Warren, 2013 (full reference in bibliography).

While carrying out these processes may seem effortless and automatic to some, it may be a taxing and complex task for others. In the next section, we briefly describe the factors that challenges that listeners encounter in their day-to-day life.

2.2.2 Challenges in speech processing and comprehension

One key aspect related to speech processing and content comprehension is the notion of *variability* (Warren, 2013). As we mentioned in the previous section, the first step in speech processing is *perception*, the process where we map the acoustic input onto the stored phonological (lexemes) and lexical (lemmas) representations (Romero-Rivas *et al.*, 2015). This process will run smoothly in all those situations where listeners' perceptual expectations will be met, and speech signals will be processed rather automatically (*Figure 4* - Van Engen & Peelle, 2014).

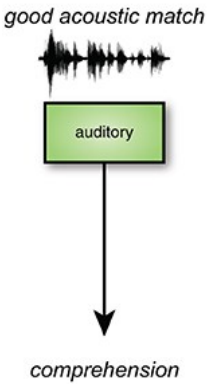


Figure 4. *Graphic representation of speech processing when the input (the speech signal) matches the perceptual expectations of listeners.*

Note. Credit for the picture: Van Engen & Peelle (2014), p. 2

The knowledge of the phonetic repertoire of a language, however, is not sufficient to successfully recognize the sounds in a speech stream. Our language systems need to be as flexible as possible to compensate for the differences in phonetic realizations. In the words of Anne Cutler (2012), "listeners cannot assume that speech will present each occurrence of a speech sound in the exact same form" (p. 36). In short, listeners need to be able to cope with *variability* - in this case, the natural differences in the realization of sounds (Warren, 2013).

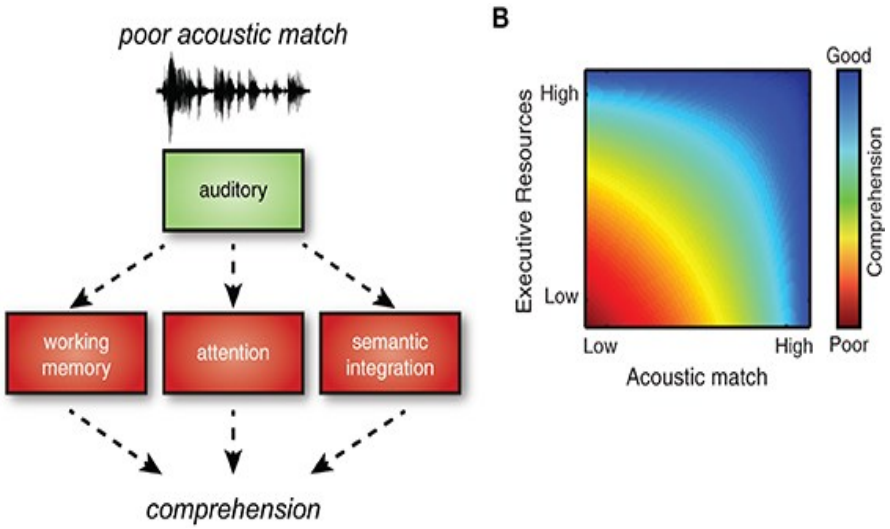
Speech variability defines the uniqueness of each speaker and can be due to individual factors, such as:

- The characteristics of speakers (e.g., physical traits that affect the articulatory gestures and the voice);
- Demographic factors (e.g., age, sex, educational background);
- Sociolinguistic factors (e.g., regional or foreign accents, regional or ethnic dialects);
- Conversational speech (where phenomena of connected speech and coarticulation can be found in higher proportions compared to read speech - Farnetani & Recasens, 1997);

- Emotions (which affect prosody and intonation) (Bäzinger & Scherer, 2005; Byrd & Mintz, 2010; Mattys *et al*., 2012; Warren, 2013; Boland *et al*., 2016).
- Acoustic mismatches between the sounds in the speech stream and the speaker's phonetic repertoire could also be caused by *external factors* (e.g., Mattys *et al*., 2012). For instance, environmental noise (reverberation, distance between speakers, etcetera) can affect intelligibility[24] and speech perception.

Although individuals can rapidly cope with natural variability and adapt to understand unfamiliar phonetic realizations (Romero-Rivas *et al*., 2015; Pisoni, 2018; Brown *et al*., 2020), the concurrent presence of one or more factors could cause problems in speech perception, potentially hindering speech comprehension (Adank *et al*., 2009; Warren, 2013). The more numerous the acoustic variations and the greater the acoustic mismatch between the signal and the phonetic repertoire, the higher cognitive resources (e.g., working memory, attention) may need to be recruited to carry out the process effectively (*Figure 5* - Van Engen & Peelle, 2014). Therefore, a high degree of acoustic mismatch could lead to an increase in the *listening effort* and *cognitive demand*, impeding recognition of the elements in the speech stream and further processing (such as word segmentation; Byrd & Mintz, 2010; Mattys & Wiget, 2011; Van Engen & Peelle, 2014; Peelle, 2017; Porretta & Tucker, 2019; Porretta *et al*., 2020).

Figure 5. *Graphic representation of speech processing when the input (the speech signal) poorly matches the perceptual expectations of listeners.*



Note. Credit for the picture: Van Engen & Peelle (2014), p. 2

[24] Intelligibility refers to the "misidentification of words" (Romero-Rivas *et al*., 2015: p. 2).

In sum, adverse conditions can lead to a failure in mapping the phonetic representations to their conceptual representation, perceptual interference both at low (within the signal) and high (between more than one signal) levels, and reduced attentional and memory capacity (Mattys *et al*., 2012).

Accented speech is a well-known factor that affects speech processing (e.g., Mattys *et al*., 2012). It is characterized by differences in segmental and suprasegmental elements from native speaker pronunciation norms (Mattys *et al*., 2012). Suppose a speaker needs to process a speech signal in an unfamiliar accent. In that case, these differences may lead to an increase in listening effort, affecting the speed of processing of the speech signal and potentially leading to a mapping failure of the phonological representations onto the lemmas, hindering overall processing (Munro & Derwing, 1995; Floccia *et al*., 2009; Mattys *et al*., 2012: p. 955; Romero-Rivas *et al*., 2015) and hampering communication. Accented speech reduces intelligibility and comprehensibility (Van Engen & Peelle, 2014); however, research has highlighted that native speakers quickly adapt to unfamiliar accents through exposure and with the aid of *perceptual learning* - an adaptation strategy listeners use "to categorize ambiguous phonemes using lexical information" (Bradlow & Bent, 2008; Romero-Rivas *et al*., 2015: p. 2).

Speech variability poses challenges for L1 speakers, but it is even more challenging for L2 speakers (Grey *et al*., 2018). In the case of L2 learners, some factors can contribute to more significant difficulties with speech processing in addition to those listed in the previous section:

- Limited or incomplete knowledge of the L2 has a detrimental effect on speech processing, which, combined with environmental and acoustic factors, can result in poor comprehension (Garcia Lecumberri *et al.*, 2010; Mattys *et al*., 2012);
- Quality and rate of exposure to authentic L2 input and its sociolinguistic variants influence familiarity with different phonetic realizations (Garcia Lecumberri *et al.*, 2010);
- The structure and knowledge of their L1 tend to affect speakers' perception of the sounds of their L2, resulting - for instance - in a greater susceptibility to linguistic phenomena such as the lexical competition effects (Grey *et al*., 2018).

L2 learners may encounter significant challenges when faced with speech signals influenced by various factors from these lists. Compared to native speakers, they will face increased listening effort and higher cognitive demands on attention and working memory (Grey *et al.*, 2018). The complexity

of these types of speech signals could, therefore, severely hinder speech processing and comprehension, potentially halting communication and creating an invisible (yet evident) barrier. However, like L1 speakers, learners can retune the perceptual categories of their L2 and improve speech processing (e.g., Mitterer & McQueen, 2009).

The transient and continuous nature of speech makes speech processing and listening comprehension complex skills to master for L2 learners. Additionally, variability within and between speakers further makes this task challenging (Jia & Hew, 2023). Therefore, L2 speakers need to automatize the word recognition process with time and exercise in order to develop and optimize their listening comprehension skills (Goh, 2000; Ke & Wang, 2022). In her seminal study, Goh (2000) investigated and discussed the ten most common difficulties L2 speakers encounter when dealing with listening comprehension. She carried out semi-structured interviews and collected self-reports from a group of L2 speakers of English (Chinese native speakers, university students), then analyzed and discussed the results within the cognitive framework Anderson developed in 1995 also on the basis of their listening abilities[25]. Both low- and high-ability listeners primarily lamented difficulties in word recognition and lack of attention. As learners of English, they struggled to chunk speech signals effectively and had trouble identifying both known and unknown words due to the poor automatization of the word recognition subprocess, which is crucial for enhancing their listening comprehension skills (Ke & Wang, 2022). Both groups also struggled to remember what they heard, likely due to listening effort and cognitive demands. Lastly, while participants in the low-ability group reported that many of their problems arose from the processing inefficiency at the low levels of comprehension (speech perception in particular), high-ability participants stated that they often understood the individual words in the speech streams but struggled to grasp the overall message of the utterances due to a lack of contextual information.

In sum, speech processing is a rather challenging skill to master - especially under certain circumstances and for diverse populations, such as L2 learners of English (Goh, 2000; Jia & Hew, 2020) - due to the:

- The transient and continuous nature of speech that makes it difficult to efficiently segment and recognize words unless these processes are explicitly exercised;

[25] Listening abilities were measured using the SLEP test. Two groups were formed based on their score in the listening test – the high- and low-ability groups: "students in the high ability group had a range of scores that was equivalent to TOEFL scores of 550±600, while those in the low ability group had a range of scores equivalent to TOEFL 440±500" (see Goh, 200: p. 66 for details).

- The natural variability across speakers and the acoustic characteristics of environments that potentially affect the speech signals;
- The high demands of cognitive resources and listening effort necessary to carry out speech processing and comprehension.

Despite these challenges, is there a way to optimize this process and aid speech comprehension and learning? In the next section, I will discuss how the simultaneous presentation of aural and written linguistic inputs (namely, the bimodal input) can be used to overcome some difficulties that listeners encounter while engaged in listening.

2.3 Effects of bimodal input on speech processing, comprehension and learning in L1 and L2 speakers

In the last two section, we have seen that speech processing and comprehension are complex tasks for all listeners. In fact, regardless of being a native or non-native speaker of a language, the natural variability among speakers and the acoustic properties of some environments can pose challenges during listening comprehension. A question arises: how can we assist these individuals? One possible answer lies in the way we present the linguistic input. For instance, we could present it using two modalities simultaneously - that is, the auditory and written modalities. This solution could help listeners carry out these tasks since written text is more stable and less transient input than speech. This strategy could also help them lower the cognitive demands and listening effort required by the sub-processes involved in speech processing and comprehension.

The term *bimodal input* is used "to describe the simultaneous presentation of matching aural and orthographic stimuli" (Charles & Trenkic, 2015: p. 4).

Research has investigated the effects of the simultaneous presentation of matching auditory and written stimuli on speech processing, comprehension, and learning to provide insights into how captions enhance these processes. For instance, Adesope and Nesbit conducted a meta-analysis in 2012 to investigate the effects of different input modes (spoken-only, written-only, concurrent spoken-written presentation of an input) on learning. The analysis results showed that bimodal input improved learning compared to spoken-only presentations. Moreover, the analysis highlighted that the magnitude of this beneficial effect was linked to the educational level of students (L1 learners) and their proficiency level of knowledge (L2 learners). Research on the effects of reading-while-

listening has also highlighted the benefits to comprehension, fluency, and incidental vocabulary learning (e.g., Conklin et al., 2020), however, findings seem to be contrasting (Clinton-Lisell, 2023).

Another example of bimodal input is represented by *subtitles* and *captions*. Some researchers used this type of input to investigate more in depth the effects of the simultaneous presence of aural and written stimuli on bottom-up processes - namely speech perception and segmentation -, perceptual retuning, and vocabulary learning (e.g., Bird & Williams, 2002; Charles & Trenkic, 2015). For instance, building from Vanderplank's 1988 study[26], and recent findings regarding the benefits of captions, Bird and Williams explored in a two-experiment study how different presentation modes (*text-only*, *sound-only*, *text + sound*) boosted word learning by measuring the improvements of spoken word recognition and recognition memory in native and non-native speakers of English. In the first experiment, Bird and Williams asked a group of 32 native ($N = 16$) and non-native speakers of English ($N = 16$) to complete a lexical decision task where participants had to state if they were familiar or unfamiliar with a list of words presented in different input modes. Accuracy (explicit memory) and reaction times (decoding speed and implicit memory) were collected as behavioral measures. In the second experiment, the researchers tested 24 advanced L2 speakers of English in rhyme monitoring[27] and recognition memory tasks, again measuring reaction times and accuracy. Overall results highlighted that exposure to the bimodal input boosted learning for nonwords – that is, when new phonological forms (i.e., lexemes) needed to be encoded. Furthermore, the results demonstrated that participants' ability to recognize words improved (higher accuracy scores and shorter reaction times) when items were presented in the bimodal input compared to the 'sound-only' condition.

Similar to Bird and Williams (2002), Charles and Trenkic designed a two-experiment study to investigate the role of bimodal input on speech segmentation in L2 learners of English. In the first experiment, ten international university students with different linguistic backgrounds (L1s and years spent learning English) were instructed to repeat sixty excerpts from TV programs and films delivered in standard British English. In the second experiment, another group of twelve international university students participated in a 4-week pre-test/treatment/post-test experiment to assess the effects of various input modalities ('bimodal', 'no subtitles', and 'no sound' conditions) on lexical segmentation. For this experiment, participants were asked to complete a pre-test phase to assess their baseline performance. Then, during the treatment phase (two weeks), they watched the videos in their assigned

[26] Vanderplank, R. (1988). The value of teletext subtitles in language learning. *ELT Journal*, 42: p. 272-281.
[27] In this task, participants are asked to decide if a target item rhymes with a cue item. Items can be words or non-words. This type of task is useful to measure word learning through an implicit learning mechanism called priming (the tendency to produce, repeat or process faster a lexical item or a syntactic structure after having encountered a related lexical item or syntactic structure – Bock, 1986; Hoey, 2012).

condition and completed immediate post-tests, before concluding the study with a delayed post-test to assess the potential improvements. During the study, they always completed *listening shadowing tasks*, which consisted of listening to short excerpts from familiar and unfamiliar videos and repeating whatever words they heard (Mitterer & McQueen, 2009; Charles & Trenkic, 2015). Results show that L2 speakers of English potentially have problems with speech segmentation and that this process can be aided by the bimodal input - i.e., captions. In fact, the participants in the 'bimodal' condition group outperformed the control groups after the treatment phase.

Two studies have investigated the role of audiovisual translation products in aiding speech perception and perceptual learning. The first one was conducted by Mitterer and McQueen in 2009. Specifically, this study aimed to assess if intralingual and interlingual subtitles could aid L2 speakers of English in learning unfamiliar accents; in other words, the researchers were interested in determining if "the retuning of phonemic categories could be induced by orthographic information" (Mitterer & McQueen, 2009: p. 2). Researchers asked 121 Dutch university students with good knowledge of English to watch a video featuring unfamiliar regionally-accented English in one of the three conditions (intralingual subtitles - English, interlingual subtitles - Dutch, no subtitles). After viewing the video, participants performed a listening shadowing task, where they had to repeat a total of 160 excerpts. This group of items included 80 excerpts taken from the video they watched, and 80 new excerpts taken from the same audiovisual content. Results highlight that this type of retuning can indeed happen: in fact, participants in the 'intralingual subtitles' condition group outperformed those in the 'no subtitles' condition group, and that enhancement was found both for old and new items. Moreover, participants in the 'interlingual subtitles' condition group performed worse than those in the 'no subtitles' condition group when repeating new items. This suggests that the text in the interlingual subtitles may have provided inconsistent phonological information compared to the audio, potentially leading to interference between the two existing phonological systems.

The second study conducted to investigate the effects of audiovisual translation products on speech perception was designed by Birulés-Muntané and Soto-Faraco in 2016, finding similar results to Mitterer and McQueen (2009). Their study aimed to investigate the benefits of intralingual subtitles on speech perception in non-proficient L2 speakers of English. Sixty university students (L1s: Catalan, Spanish, or Italian) were assigned to one of the experimental conditions (‘intralingual English subtitles’, ‘interlingual Spanish subtitles’, ‘no subtitles’) and watched an episode of a TV series. Then, they were asked to complete three tests: a listening task, a vocabulary task, and a comprehension test to assess speech perception, vocabulary acquisition, and content comprehension. Results showed that students who watched the video with the 'intralingual subtitles' condition

improved their speech perception skills and vocabulary compared to the other conditions (‘interlingual subtitles’, ‘no subtitles’), confirming the hypothesis that intralingual subtitles support perceptual learning, aiding speakers in phonological retuning.

Up to this point, we have reviewed the most relevant studies concerning the use of bimodal input to aid speech processing. We will now move to discuss more generally the benefits of using audiovisual translation products – specifically, intralingual subtitles and captions - to support speech processing, comprehension, and learning in L2 speakers of English in their day-to-day lives and educational contexts.

2.4 The use of audiovisual translation products to aid speech processing, content comprehension, and learning

2.4.1 Brief overview of the audiovisual translation products

Audiovisual translation[28] has always served a social function by making audiovisual products accessible to those individuals who would otherwise lack access to such materials and the associated information (Díaz Cintas & Remael, 2014). The main products of audiovisual translation are *intralingual* (or *same-language*) *subtitles*, *interlingual subtitles*, and *captions*.

To create subtitles, professionals need to modify (i.e., condense and reduce) the text of a script or speaker's speech due to time and space constraints in the audiovisual content (Diaz Cintas & Remael, 2014; Szarkowska *et al*., 2024). Subtitles can be in the same (*intralingual subtitles*) or in a different language of the oral speech (usually, a person’s native language – *interlingual subtitles*) and can be created offline or in real-time during live events (3PlayMedia, 2023) (see section 2.4.3 for a more detailed description of *live subtitles*). *Captions*, on the other hand, are transcripts of speakers' original speech, which also contain indications of the music and sound effects in the video (Díaz Cintas & Remael, 2020). Similarly to subtitles, captions can be created offline or in real-time by ASR systems (3PlayMedia, 2023). The differences between these two audiovisual products also lie in their *display format* - that is, in the way they are displayed on screen. For instance, while subtitles are usually

[28] Audiovisual translation is one of the sub-disciplines included in translation studies (see Wang & Daghigh, 2024 for a review of the literature). Here we report the definition of audiovisual translation in the words of Díaz Cintas and Remael: “[it is an umbrella term] used to encapsulate different translation practices used in the audiovisual media in which there is a transfer from a source to a target language, which involves some form of interaction with sound and images” (Díaz Cintas & Remael, 2014: p. 12).

shown in a white font with no background on the screen, captions are shown as white text in a black box to maximize readability (3PlayMedia, 2023).

The use of audiovisual translation products to aid content comprehension and access to information has a long history. The first instances of this practice go back to the silent movies in 1920s with the use of *intertitles,* title cards used to display spoken utterances, dialogues, and descriptions of the narration that were inserted between sequences of the movies (Nagels, 2012). With the introduction of sound in films, intertitles were replaced by interlingual subtitles, an economical alternative to *dubbing* that ensured access to audiovisual products for individuals speaking a language different from the one spoken in the movie (Ivarsson, 2009). Another step forward was made with the advent of television in the 1970s when real-time captions began appearing in TV shows to guarantee access to information for Deaf and Hard-Of-Hearing individuals (Kuo, 2004; Venturini, unpublished master's thesis, 2022). Nowadays, a large number of viewers use audiovisual translation products to aid speech processing and content comprehension (Gernsbacher, 2015). Captions are primarily used to guarantee access to information for diverse users (3PlayMedia, 2024). In some countries, the use and production of this type of audiovisual translation products are regulated by laws to guarantee access to information to people with disabilities (e.g., see the *21st Century Communications and Video Accessibility Act* in the US)[29] (Venturini, unpublished master's thesis, 2022). However, technological advancements such as the diffusion of DVDs and streaming platforms on the Internet have contributed to an increased usage of captions among all users. A survey conducted by the web platform Preply revealed that 51% of the Americans out of the 1500 individuals who were interviewed frequently use audiovisual translation products when watching audiovisual content (especially the younger generations of viewers) (Preply, 2024). They use captions and subtitles not only to improve content comprehension, but also to aid speech processing (for instance, to become familiar with unfamiliar accents; Preply, 2023; 2024).

2.4.2 Benefits of audiovisual translation products on L2 speech processing, comprehension and language learning

Over the past 25 years, technological advancements have greatly increased daily access to the English language. As a result, individuals around the world are now more frequently exposed to authentic linguistic input (Montero Perez, 2022; Jia & Hew, 2023). Portable devices, such as

[29] *Twenty-First Century Communications and Video Accessibility Act | Federal Communications Commission*. (2010). Retrieved March 15, 2025, from https://www.fcc.gov/cvaa.

smartphones and tablets, along with social media and video streaming platforms on the Internet, are significantly contributing to this trend. Moreover, the fact that streaming platforms provide various types of audiovisual translation products to improve information accessibility for all viewers has led L2 learners and speakers to start using these tools to support speech processing, learn new languages or improve their already-existing linguistic knowledge (Dizon & Thanyawatpokin, 2021; Jia & Hew, 2023; 3PlayMedia, 2019; 2024). For instance, research by Dizon (2016) has highlighted how L2 speakers benefit from using intralingual and interlingual subtitles to enhance listening comprehension and vocabulary learning while watching audiovisual products in English on a streaming service.

Research on the benefits of using audiovisual translation products to improve language knowledge has been conducted since the 1980s (see Vanderplank, 2010 and 2013 for a review of the literature over 30 years). As of today, the main findings on the topic highlight that audiovisual translation products have positive effects on speech processing, listening comprehension, memory for content, and overall literacy in many populations (Montero Perez *et al*., 2013; Gernsbacher, 2015; Montero Perez, 2022). Specifically, existing research has revealed benefits for (see Montero Perez, 2022 for a recent review and Montero Perez *et al*., 2013 for a meta-analysis):

- Vocabulary learning (e.g., Sydorenko, 2010; Montero Perez *et al*., 2014; Gass *et al*., 2019; Montero Perez, 2020; Teng, 2022);
- Speech perception and segmentation (e.g., Bird & Williams, 2002; Mitterer & McQueen, 2009; Charles & Trenkic, 2015; Birulés-Muntané & Soto-Faraco, 2016);
- Pronunciation (e.g., Wisniewska & Mora, 2020);
- Overall improved comprehension of the content of the video (e.g., Montero Perez *et al*., 2013);
- Overall development of listening abilities (Yeldham, 2018).

Studies on the effects of on-screen text on grammar and syntax acquisition remains limited, but, in general, positive findings have emerged (e.g., Ghia, 2012; Pattermore & Muñoz, 2020; Montero Perez, 2022).

Even though a large part of the research on the topic has assessed the positive effects of audiovisual translation products on comprehension and acquisition, there is still a need for further investigation into the influence of individual characteristics and natural variability among learners and speakers. For instance, it remains unclear what the impact of language proficiency, language background, and cognitive processes (e.g., working memory, attention) is on the magnitude of the beneficial effects of audiovisual translation products on speech processing and content comprehension (e.g., Gass *et al*.,

2019). This difficulty in generalizing results is due to the varying methods and definitions used across studies, therefore further research is needed (Yeldham, 2018; Montero Perez, 2022; To, 2024). Regarding the role of *proficiency* in moderating the effects of captions on listening comprehension, research has yielded mixed results. For instance, the meta-analysis conducted by Montero Perez and colleagues in 2013 did not find any benefit in listening comprehension and vocabulary learning of using captions for beginning learners of English as an L2, but only for intermediate and advanced learners. Similarly, in Taylor's study (2005), low-proficient participants did not benefit from captions; they felt the captions were distracting and struggled to process the various inputs from the captioned video. Conversely, Venturini and colleagues (2022) discovered that low-proficient learners experienced greater improvements in content comprehension when using word-by-word, incremental automatic captions compared to traditional two-line, human-created captions. In contrast, high-proficient speakers of English as an L2 found the classic two-line captions to be more beneficial for content comprehension. In summary, research still needs to clarify how L2 speakers' proficiency affects the potential benefits of audiovisual translation products on speech processing and content comprehension, especially at lower levels of linguistic competence (e.g., Matielo *et al*., 2015; Pujadas, 2019).

2.4.3 The use of live subtitles and automatic captions in educational settings

While many studies suggest that audiovisual translation products enhances speech processing and content comprehension and facilitates word learning, other studies in the context of multimedia learning (Mayer, 2002; 2005) have reported contrasting findings. Specifically, previous research has not yet fully clarified the effects of audiovisual translation products on speech processing and comprehension in specific settings, such as in *educational settings* (e.g., during university lectures).

Attending a lecture where a professor speaks while slides are projected is a typical setup in educational settings, especially at the academic level. The *multimedia learning* theory states that in these settings, a learner builds a mental representation from the combined action of words (the explanations of the professor) and pictures (in the slides), therefore benefitting from a rich multimodal input (Mayer, 2002; 2024). But what about the use of audiovisual translation products in these settings? Would they help students with speech processing, content comprehension, and learning like other audiovisual translation products do in other settings? Existing literature on the topic has

preliminarily highlighted that the use of live subtitles[30] or automatic captions in this context could be beneficial for all students in the courses delivered in English, which has become a language of instruction due to the internationalization of academic institutions and the presence of a multilingual and multicultural student population (van Gauwbergen *et al*., 2024). In this context, both intrinsic speech variability in instructors (e.g., accent, speech rate), individual differences in students (e.g., language proficiency), as well as the acoustic characteristics of lecture halls, may challenge L2 speakers when processing speech, causing an increase in the load of the cognitive processes (Goh, 2000; Mattys & Wiget, 2011; Mattys *et al*., 2012 for a review; van den Heuij *et al*., 2018). Therefore, based on previous literature and findings, the use of audiovisual translation products in educational contexts should be helpful, and L2 speakers may benefit from the bi- and multimodal input. However, some researchers may argue that presenting linguistically redundant input (subtitles/captions) along with spoken narration and diagrams/images/written text on the slides has the potential to increase the cognitive load (i.e., on working memory and attention) imposed on the learners' processes to the extent that it hinders comprehension and learning (*redundancy effect* - Sweller, 2005; Kalyuga & Sweller, 2014; Kalyuga *et al*., 2004 and *split-attention effect* - Ayres & Sweller, 2014). For instance, a study conducted by Mayer and colleagues in 2014 showed that adding captions to a video lecture in English did not provide any benefit to L2 speakers of English in terms of improving content comprehension. In contrast, a study conducted by Chan and colleagues in 2022 found that L2 speakers of English understood the content of a video lecture better when interlingual subtitles were present, rather than when interlingual subtitles were provided. Additionally, the results indicated that there was no increase in cognitive load for the students in either condition, suggesting that the redundancy effect did not apply in this context. In sum, research on the effects of audiovisual translation products with this type of multimodal input has found contrasting results.

Similarly, research on the effects of *live subtitles* and automatic, ASR-generated captions on content comprehension in educational contexts is still ongoing (Chan *et al.*, 2019; van Gauwbergen *et al*., 2024).

Many studies investigating the usefulness of automatic captions and transcriptions in educational contexts were conducted during the *Liberated Learning Project* (LLP - see Chapter 1, section 1.6 for

[30] Nowadays, live (intralingual and interlingual) subtitles are produced in real time by *respeakers*, professionals who "listens to the original sound of a (live) program or event and respeaks it, including punctuation marks (…), to a speech recognition software, which turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay" (Romero-Fresco, 2011: p. 1 in Robert *et al*., 2021: p. 54). The effects of live subtitles on content comprehension in diverse users is still a novel topic of research (Robert *et al*., 2021; van Gauwbergen *et al*., 2024). Respeakers tend to be more accurate when transcribing speech in comparison with the performance of ASR systems (Romero-Fresco & Fresno, 2023), and for this reason, this technique may be used more extensively in educational contexts (Bolaños García-Escribano *et al*., 2024).

details) (e.g., Bain *et al*., 2002; Wald & Bain, 2008). The project was developed to investigate the reliability of the automatic speech recognition technology to "provide universal access to lecture material for students with diverse backgrounds" (Ryba *et al*., 2006: p. 1) through the use of automatic captions during lectures and automatic transcriptions to support notetaking and revision of previous notes (Ryba *et al*., 2006; Wald, 2006a; Wald, 2006b; Wald, 2007; Wald & Bain, 2008). Experimental studies conducted in this project have highlighted the benefits for students (e.g., an increase in word recognition and better comprehension of the content of lectures), but they have also sought to surface common problems with current ASR systems. The major problem, however, was *the lack of accuracy in transcriptions* (Ryba *et al*., 2006; Wald, 2006a; Wald, 2006b; Wald, 2007; Wald & Bain, 2008; Butler *et al*., 2019).

In this framework, one of the most important goals was also to develop and optimize a technological solution that would guarantee access to accurate information. Indeed, Bain and colleagues, in their 2002 paper, stated that "word accuracy is arguably the project's most important critical success factor, whether for display in the classroom, used as lecture notes, or both," as well as an acceptable timing of generation and display of the transcribed words (Bain *et al*., 2002: p. 194). In many studies of the *Liberated Learning Project*, a speaker-dependent[31] ASR system was used to generate reliable automatic captions and transcriptions. For instance, Ryba and colleagues (2006) conducted a study in the framework of LLP on the use of a speaker-dependent ASR system to create automatic captions and transcriptions in the academic setting. They enrolled 160 L1 and L2 speakers of English in a university course to participate in the study. Then, researchers asked participants to attend some lectures of their course, use the transcriptions generated by the ASR system, and complete three activities in which they reported their opinions and insights on the *Liberated Learning* project. Most importantly, Ryba and colleagues reported that students vocalized the importance of providing accurate transcriptions to aid speech processing, content comprehension, and learning.

The role of accuracy of automatic captions and transcriptions in educational contexts has also been investigated outside the LLP context. For instance, Butler and colleagues (2019) collected the opinions of deaf and hard-of-hearing students about the use of ASR systems to aid content comprehension and their confidence in using this system in class during lectures. Similarly to previous research, participants reported that they found automatic captions helpful for enhancing content comprehension. They also emphasized the importance of providing accurate and well-timed

[31] *Speaker-dependent* ASR systems are designed to be trained with the input of an individual user. On the contrary, *speaker-independent* systems are designed "to deal with the acoustic variability intrinsic in the speech signals coming from many different talkers, often with notably different regional accents." (Juang & Rabiner, 2005: p. 11).

transcriptions of the instructors' speech so that automatic captions become their primary source of information during lectures.

Cao and colleagues (2018) examined the effects of ASR-generated captions on non-native speakers' listening comprehension by designing a two-experiment study. Following the study Goh conducted in 2000, the first experiment aimed at highlighting the listening difficulties of non-native speakers, understanding which processes could be aided by using automatic captions, and the strategies of use of automatic captions of these speakers. Twenty non-native speakers of English were asked to listen to a series of ASR-generated captioned and uncaptioned audio clips (average WER score: 10%) while an eye tracker monitored their eyes, and to press a button every time they encountered a comprehension problem. Results showed that participants mainly struggled with speech segmentation and word identification, and that errors in the transcriptions confused and distracted them. Specifically, the errors in the transcriptions hindered the process of solving the listening problems encountered while listening to the experimental items. Data collected using the eye-tracking methodology showed that participants mainly used two strategies when using automatic captions. Non-native speakers either read the entire string while listening (to increase confidence in their listening abilities) or occasionally looked at the text on the screen based on their difficulties in listening comprehension (for instance, when they wanted to confirm what they heard in the utterance). Importantly, some participants reported a higher cognitive load when speech and text were not aligned (i.e., not appearing simultaneously - a similar outcome was also reported by Shimogori and colleagues in their study, 2010, where participants emphasized the importance of having the text closely aligned with the speech signal) and difficulties in processing the multimodal input. On the other hand, the second experiment aimed to investigate further the strategies displayed by non-native speakers in the first experiment. Twenty-two participants were asked to complete a second listening task where transcriptions were presented in two conditions. In the first condition, the text was aligned with the speech, and the interim output generated by the ASR system was displayed as soon as possible on the screen (*speed-oriented* display format). In the second condition, a delayed written output was provided, showing only the final results of the ASR system (*accuracy-oriented* display format). Results showed that non-native speakers who focused on the aural input preferred the speed-oriented display format, while those who read the written text while listening favored the accuracy-oriented display format. Results for this second experiment again underlined the importance of providing accurate transcriptions to non-native speakers of a language to aid listening comprehension and abilities in general.

Lastly, Chan and colleagues (2019) compared the effects of human- and ASR-generated subtitles on cognitive load and learning in L1 and L2 speakers. A total of 92 students were assigned to one of the three experimental conditions ('no subtitles', 'automatically generated subtitles', and 'corrected subtitles'). They were then asked to 1) complete a pre-test task to assess their previous knowledge on the topic, 2) watch a 25-minute video clip extracted from a longer, seminar-style video lecture on a topic related to Business and Economics, 3) complete a comprehension task and a self-evaluation report on cognitive load to measure the perceived effort while watching the video. Aligned with previous studies, these results indicate that accuracy (as well as the speed of presentation of the subtitles, which was too fast for students to read the text accurately) significantly affected content comprehension and cognitive load.

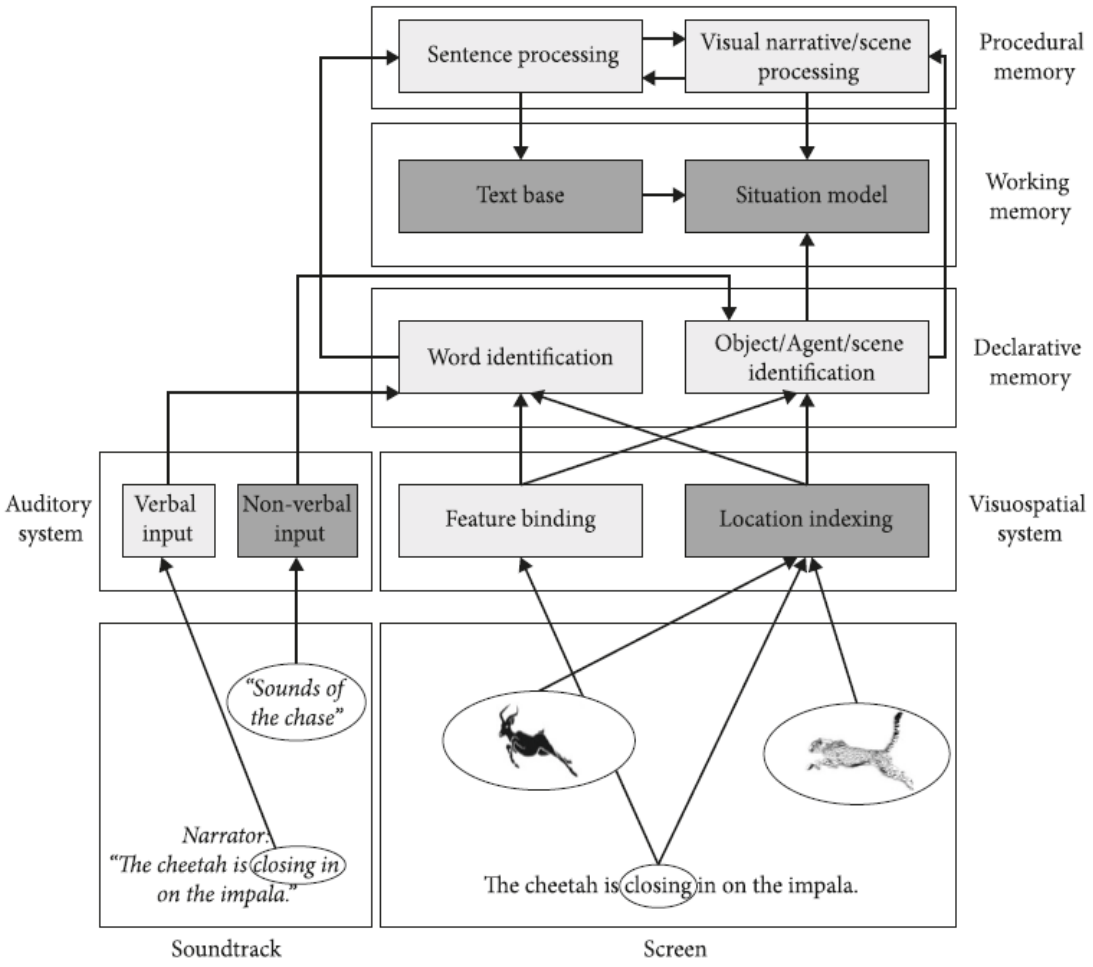
In summary, existing literature indicates that automatic captions and transcriptions can negatively impact speech processing and comprehension if the text contains a high percentage of errors or is poorly aligned with speech.

2.4.4 Cognitive theories of multi-modal input processing

Before closing this chapter, I will briefly introduce the most recent model of multi-modal input processing. Audiovisual content, such as movies, comprises various types of inputs: spoken dialogue, written text, and visual imagery (Montero Perez, 2022). Watching movies is a leisure activity; however, it can be cognitively demanding for viewers, as they must coordinate several cognitive tasks simultaneously - namely listening, reading, and watching the visuals - while processing a *multimodal input*. Formally, *multimodality* is defined as "... the use of several semiotic modes in the design of a semiotic product or event, together with the particular way in which these modes are combined – they may, for instance, reinforce each other (‘say the same thing in different ways’), fulfill complementary roles [e.g., when pictures complement a story in a children’s book] or be hierarchically ordered [e.g., dramatic music in an action movie]" (Kress & Van Leeuwen, 2001: p. 20 in Montero Perez, 2022: p. 3).

In the last 25 years, researchers have proposed several cognitive models to describe the mechanisms involved in the processing of multimodal audiovisual content (see Kruger & Liao, 2022 for a review). The most recent model is the *Multimodal Integrated-Language Framework* developed by Liao and colleagues (2021) (*Figure 6*).

Figure 6. *Diagram representing the Multimodal Integrated-Language Framework developed by Liao and colleagues (2021).*



Note. Credit for the picture: Kruger & Liao (2022): p. 16.

The diagram unifies the models of reading developed by Reichle in 2021 and *multimedia learning* (Mayer, 2014) to describe how viewers process and understand multimodal input. Specifically:

- The *Über-Reader model* (Reichle, 2021) is a computational model that represents "the perceptual, cognitive, and motor (eye movement) processes involved in reading" (Kruger & Liao, 2022: p. 15) and postulates that "attention is allocated in a serial manner" (ibid.: p. 16), with word recognition as the element in the text that drives the eye movements and the shifts of attention during reading.
- The *cognitive theory of multimedia learning* model assumes that the visual and aural inputs are processed in two different channels ("soundtrack" and "screen" in *Figure 6*; Mayer, 2001) following the structure of working memory theorized by Baddeley and Hitch (1974), the *dual-*

coding theory and the *dual-channel assumption* postulated by Paivio respectively in 1986 and 1991 (Kruger & Liao, 2022; Montero Perez, 2022). The simultaneous presence of the two types of input (audio + text or audio + image) activates the two channels, fostering learning (in the case of audiovisual translation products, vocabulary acquisition, for instance). This mechanism is based on one of the principles of *multimedia learning*, which states that when "information is repeated by means of different channels (e.g., words and images versus words alone), readers and learners retain better information" (Chan *et al*., 2019: p. 243) (Kruger & Liao, 2022).

The *Multimodal Integrated-Language Framework* incorporates both serial and parallel processing (respectively represented by the light grey and dark grey boxes in *Figure 6*) to explain the interactive mechanisms behind the processing of multimodal input (Kruger & Liao, 2022). It also highlights the limitations of our cognitive system regarding the amount of information that can be processed in each channel, particularly in working memory (*limited capacity assumption*: Baddeley, 2012).

2.5 Preliminary conclusions

Speech processing and comprehension are complex cognitive tasks that listeners carry out every day, mostly in sub-optimal conditions (Van Engen & Peelle, 2014). Both the natural variability among speakers and the acoustic conditions of the environment determine the load imposed on our cognitive system, both in L1 and L2 speakers (Grey *et al.*, 2018). One way to support these processes is to use audiovisual translation products (captions, intralingual, and interlingual subtitles). Current research has highlighted the benefits of providing different types of audiovisual translation products to diverse users (e.g., Gernsbacher, 2015). Specifically, L2 learners can use this tool to support speech processing (perception and segmentation), improve content comprehension, and enhance their language knowledge (e.g., Montero Perez et al., 2013; Montero Perez, 2022). However, research still needs to clarify why these mixed results come up and what role individual characteristics (e.g., proficiency, working memory capacity, attitudes) play in the magnitude of the positive effects on speech processing, comprehension, and learning. Moreover, it is still open to debate whether bimodal (spoken audio + written text) and multimodal (spoken audio + written text + images) inputs bring more benefits or drawbacks to L1 and L2 speakers, especially in educational settings (e.g., Mayer *et al*., 2014; Chan *et al*., 2022). In this context, research on the effects of ASR-generated captions on these processes is still scarce and needs to be expanded since it is not clear the impact of accuracy

and transcription errors on speech processing, comprehension, and overall cognitive processes (especially attention) (Chan *et al*., 2019; Van Gauwbergen *et al*., 2024).

In sum, this chapter provided a theoretical overview of the psycholinguistics processes involved in speech processing and comprehension in L1 and L2 speakers. It also described the benefits and potential drawbacks of using audiovisual translation products in various settings, focusing on the use of automatic captions in educational contexts.

3 How L2 speakers of English use multimodal input

An investigation into the viewing habits and the use of audiovisual translation products to aid speech processing and comprehension in second-language English speakers

3.1 Introduction

This chapter reports on the first part of the research project. We present a study on the viewing habits of audiovisual products in English by Italian university students enrolled in various degree programs, speakers of English as an L2. This study also examines the students' use of audiovisual translation products (captions, intralingual, and interlingual subtitles) to support speech processing and content comprehension of their L2. Finally, it investigates their opinions on the potential use of live automatic captions in educational settings, along with their preference for the characteristics of the text of captions.

It is currently unclear whether Italian viewers, who can also be L2 speakers of English, exhibit similar viewing habits for audiovisual products in English as their American counterparts. If they do, it is important to understand why and how they utilize audiovisual translation products as L2 speakers of English. Moreover, while previous research has uncovered the benefits of intralingual and interlingual subtitles on L2 learning and comprehension (e.g., Gernsbacher, 2015), research on the impact of automatic captions on speech processing and comprehension in educational settings is scarce and still needs investigation (Cao *et al*., 2018; Chan *et al*., 2019). This study also marks the first step in developing markups that graphically show the confidence levels of the ASR system tested in a subsequent study (see Chapters 4 and 5).

Before moving to discuss the first of the four studies, we want to clarify that from now on (specifically, only in the chapters reporting the studies) we will use the umbrella term "supporting

written content" to refer to captions, intralingual (or same language) subtitles, and interlingual (translated) subtitles as a group of textual aids that viewers can turn on while watching audiovisual products (e.g., videos on streaming platforms, movies, etcetera). This term was included in the questionnaires presented to participants so as to simplify what could be a highly technical term, such as “audiovisual translation products”. We also want to clarify that in the framework of this research project we will use the term “automatic captions” to refer to the type of written supporting content used in our studies. We opted to use this term because the text produced by the partner company's ASR system offers a written transcription of a speaker's speech. Additionally, the display format used to display the text on screen in the third study follows the traditional style for captions, featuring white text on a black background to enhance readability (see Chapter 5).

3.2 Design, objectives and research questions

We designed a study to investigate Italian university students' viewing habits for audiovisual products in English outside of the classroom, including the habit of using different types of supporting written content (captions, intralingual, and interlingual subtitles). We also asked them to share their opinions on using live, automatically generated captions for lectures conducted in English at the University. Additionally, we inquired about their preferences regarding the display formats of these captions. Specifically, we aimed to address the following research questions:

RQ1. Do students regularly watch audiovisual products in English and use supporting written content to aid speech processing and content comprehension of the L2 in their day-to-day lives?

RQ2. Would they be inclined to use live automatic captions in class during lectures in English?

RQ3. Do they think it is helpful to have the text display transcription errors/the confidence level via a color-coded markup?

3.3 Methods

3.3.1 Participants

Forty-two university students, who were all native speakers of Italian (34 F; M_{age} = 24yo, SD = 6,55; age range: 18-57 years old) took part in the study. Twenty-four participants were enrolled in degree programs ranging from Archeology to Medicine (labeled as $Students_{Uni}$ in the Table below) (*N* = 24, 57,1%) and 18 participants were students enrolled in a degree program in *Foreign Languages and Cultures* (labeled as $Students_{ENG\_L2}$ in the Table below) (*N* = 18, 42,9%). All of the 18 participants were studying English as content area. *Table 1* provides a summary of participants' demographic data based on the course they are enrolled in at the university.

Table 1. *Participants' demographic data.*

Group	*N*	Gender	M_{age}	SD_{age}	$Range_{age}$	Education
$Students_{ENG\_L2}$	18	15 F; 3 M	20	2,9	18-23	HS: 18
$Students_{Uni}$	24	19 F; 5 M	26	7,6	21-57	HS: 8 BA: 10 MA: 6

Note. Labels in the *Education* column: HS: High School; BA: Bachelor's Degree; MA: Master's Degree.

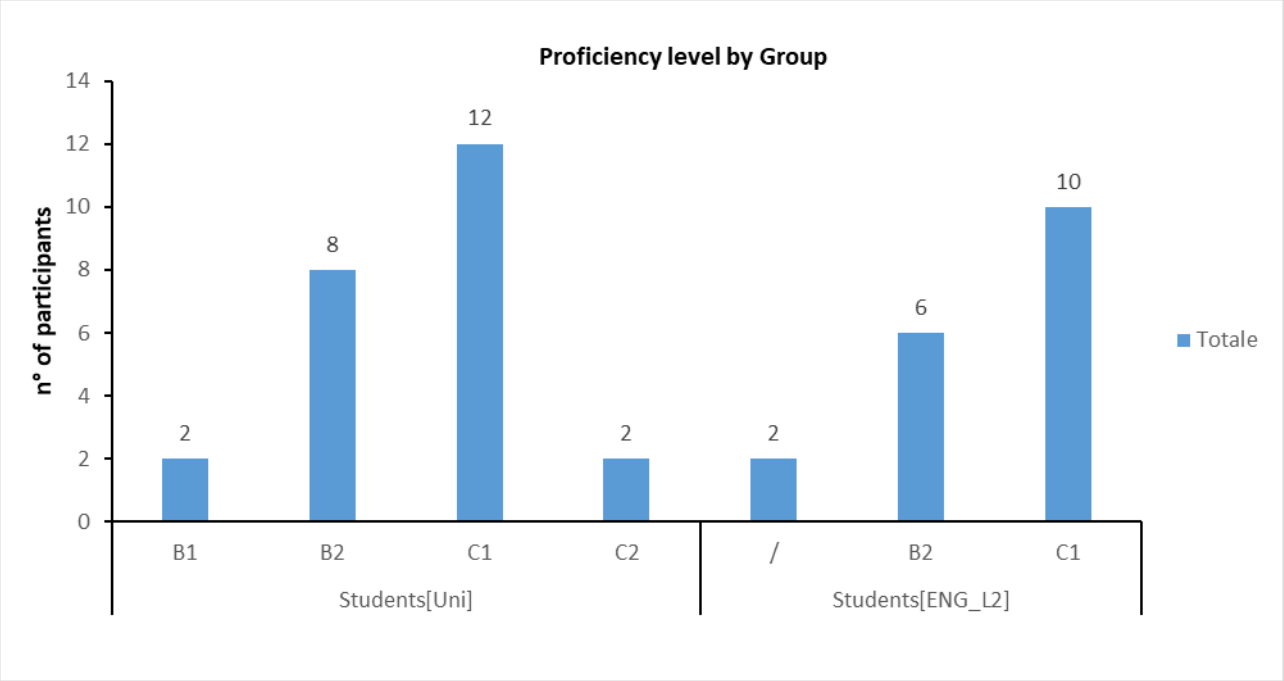
We asked students to self-report their proficiency level of English using the levels of the Common European Framework of Reference for Languages (CEFR[32]) scale. *Figure 7* shows the distribution of the self-rated proficiency level of participants by group. Half participants from the $Students_{Uni}$ group (*N* = 12, 50%) attested that their proficiency level was C1, 33,3% (*N* = 8) B2, while the rest of participants declared that their proficiency was attested respectively at B1 (*N* = 2, 8,3%) and C2 (*N* = 2, 8,3%) levels. More than half participants in the $Students_{ENG\_L2}$ group self-reported to have a C1 level of proficiency in English (55,5%) and six participants declared a B2 level (33,3%). Two participants in this group did not declare their proficiency level.

[32] Council of Europe: Common European Framework of Reference for Languages (CEFR) (https://www.coe.int/en/web/common-european-framework-reference-languages/uses-and-objectives).

Overall, the majority of participants were advanced speakers of English as a foreign language (CEFR levels: C1-C2; *N* = 24, 57,1%), while the remaining ones were intermediate speakers of English (CEFR levels: B1-B2; *N* = 16, 38,1%).

All students provided their written consent prior to participating in the study.

Figure 7. *Distribution of participants' level of proficiency of English reported by group (Students$_{Uni}$: university students enrolled in various degree programs, Students$_{ENG_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).*



3.3.2 Materials and Procedure

Prior to participating in the experimental activities, all participants signed an informed consent form.

Participants were then asked to complete three activities:

1. A task to measure their listening comprehension abilities.
2. A questionnaire investigating their habits of using supporting written content (captions, intralingual and interlingual subtitles) while watching audiovisual products in English.

3. A questionnaire investigating their opinions on the potential use and their characteristics of automatically generated live captions to be shown in class during lectures in English.

First, we asked participants to complete the listening comprehension task (MTELP) via *Pavlovia*[33]. The listening comprehension task was taken from the Michigan Test of Language Proficiency (MTELP), a standardized test that measure the proficiency level of English in L2 speakers. The task consists of listening to a set of 45 multiple choice questions and choosing the correct answer among three options. The minimum attainable score is 0, while the maximum score is 45 (one point for each correct answer).

We then asked students to complete a questionnaire to investigate their viewing habits, aiming to reveal the frequency of their exposure to the English language. Some questions investigated students' preferred type of supporting written content (captions, intralingual, and interlingual subtitles) while watching audiovisual products in English. Lastly, some questions were aimed at revealing whether captions/subtitles were used solely to enhance content comprehension or to also assist in speech processing (perception, segmentation) and learning. This questionnaire included eleven multiple-choice questions. We included the definitions of "captions" and (intralingual and interlingual) "subtitles" in Question n° 6 of the questionnaire to make sure that participants knew the meaning of the terms (see *Appendix A.I*).

Finally, students were asked to complete a questionnaire to collect their opinions on the potential use of live automatic captions in class. This questionnaire included eight multiple-choice questions (*Appendix B.I*). Some of the questions were aimed at investigating if students would find it helpful to have live automatic captions in class to support speech processing and content comprehension during lectures delivered in English. We also asked them about their perception of the impact of errors in the transcription on comprehension. Lastly, we asked them if they would find it helpful if the text of the live automatic captions would display errors or the level of confidence the ASR system has in its transcription using a color-coded markup.

Data was collected using two slightly different versions of the questionnaires on viewing habits and on the potential use of live automatic captions in class (see *Appendices A.I, A.II* and *Appendices B.I, B.II*). The first version of both questionnaires (A.I and B.I) was used to collect data from the students enrolled in the *Foreign Languages and Cultures* degree program, while versions A.II and B.II were

[33] https://www.pavlovia.org

used to collect data from the Students$_{Uni}$ group. *Table 2* gathers all the questions and options that differ in the two versions of the questionnaire.

Table 2. *List of questions and answers that differ in the two versions of the questionnaires employed in this study.*

Questionnaire on viewing habits		
Question n°	**First version (appendix A.I)**	**Second version (appendix A.II)**
6[34]	***Question*** Do you prefer using captions or subtitles? ***Options*** a) Captions b) Subtitles c) I don't use captions/subtitles	***Question*** Which type of supporting written content (captions, subtitles, etc.) do you prefer using when watching audiovisual content in English? ***Options*** a) Captions (written transcript of the audio) b) Subtitles (same language of the audio) c) Subtitles (translation of the audio in your native language) d) It depends e) I don't use captions/subtitles f) Other answer
7	-	***Question*** Why do you prefer that type of supporting written content (captions, subtitles, etc.)? Could you please motivate your previous answer? You can choose more than one answer. ***Options*** a) I want to know what's the translation in my native language of the words in the speech. b) I want to know what's the meaning of a word in my native language. c) I prefer reading each word as soon as it is pronounced by speakers. d) I prefer reading the text in my native language. e) It requires less effort to read the text in my native language.
8	***Question*** How often do you use captions/subtitles when watching audiovisual content in English? ***Options*** a) Never b) Rarely c) A few times a week d) Every day	***Question*** How often do you use these formats when watching audiovisual content in English? *[one answer for each type of written supporting content, i.e. captions, intralingual subtitles, interlingual subtitles]* ***Options*** a) Never b) Rarely c) Frequently d) Always

[34] In addition to this question, in the first version of this questionnaire we asked participants "In which language do you prefer subtitles/captions?". To answer this question, they had to select one option among the following: a) I don't use captions/subtitles; b) In my native language (subtitles); c) In the same language of the oral speech (captions); d) In the same language of the oral speech (subtitles).

Questionnaire on the potential use of live automatic captions during lectures		
Question n°	**First version (appendix A.I)**	**Second version (appendix A.II)**
7	***Question*** In your opinion, would it be helpful if a specific display format signaled transcription errors? ***Options*** a) Yes b) No c) I'm not sure	***Question*** In your opinion, would it be helpful if a specific display format signaled how confident the system was in its transcription? ***Options*** a) Yes b) No c) I'm not sure

Note. For Question n° 7 (questionnaire on viewing habits), we report only the options that were not included in the first version of the questionnaire (see column "Second version").

While the first version of the questionnaires was uploaded to *Qualtrics XM*, the second version was delivered via *Google Forms*. The first version of the questionnaires was used to collect data from the students enrolled in the *Foreign Languages and Cultures* degree program, while the second version was used to collect data from all the other participants.

The activities included in this study were approved by the Ethics Committee of Ca' Foscari University, Venice.

3.4 Results

In this section, I present the descriptive statistics for the listening comprehension task (MTELP) and the two questionnaires. The results from the collected data will be discussed one question at a time and separately for the two groups ($Students_{ENG\_L2}$ and $Students_{Uni}$), because participants may critically differ on many levels (e.g., language use, learning strategies), but also considering the differences in how some questions were formulated in the two versions of the questionnaires (for details, see the section *Materials*, §3.3.2; see also the section *Limitations*, §3.6).

3.4.1 Listening Comprehension Task (MTELP)

Scoring for the listening comprehension task was assessed automatically in *PsychoPy* by assigning the value 1 to the correct answer and assigning the value 0 to the wrong answers for each trial. *Table*

2 summarizes the descriptive statistics for participants' scores in the listening comprehension task, while *Figure 8* shows the distribution of scores.

Students enrolled at the *Foreign Languages and Cultures* degree program ($Students_{ENG\_L2}$) had an average MTELP score of 42,83 (SD = 2,01), while the university students enrolled in various degree programs ($Students_{Uni}$) had an average MTELP score of 42,71 (SD = 3,33). Taking a closer look at the two groups, *Figure 8* shows the distribution of scores from the listening comprehension task for each group of participants, showing that the distribution is skewed towards the highest values (≥ 40) in both groups (see also *Table 3*). While the range of scores is more homogeneous in the group of students enrolled at the *Foreign Languages and Cultures* degree program ($Students_{ENG\_L2}$) (range: 40-45), the group of students enrolled at university ($Students_{Uni}$) have a wider range (30-45). We then compared the mean MTELP scores for the two groups using the Mann-Whitney U test to check if there was a difference in their listening comprehension abilities. Results showed that the difference in the mean MTELP scores was not significant (Mann–Whitney $U = 231$, $p = 0.706$), therefore participants in the two groups did not differ in their listening abilities.

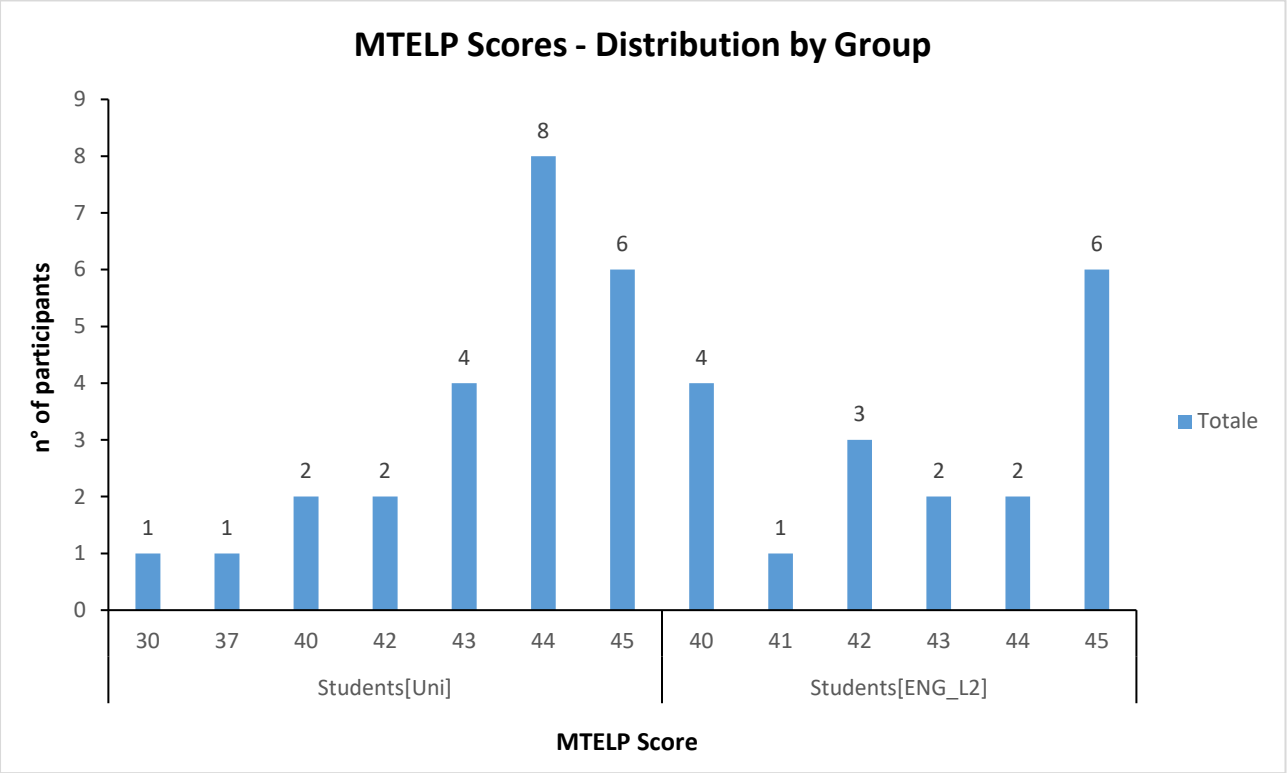
These results suggest that most participants possess very good listening skills, regardless of the degree program they are enrolled in.

Table 3. *Descriptive statistics for participants' scores in the listening comprehension task reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).*

	N	M_{MTELP}	SD_{MTELP}	$Median_{MTELP}$	$Range_{MTELP}$	Skewness
$Students_{ENG\_L2}$	18	42,83	2,01	43	40-45	-0.284
$Students_{Uni}$	24	42,71	3,33	44	30-45	-2.814

Note. The listening skills were measured with a portion of the Michigan Test of Language Proficiency (MTELP).

Figure 8. *Distribution of the MTELP scores for all participants reported by group (Students$_{Uni}$: university students enrolled in various degree programs, Students$_{ENG_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).*



3.4.2 Questionnaire on participants' viewing habits and use of supporting written content while watching audiovisual products in English

Q1. *Do you watch audiovisual content in English at home, university, etc.?*

In their everyday life, almost every participant confirmed that they watch audiovisual content in English (*N* = 39, 92,9%), except three participants from the Students$_{Uni}$ group (7,1%) (*Table 4*).

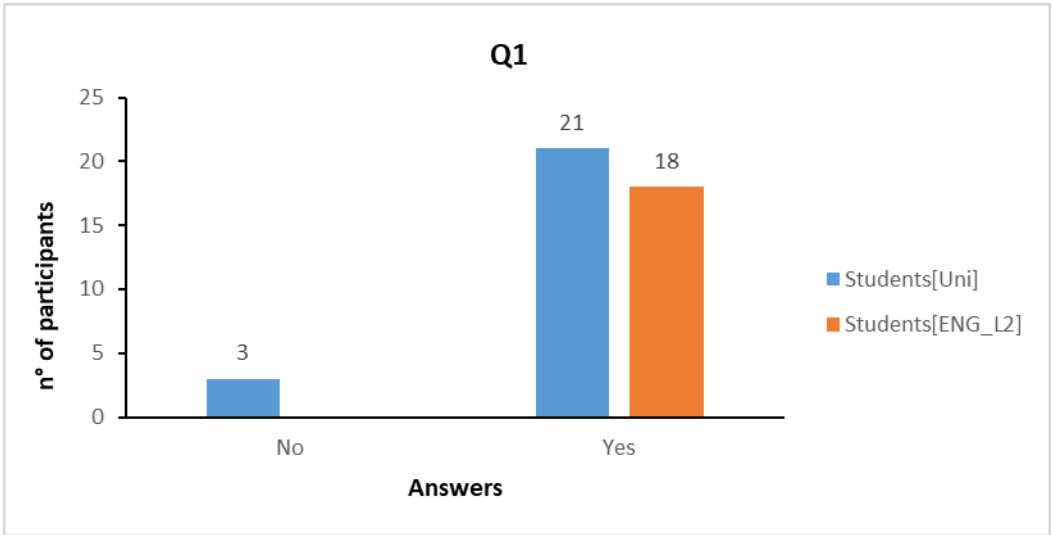
Table 4. *Number (count, percentage) of answers for Question n° 1 in the questionnaire on viewing habits and supporting written content use.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
No	-	3 (12,5%)	3 (7,1%)
Yes	18 (100%)	21 (87,5%)	39 (92,9%)

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Figure 9 shows the distributions of answers for Q1 by group. While all participants (N = 18) in the group of students enrolled in the *Foreign Languages* course ($Students_{ENG\_L2}$ – orange bar) stated that they watch audiovisual products in English, twenty-one participants (87,5%) in the group of students enrolled in various courses at the university ($Students_{Uni}$ – dark blue bar) stated the same. Only three participants from this group selected "No" as their answer, stating that they don't watch audiovisual products in English.

Figure 9. *Distribution of answers for Question n° 1 in the questionnaire on viewing habits and supporting written content use by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q2. *How often do you watch audiovisual content in English in your day-to-day life?*

The majority of participants watch audiovisual content in English every day (N = 30, 71,4%) or a few times a week (N = 10, 23,8%) (*Table 5*). Only two participants from the $Students_{Uni}$ group answered that they rarely watch audiovisual content in English.

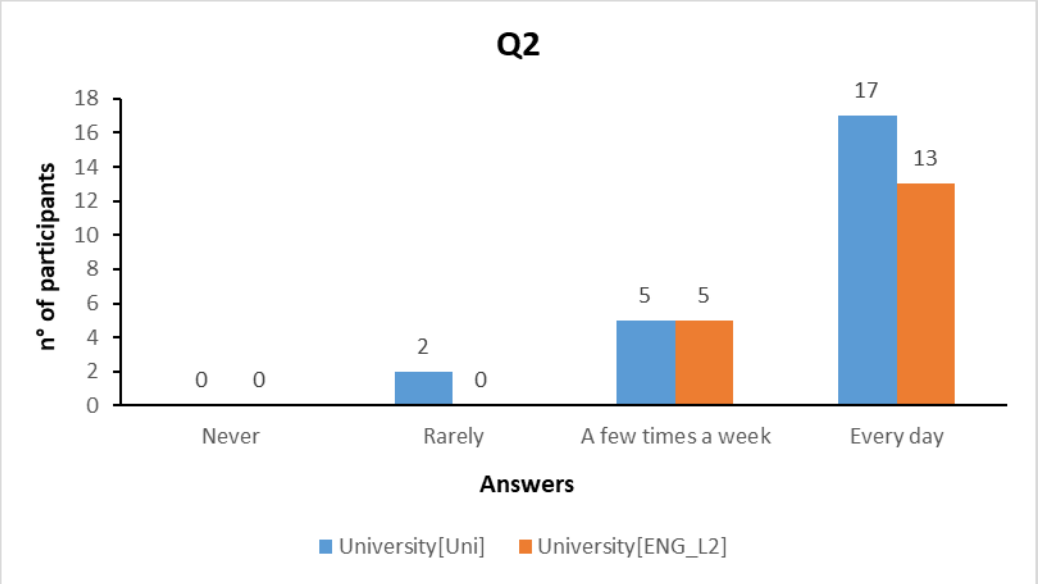
Table 5. *Number (count, percentage) of answers for Question n° 2 in the questionnaire on viewing habits and supporting written content use reported by group.*

Answer	**$Students_{ENG\_L2}$**	**$Students_{Uni}$**	**Total**
Never	-	-	-
Rarely	-	2 (8,3%)	2 (4,8%)
A few times a week	5 (27,8%)	5 (20,8%)	10 (23,8%)
Every day	13 (72,2%)	17 (70,8%)	30 (71,4%)

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

The trend of answers is similar for both groups of participants (*Figure 10*). Most participants from the two groups watch audiovisual products in English every day ($Students_{ENG\_L2}$, N = 13, 72,2%; $Students_{Uni}$, N = 17, 70,8%). Five participants per group ($Students_{ENG\_L2}$, 27,8%; $Students_{Uni}$, 20,8%) watch audiovisual content in English a few times a week, while only two participants from the $Students_{Uni}$ group (8,3%) rarely do so.

Figure 10. *Distribution of answers for Question n° 2 in the questionnaire on viewing habits and supporting written content use reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q3. *What kind of audiovisual content in English do you watch? [More than one choice per participant]*

Almost all participants watch TV series (*N* = 37, 88,1%), movies (*N* = 31, 73,8%), short clips on various video platforms (*N* = 36, 85,7%) and vloggers on YouTube (*N* = 29, 69%). Almost half of the participants also watch news in English on the TV/on the web (*N* = 20, 47,6%). The least watched audiovisual content in English are documentaries (*N* = 13, 30,9%), online lectures (*N* = 14, 33,3%), and learning videos from textbooks (*N* = 4, 9,5%) (*Table 6*).

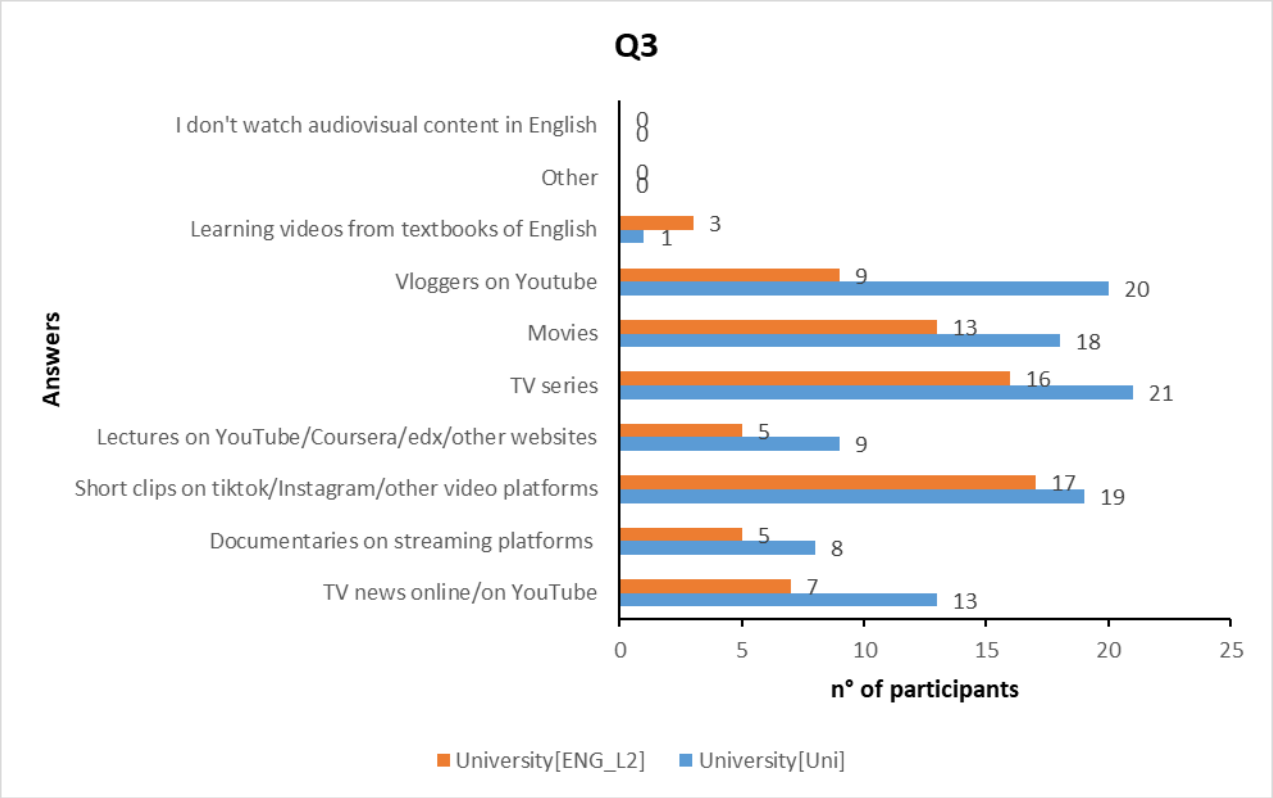
Table 6. *Number (count, percentage) of answers for Question n° 3 in the questionnaire on viewing habits and supporting written content use reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
News on TV/on the web	7 (38,9%)	13 (54,2%)	20 (47,6%)
Documentaries on streaming platforms	5 (27,8%)	8 (33,3%)	13 (30,9%)
Short clips on video platforms	17 (94,4%)	19 (79,2%)	36 (85,7%)
Online lectures	5 (27,8%)	9 (37,5%)	14 (33,3%)
TV series	16 (88,9%)	21 (87,5%)	37 (88,1%)
Movies	13 (72,2%)	18 (75%)	31 (73,8%)
Vloggers	9 (50%)	20 (83,3%)	29 (69%)
Learning videos from textbooks	3 (16,7%)	1 (4,2%)	4 (9,5%)
Other	-	-	-
I don't watch audiovisual content in English	-	-	-

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Figure 11 shows the distributions of answers for each group (orange bars: $Students_{ENG\_L2}$; dark blue bars: $Students_{Uni}$). Most participants from both groups watch movies ($Students_{ENG\_L2}$, $N = 15$, 83,3%; $Students_{Uni}$, $N = 21$, 87,5%), TV series ($Students_{ENG\_L2}$, $N = 15$, 83,3%; $Students_{Uni}$, $N = 21$, 87,5%), and short clips on various platforms on the web ($Students_{ENG\_L2}$, $N = 15$, 83,3%; $Students_{Uni}$, $N = 21$, 87,5%). A higher number of participants from the $Students_{Uni}$ group ($N = 20$, 83,3%) compared to the number of participants from the $Students_{ENG\_L2}$ group ($N = 9$, 50%) watches vloggers on web platforms. The same trend is evident in the selection of the "news on the web/TV" ($Students_{Uni}$, $N = 13$, 54,2%; $Students_{ENG\_L2}$, $N = 7$, 38,9%), "documentaries on streaming platforms" ($Students_{Uni}$, $N = 8$, 33,3%; $Students_{ENG\_L2}$, $N = 5$, 27,8%) and "lectures online" ($Students_{Uni}$, $N = 9$, 37,5%; $Students_{ENG\_L2}$, $N = 5$, 27,8%). Lastly, more participants from the $Students_{ENG\_L2}$ group ($N = 3$, 16,7%) usually watch learning videos from textbooks compared to the $Students_{Uni}$ group ($N = 1$, 4,2%).

Figure 11. *Distribution of answers for Question n° 3 in the questionnaire on viewing habits and supporting written content use reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q4. *When you watch audiovisual content in English, you mostly watch it (on which device?) [More than one choice per participant]*

Most participant watch audiovisual products in English on their smartphone (*N* = 36, 85,7%) and laptops (*N* = 21, 50%), while a small percentage use tablets (*N* = 11, 26,2%) and desktop computers (*N* = 12, 28,6%) (*Table 7*).

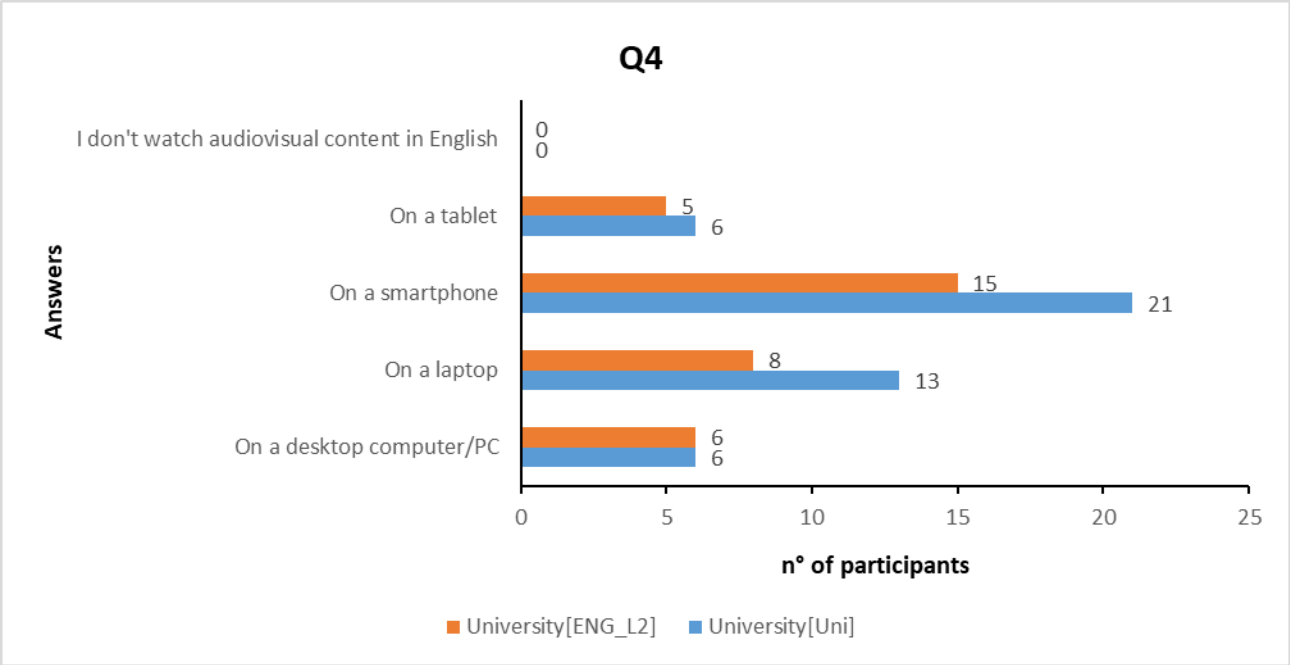
Table 7. *Number (count, percentage) of answers for Question n° 4 in the questionnaire on viewing habits and supporting written content use reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
On a tablet	6 (33,3%)	6 (25%)	12 (28,6%)
On a smartphone	15 (83,3%)	21 (87,5%)	36 (85,7%)
On a laptop	8 (44,4%)	13 (54,2%)	21 (50%)
On a desktop computer/PC	5 (27,8%)	6 (25%)	11 (26,2%)
I don't watch audiovisual content in English	-	-	-

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Most participants from both groups use their smartphones to watch audiovisual products in English ($Students_{ENG\_L2}$, $N = 15$, 83,3%; $Students_{Uni}$, $N = 21$, 87,5%), as well as their laptops ($Students_{ENG\_L2}$, $N = 8$, 44,4%; $Students_{Uni}$, $N = 13$, 54,2%) (*Figure 12*). A small number of participants in both groups, on the other hand, use their tablets ($Students_{ENG\_L2}$, $N = 6$, 33,3%; $Students_{Uni}$, $N = 6$, 25%) and desktop PCs ($Students_{ENG\_L2}$, $N = 5$, 27,8%; $Students_{Uni}$, $N = 6$, 25%) to watch audiovisual content in English.

Figure 12. *Distribution of answers for Question n° 4 in the questionnaire on viewing habits and supporting written content use reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q5. *Do you use supporting written content (captions, subtitles, etc.) when watching videos in English?*

Table 8 shows that the majority of participants currently use supporting written content (*N* = 33, 78,6%) when watching audiovisual products in English, with only 21,4% of them (*N* = 9) not using captions or intra/interlingual subtitles.

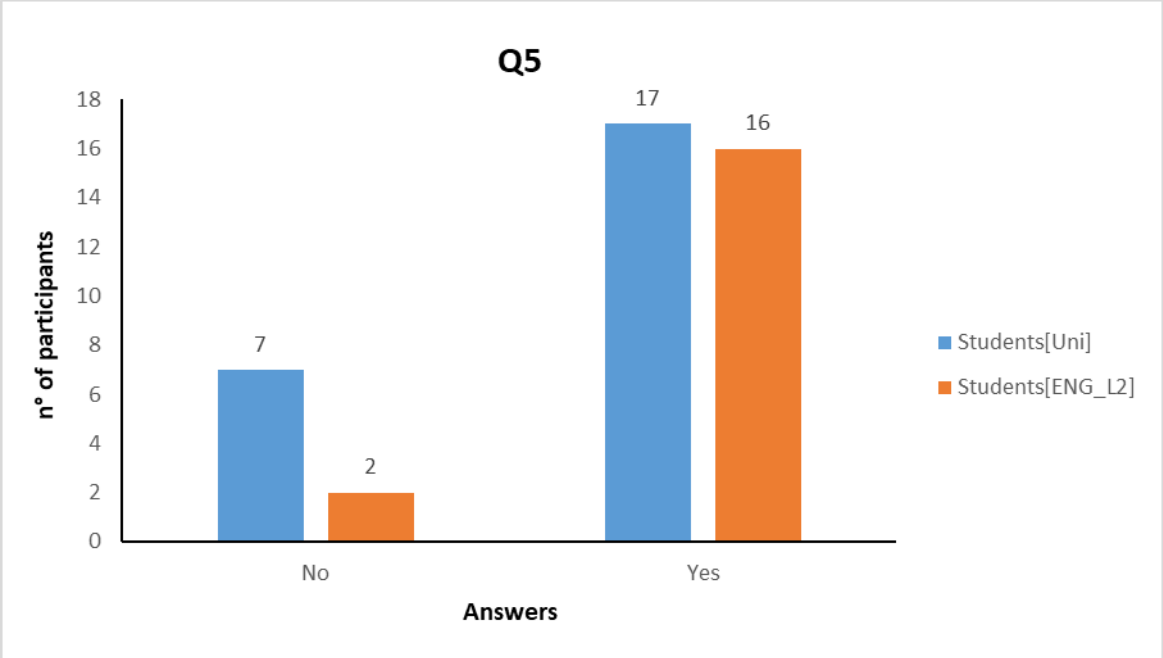
Table 8. *Number (count, percentage) of answers for Question n° 5 in the questionnaire on viewing habits and supporting written content use reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
No	2 (11,1%)	7 (29,2%)	9 (21,4%)
Yes	16 (88,9%)	17 (70,8%)	33 (78,6%)

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

The numbers are similar in both groups (*Figure 13*), where the majority of participants from both groups use supporting written content when watching audiovisual content in English ($Students_{ENG\_L2}$, *N* = 16, 88,9%; $Students_{Uni}$, *N* = 17, 70,8%). A higher number of participants from the $Students_{Uni}$ group (*N* = 7, 29,2%) compared to the number of participants from the $Students_{ENG\_L2}$ group (*N* = 2, 11,1%) do not use captions nor interlingual/intralingual subtitles when watching audiovisual content in English.

Figure 13. *Distribution of answers for Question n° 5 in the questionnaire on viewing habits and supporting written content use reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q6. *Which supporting written content (captions, subtitles, etc.) do you prefer using when watching videos in English?*

Overall, half participants from the sample preferred captions (*N* = 21, 50%), while the other half preferred using subtitles (*N* = 21, 50%). Only a small number of participants (N = 7, 16,7%) stated that they do not use supporting written content when watching audiovisual products in English. Three participants (7,1%) stated that they use supporting written content based on the circumstances. *Table 9* summarizes the data.

Table 9. *Number (count, percentage) of answers for Question n° 6 in the questionnaire on viewing habits and supporting written content use reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
Captions	14 (77,8%)	7 (29,2%)	21 (50%)
Interlingual subtitles	1 (5,6%)	7 (29,2%)	8 (19%)
Intralingual subtitles	2 (11,1%)	11 (45,8%)	13 (31%)
It depends (on some factors)	-	3 (12,5%)	3 (7,1%)
I don't use captions/subtitles	1 (5,6%)	6 (25%)	7 (16,7%)
Other	-	2 (8,3%)	2 (4,8%)

Note 1. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Note 2. The column $Students_{ENG\_L2}$ in this table contains the answers to two questions. In the first version of the questionnaire (used to collect data from the $Students_{ENG\_L2}$ group), Question n° 6 included three options: "captions", "subtitles", and "I don't use captions/subtitles". Then, in Question n° 9, we asked participants to specify in which language they preferred reading the text. The options for this question were: "In my native language (subtitles)", "In the same language of the oral speech (captions)", "In the same language of the oral speech (subtitles)", "I do not use captions/subtitles". The second version of the questionnaire instead did not include Question n° 9, while Question n° 6 included six options: "captions (written transcript of the audio)", "subtitles (same language of the audio)", "subtitles (translation of the audio in your native language)", "It depends", "Other (please, specify it)", and "I don't use captions/subtitles". For details, see *Table 2* (section 3.3.2) and *Appendices A.I* and *A.II*.

Figure 14 reports the distributions of answers by group.

Regarding the $Students_{ENG\_L2}$ group (orange bars in the graph), the majority of participants answered that they use captions (N = 14, 77,8%), two participants stated that they use intralingual subtitles (11,1%), one participant stated that they use interlingual subtitles (5,6%), and only one participant said that they do not use captions nor subtitles (5,6%).

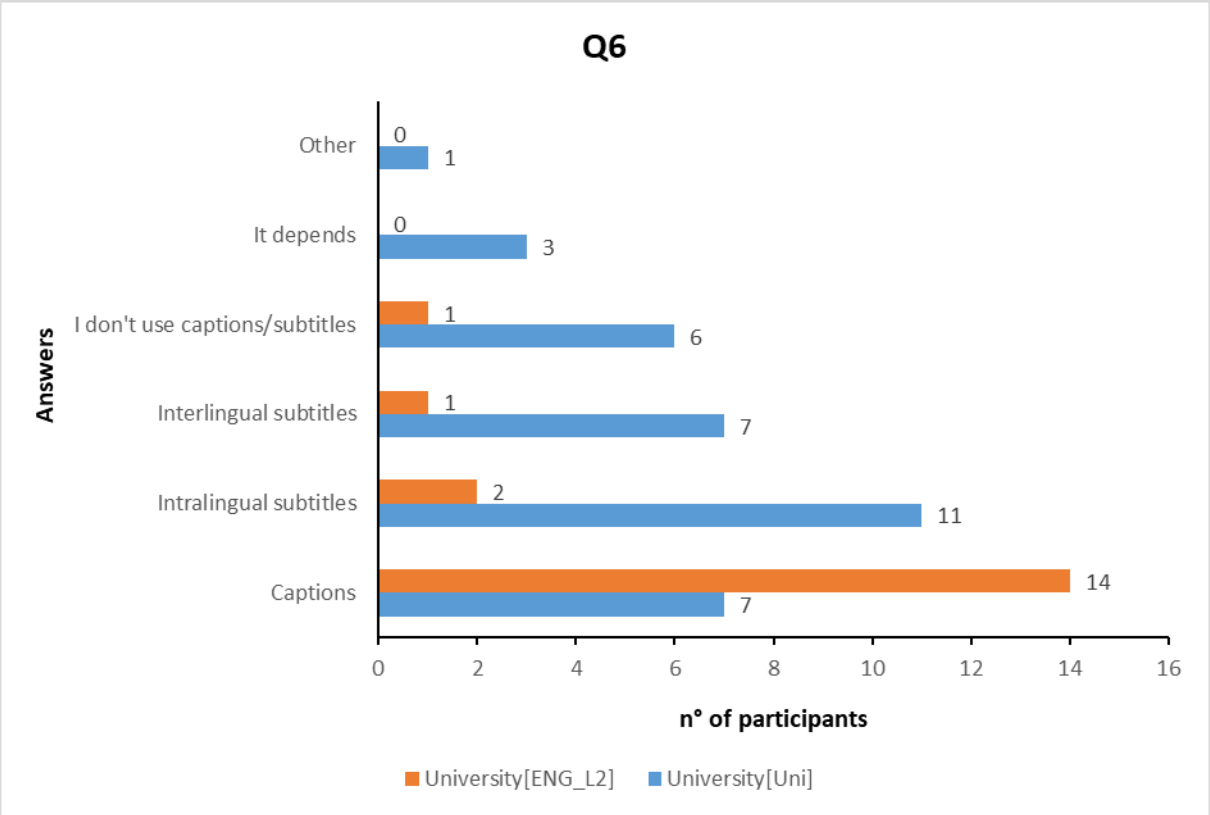
Participants from the $Students_{Uni}$ group (dark blue bars in the graph) mostly use intralingual subtitles (N = 11, 45,8%). One third of them use captions and interlingual subtitles (N = 7, 29,2%). Six participants stated that they do not use any supporting written content (25%), while three participants (12,5%) stated that they use supporting written content based on the circumstances. One participant from the $Students_{Uni}$ group declared the following:

It depends on the context: videos and TikTok captions are fine but I prefer subtitles in the same language of the audio with tv series and movies (I think it depends on the seriousness of the content I'm watching).

Another participant from the same group specified that they do not use any supporting written content, but they only turn on captions for specific reasons:

I don't use captions/subtitles, I only use captions if the spoken word is not clear due to accents or fast pronunciations.

Figure 14. *Distribution of answers for Question n° 6 in the questionnaire on viewing habits and supporting written content use reported by group (Students$_{Uni}$: university students enrolled in various degree programs, Students$_{ENG_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).*



Note. Results are reported by group (Students$_{Uni}$: university students enrolled in various degree programs – blue bars; Students$_{ENG_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q7. *Could you please motivate your previous answer? [More than one choice per participant]*

In general, participants from both groups showed a similar pattern of answers (*Table 10*). Participants tend to agree that they use the supporting written content of their choice mainly because they want to make sure they understand the content of what they are watching (*option H*: $N = 28$, 66,7%; Students$_{ENG_L2}$, $N = 14$, 77,8%; Students$_{Uni}$, $N = 14$, 58,3%), but also because they want to learn new words (*option F*: $N = 22$, 52,4%; Students$_{ENG_L2}$, $N = 11$, 61,1%; Students$_{Uni}$, $N = 11$, 45,8%)

or learn/improve their pronunciation of words (*option G*: *N* = 21, 50%; Students$_{ENG_L2}$, *N* = 11, 61,1%; Students$_{Uni}$, *N* = 10, 41,7%). Only a small number of participants (mainly students enrolled in the *Foreign Languages and Cultures* degree program at the university - Students$_{ENG_L2}$, *N* = 4, 22,2%; Students$_{Uni}$, *N* = 2, 8,3%) said that they use supporting written content to identify where words begin and end by reading the written transcript (*option E*: *N* = 6, 14,3%).

A few participants from the Students$_{Uni}$ group stated that they do not use supporting written content because they do not need it (*option D*: *N* = 6, 14,3%). Participants that preferred not using supporting written content said that they prefer listening to the speaker(s) (*option C*: *N* = 7; Students$_{ENG_L2}$, *N* = 1, 5,6%; Students$_{Uni}$, *N* = 6, 25%), to focus on listening improve their listening skills rather than using captions/subtitles to do so (*option B*: *N* = 3, 7,1%; Students$_{ENG_L2}$, *N* = 1, 5,6%; Students$_{Uni}$, *N* = 2, 8,3%), but also because they find it hard to pay attention both to the speaker(s) talking and the text on the screen (*option A*: *N* = 3, 7,1%; Students$_{ENG_L2}$, *N* = 1, 5,6%; Students$_{Uni}$, *N* = 2, 8,3%).

Table 10. *Number (count, percentage) of answers for Question n° 7 in the questionnaire on viewing habits and supporting written content use reported by group.*

Option	Students$_{ENG_L2}$	Students$_{Uni}$	Total
A	1 (5,6%)	2 (8,3%)	3 (7,1%)
B	1 (5,6%)	2 (8,3%)	3 (7,1%)
C	1 (5,6%)	6 (25%)	7 (16,7%)
D	-	6 (25%)	6 (14,3%)
E	4 (22,2%)	2 (8,3%)	6 (14,3%)
F	11 (61,1%)	11 (45,8%)	22 (52,4%)
G	11 (61,1%)	10 (41,7%)	21 (50%)
H	14 (77,8%)	14 (58,3%)	28 (66,7%)
I	-	3 (12,5%)	3 (7,1%)
L	-	4 (16,7%)	4 (9,5%)
M	-	3 (23,1%)	3 (7,1%)

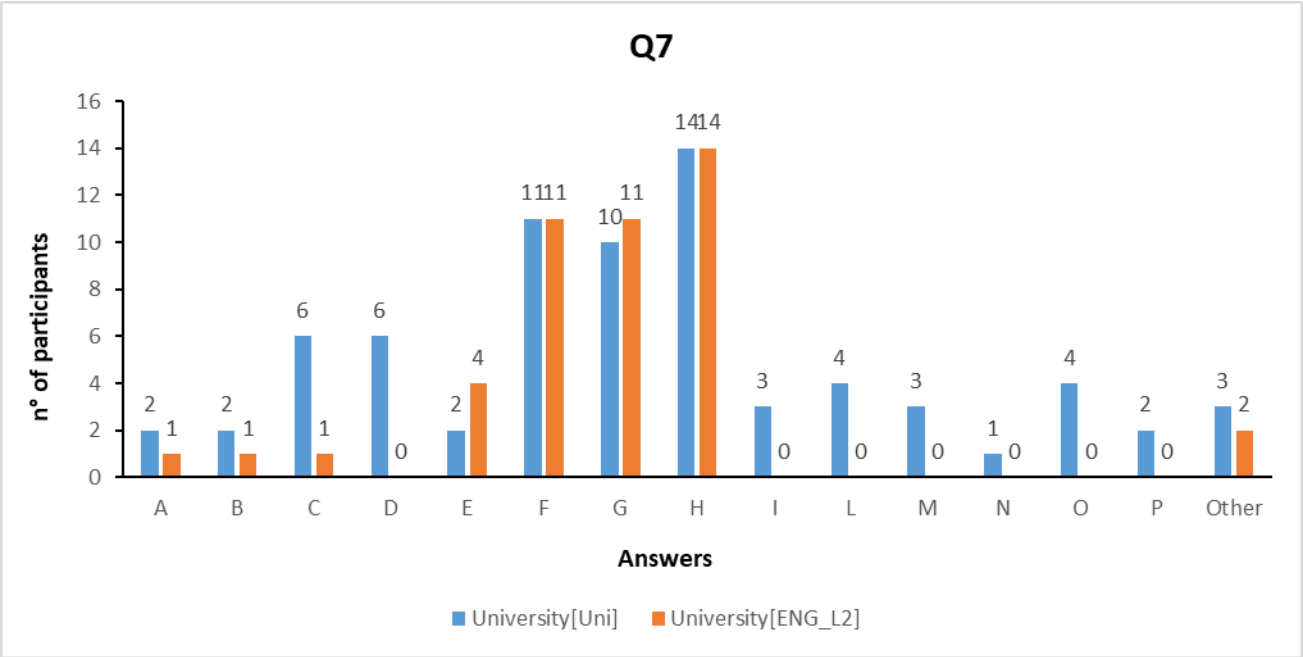
N	-	1 (4,2%)	1 (2,4%)
O	-	4 (16,7%)	4 (9,5%)
P	-	2 (8,3%)	2 (4,8%)
Other	2 (11,1%)	3 (12,5%)	5 (11,9%)

Note 1. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Note 2. Labels in the *Answer* column: *Option A*: I find it hard to concentrate if I have pay attention both to the audio and the text; *Option B*: Yes: I do not turn captions/subtitles on because I try to improve my listening skills, and I don't think that written support helps; *Option C*: I don't use captions/subtitles because I prefer listening to what is being said rather than reading the text; *Option D*: I don't use captions/subtitles because I don't need them; *Option E*: I need to identify where words begin and end by reading the written transcript; *Option F*: I want to learn new words; *Option G*: I want to improve/learn the pronunciation of words; *Option H*: I want to make sure I understand the content of what I am watching/listening; *Option I*: I want to know what's the translation in my native language of the words in the speech; *Option L*: I want to know what's the meaning of a word in my native language; *Option M*: I prefer reading each word as soon as they are pronounced by speakers; *Option N*: I prefer reading the exact words pronounced by speakers; *Option O*: I prefer reading the text in my native language; *Option P*: It requires less effort to read the text in my native language. *Options I – P* were included only in the second version of the questionnaire.

Among the participants in the $Students_{Uni}$ group, four of them stated that they use interlingual subtitles because they want to find out the meaning of a word in their L1 (*option L* - 16,7% out of the total number of participants in the group) and they prefer reading the text of the captions/subtitles in their L1 (*option O* - 16,7% out of the total number of participants in the group). Three participants stated that they want to know what's the translation in their L1 of the words in the speech (*option I* – 12,5%) and that they prefer reading each word as soon as they are pronounced by the speaker (*option M* – 12,5%). Two participants said that it requires less effort to read the text of the subtitles in their L1 (*option P* – 8,3%). Only one participant stated that they prefer reading the exact words pronounced by speakers (*option N* - 5,6%). See *Figure 15* to see the distribution of answers for this question.

Figure 15. *Distribution of answers for Question n° 7 in the questionnaire on viewing habits and supporting written content use reported by group.*



Note 1. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Note 2. Labels in the x axis: *Option A*: I find it hard to concentrate if I have pay attention both to the audio and the text; *Option B*: Yes: I do not turn captions/subtitles on because I try to improve my listening skills, and I don't think that written support helps; *Option C*: I don't use captions/subtitles because I prefer listening to what is being said rather than reading the text; *Option D*: I don't use captions/subtitles because I don't need them; *Option E*: I need to identify where words begin and end by reading the written transcript; *Option F*: I want to learn new words; *Option G*: I want to improve/learn the pronunciation of words; *Option H*: I want to make sure I understand the content of what I am watching/listening; *Option I*: I want to know what's the translation in my native language of the words in the speech; *Option L*: I want to know what's the meaning of a word in my native language; *Option M*: I prefer reading each word as soon as they are pronounced by speakers; *Option N*: I prefer reading the exact words pronounced by speakers; *Option O*: I prefer reading the text in my native language; *Option P*: It requires less effort to read the text in my native language. *Options I – P* were included only in the second version of the questionnaire.

Five participants ($Students_{ENG\_L2}$: $N = 2$; $Students_{Uni}$: $N = 3$) added some comments or wrote different answers from the ones presented in the options. One of the participants from the $Students_{ENG\_L2}$ group stated that

I do not always turn captions/subtitles on, but when I do I prefer captions because I find it confusing when the text does not correspond exactly to the audio.

Similarly, another participant from the same group reported stated

I don't like it when captions/subtitles do not match the audio.

On the other hand, one of the participants from the Students$_{Uni}$ group stated that

"(...) all the answers [they gave] are true, it just depends on the context: tiredness, dialect/pronunciation, if I'm willing to put an extra effort to understand if something is not clear, etc.".

Similarly, another participant from the same group also specified that

"I only use captions when listening to a particular dialect".

One last participant from the Students$_{Uni}$ group said that

"[I want to make sure I understand the content of what I am watching/listening][35] this meaning if there are some words I don't know or if I'm doing something else while I'm watching (like eating, which makes some noise that might make me lose some words)".

Q8. *How often do you use [supporting written content] when watching audiovisual content in English?*

This question was formulated differently in the two questionnaires. In the first questionnaire we asked participants to state how often they used captions/subtitles while watching audiovisual products in English (*Table 11*). In the second version of the questionnaire, participants were asked to answer how frequently they used each audiovisual translation product (captions, intralingual, and interlingual subtitles) in the same setting (*Table 12*). Therefore, participants from the Students$_{ENG_L2}$group and the Students$_{Uni}$ group answered two slightly different versions of the same question: for this reason, we will not present an overall analysis of the data, but we will keep the discussion of the results separate by group.

Table 11 reports the answers given by participants in the group of university students enrolled in the *Foreign Languages and Cultures* degree program. The majority of them (*N* = 10, 55,6%) answered that they use captions/subtitles a few times a week while watching audiovisual content in English. Six participants (33,3%) use supporting written content every day, while only two participants never or rarely use captions/subtitles while watching audiovisual products in English. *Figure 16* shows the distribution of answers for Question n° 8.

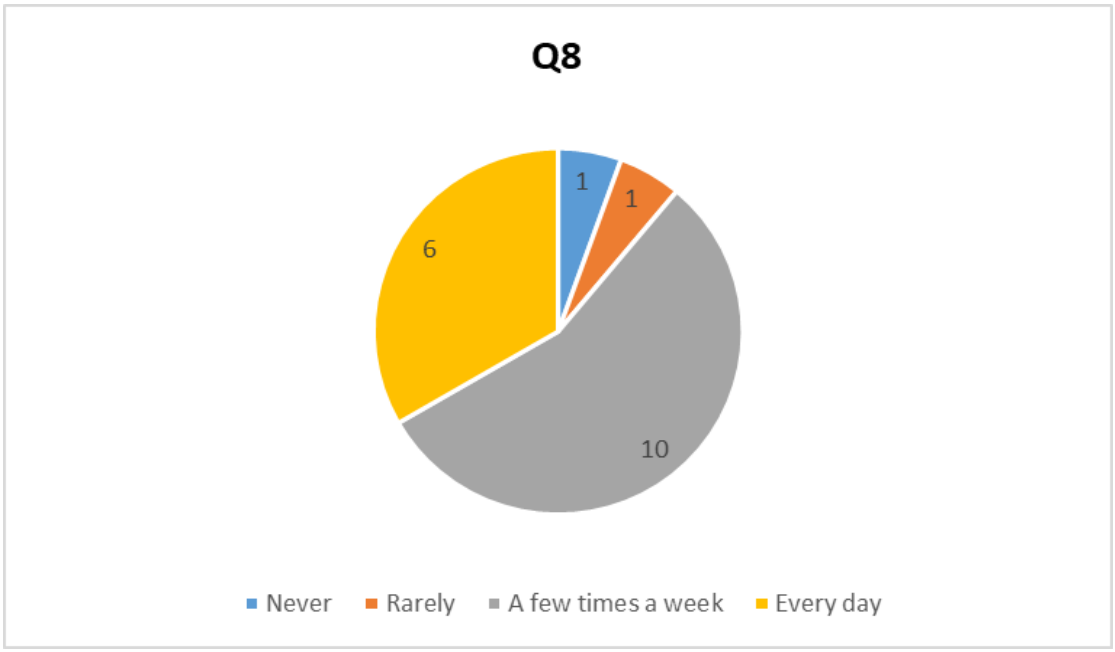
[35] The text between parentheses is the option selected by the participant and added to the text to show the complete answer.

Table 11. *Number (count, percentage) of answers for Question n° 8 in the questionnaire on viewing habits and supporting written content use reported for the group "Students$_{ENG_L2}$" (university students enrolled in a Foreign Languages and Cultures degree program).*

Answer	Total (n°)	Total (%)
Never	1	5,6%
Rarely	1	5,6%
A few times a week	10	55,6%
Every day	6	33,3%

Note. Participants had to select one of the options ("Never", "Rarely", "A few times a week", "Every day") to answer the following question: "How often do you use captions/subtitles when watching audiovisual content in English?".

Figure 16. *Distribution of answers for Question n° 8 in the questionnaire on viewing habits and supporting written content use reported for the group "Students$_{ENG_L2}$" (university students enrolled in a Foreign Languages and Cultures degree program).*



Participants from the Students$_{Uni}$ group were asked to specify how frequently they use captions, intralingual, and interlingual subtitles while watching audiovisual content in English (*Table 12*).

Most of participants from this group rarely use captions when watching audiovisual content in English (*N* = 14, 58,3%). Seven participants (29,2%) frequently use this type of supporting written content, and only three participants stated that they never (*N* = 2, 8,3%) or always (*N* = 1, 4,2%) use captions while watching audiovisual products in English.

Almost half of the participants from the $Students_{Uni}$ group frequently use intralingual subtitles (*N* = 11, 45,8%). Eight of them never use this type of supporting written content (33,3%), two participants (8,3%) rarely use it, and three participants (12,5%) always use it.

Last, but not least, eleven participants (45,8%) never use interlingual subtitles while watching audiovisual content in English. Similarly, only six participants (25%) from the $Students_{Uni}$ group rarely use this type of supporting written content. A few participants frequently (*N* = 5, 20,8%) and always (*N* = 2, 8,3%) use interlingual subtitles.

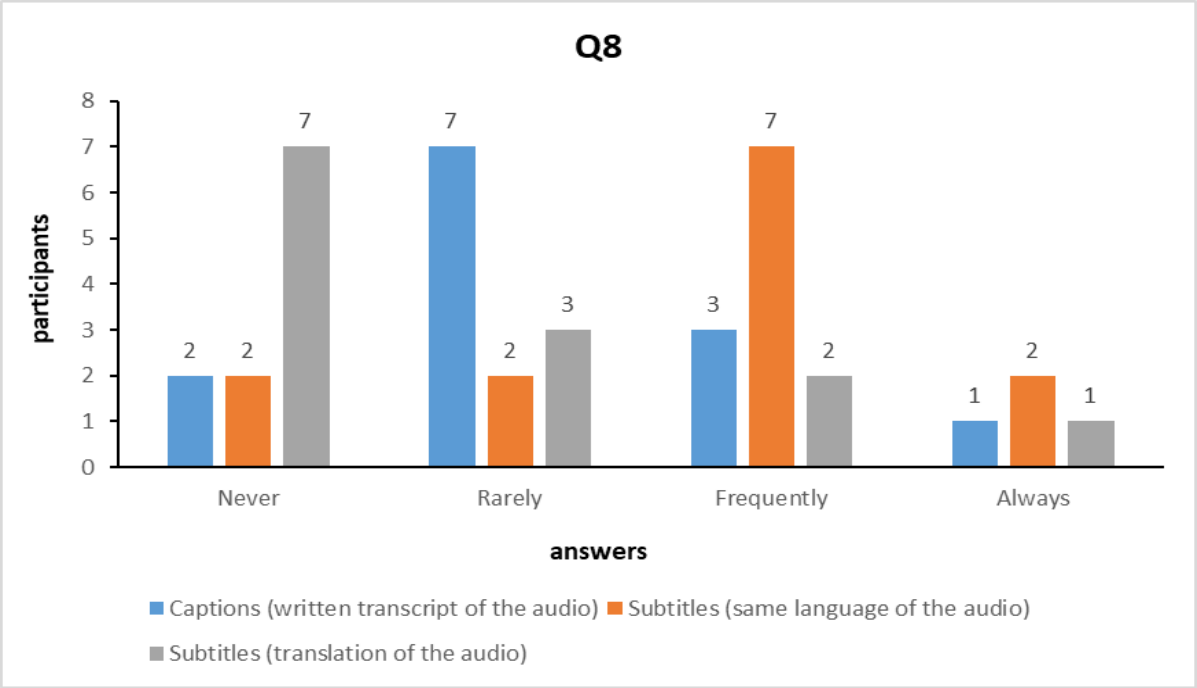
Table 12. *Number (count, percentage) of answers for Question n° 8 in the questionnaire on viewing habits and supporting written content use reported for the group "$Students_{Uni}$" university students enrolled in various degree programs.*

	Never	**Rarely**	**Frequently**	**Always**
Captions	2 (8,3%)	14 (58,3%)	7 (29,2%)	1 (4,2%)
Intralingual subtitles	8 (33,3%)	2 (8,3%)	11 (45,8%)	3 (12,5%)
Interlingual subtitles	11 (45,8%)	6 (25%)	5 (20,8%)	2 (8,3%)

Note. Participants had to select one of the options ("Never", "Rarely", "Frequently", "Always") to answer the following question: "How often do you use these formats when watching audiovisual content in English?".

Figure 17 shows the distribution of answers for Question n° 8, displaying frequency of use for each type of supporting written content (captions – blue bars; intralingual subtitles – orange bars; interlingual subtitles – grey bars).

Figure 17. *Distribution of answers for Question n° 8 in the questionnaire on viewing habits and supporting written content use reported for the group "Students$_{Uni}$" (university students enrolled in various degree programs).*



Q9. *If you turn captions/subtitles on, how important is it for you to read the same words that you are hearing when watching audiovisual content in English?*

The majority of participants deemed this type of (graphical) alignment important (*N* = 26, 61,9%). It was absolutely important for eleven participants (26,2%), and only five participants (11,9% out of the total number) deemed it as not important (*Table 13*).

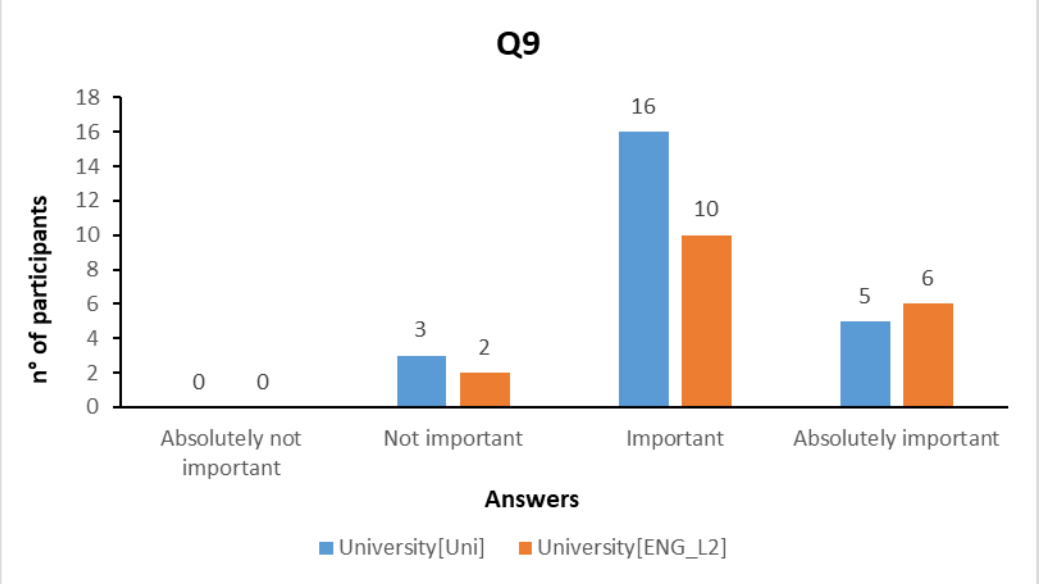
Table 13. *Number (count, percentage) of answers for Question n° 9 in the questionnaire on viewing habits and supporting written content use reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
Absolutely not important	-	-	-
Not important	2 (11,1%)	3 (12,5%)	5 (11,9%)
Important	10 (55,6%)	16 (66,7%)	26 (61,9%)
Absolutely important	6 (33,3%)	5 (20,8%)	11 (26,2%)

Note 1. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Figure 18 shows the distributions of answers for this question for each group. The trend is similar both groups, with the majority of participants that deems this characteristic of captions/subtitles important when watching audiovisual products in English ($Students_{ENG\_L2}$, N = 10, 55,6%; $Students_{Uni}$, N = 16, 66,7%). Six participants from the $Students_{ENG\_L2}$ group (33,3%) and five participants from the $Students_{Uni}$ group (20,8%) find this type of alignment absolutely important, while two participants from the $Students_{ENG\_L2}$ group (11,1%) and three participants from the $Students_{Uni}$ group (12,5%) find this type of alignment not important.

Figure 18. *Distribution of answers for Question n° 9 in the questionnaire on viewing habits and supporting written content use reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q10. *If you turn captions/subtitles on, how important is it for you to see written words appear at the same time they are pronounced when watching audiovisual content in English?*

Table 14 shows that almost half participants (*N* = 20, 47,6%) deems important this characteristic of captions/subtitles when watching audiovisual products in English. Almost 30% of participants, on the other hand, deem this feature not important.

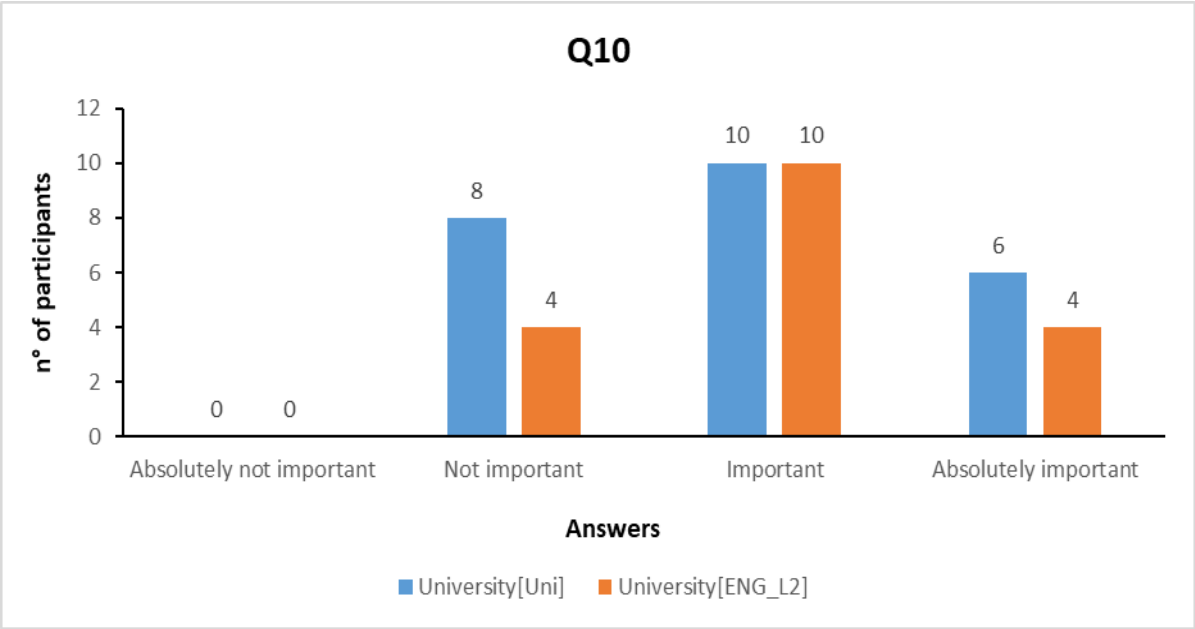
Table 14. *Number (count, percentage) of answers for Question n° 10 in the questionnaire on viewing habits and supporting written content use reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
Absolutely not important	-	-	-
Not important	4 (22,2%)	8 (33,3%)	12 (28,6%)
Important	10 (55,6%)	10 (41,7%)	20 (47,6%)
Absolutely important	4 (22,2%)	6 (25%)	10 (23,8%)

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Taking a look at the distribution of answers by group (*Figure 19*), we observe that the same number of participants from the two groups (*N* = 10 - $Students_{ENG\_L2}$, 55,6%; $Students_{Uni}$, 41,7%) think that this type of alignment is important. A higher number of participants from the $Students_{Uni}$ group (*N* = 8, 33,3%) compared to the number of participants from the $Students_{ENG\_L2}$ group (*N* = 4, 22,2%) think that this feature is not important. Lastly, four participants from the $Students_{ENG\_L2}$ group (22,2%) and six participants from the $Students_{Uni}$ group (25%) deem this alignment as absolutely important.

Figure 19. *Distribution of answers for Question n° 10 in the questionnaire on viewing habits and supporting written content use reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

3.4.3 Questionnaire on the potential use of live automatic captions in educational settings

Q1. *Live captions can be generated using an automatic speech recognition (ASR) system. Would you consider using it in class if your university provided this service?*

Table 15 summarizes the answers collected for the first question. More than half of students who were interviewed ($N = 27$, 64,3%) said that they would use live captions in class during lectures if their university provided this service. Fourteen other participants (33,3%) would not use live captions in class.

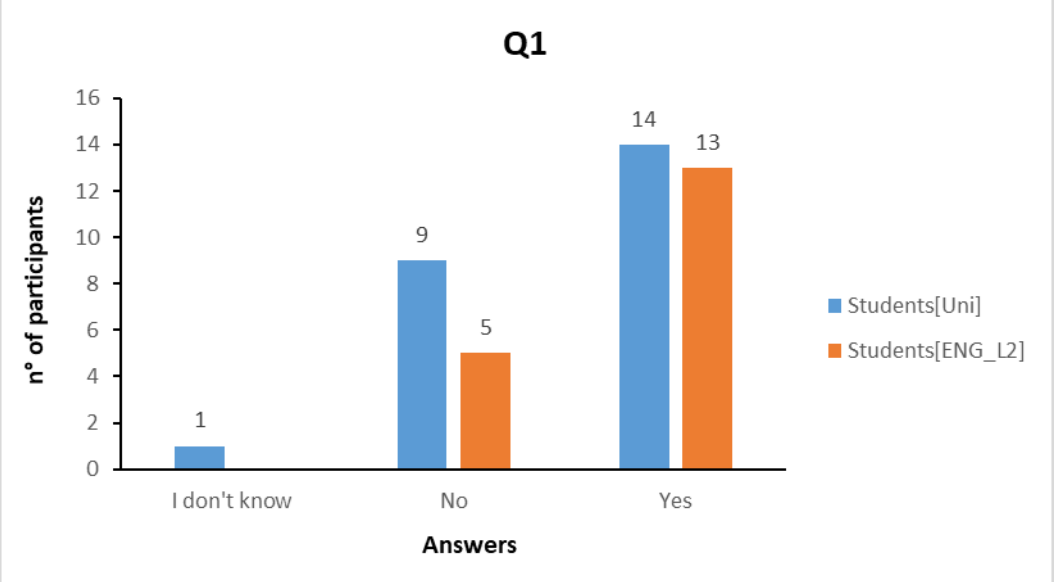
Table 15. *Number (count, percentage) of answers for Question n° 1 in the questionnaire on the use of automatic live captions in educational settings reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
No	5 (26,3%)	9 (37,5%)	14 (33,3%)
Yes	14 (72,2%)	13 (58,3%)	27 (64,3%)
I don't know	-	1 (4,2%)	1 (2,4%)

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Figure 20 shows the distribution of such answers divided by groups of participants. The trend is similar for both groups, with more than half of students who were interviewed stated that they would use live captions in class during lectures if their university provided this service (N = 27, 64,3%; $Students_{ENG\_L2}$, N = 13, 72,2%; $Students_{Uni}$, N = 14, 58,3%). Fourteen other participants (33,3%; $Students_{ENG\_L2}$, N = 5, 27,8%; $Students_{Uni}$, N = 9, 37,5%) would not use live captions in class, while one participant from the university students' group was not sure of using live captions if the service would be provided.

Figure 20. *Distribution of answers for Question n° 1 in the questionnaire on the use of automatic live captions in educational settings reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q2.*If live captions were available in class [for courses taught in English], which display format would you prefer?*

Table 16 summarizes participants' answers for Question n° 2: half participants (N = 21, 50%) would prefer captions to be displayed on two lines (see also *Figure 21* to check the distribution of the answers). Nine participants would prefer the speech-synchronized, incremental format of word-by-word captions (21,4%), while eight participants (19,1%) affirmed that they would be distracted by captions. Four participants (9,5%) chose the "captions + interlingual subtitles" option, stating that they would prefer the simultaneous presence of the speech transcript and its translation in their L1.

Table 16. *Number (count, percentage) of answers for Question n° 2 in the questionnaire on the use of automatic live captions in educational settings reported by group.*

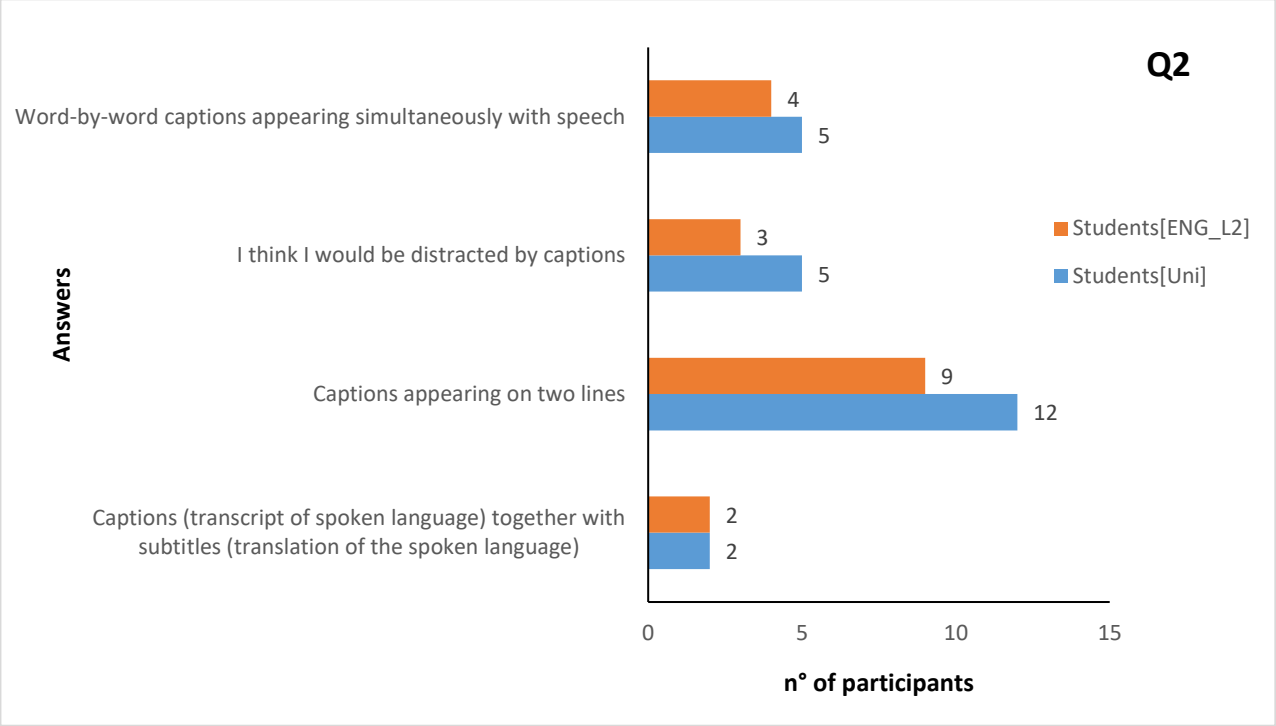
Answer	$\text{Students}_{\text{ENG\_L2}}$	$\text{Students}_{\text{Uni}}$	Total
Word-by-word	4 (22,2%)	5 (20,8%)	9 (21,4%)
Two-line captions	9 (50%)	12 (50%)	21 (50%)
Captions + Subtitles	2 (11,1%)	2 (8,3%)	4 (9,5%)
I think I would be distracted by captions	3 (16,7%)	5 (20,8%)	8 (19,1%)

Note. Results are reported by group ($\text{Students}_{\text{Uni}}$: university students enrolled in various degree programs, $\text{Students}_{\text{ENG\_L2}}$: university students enrolled in a Foreign Languages and Cultures degree program).

Figure 16 shows the distributions of the answers for Question n° 2 by group. The graph shows a similar pattern of responses for both groups, with the majority of participants preferring captions to appear on two lines ($\text{Students}_{\text{ENG\_L2}}$, N = 9, 50%; $\text{Students}_{\text{Uni}}$, N = 12, 50%). More participants from the university students enrolled in other courses ($\text{Students}_{\text{Uni}}$, N = 5, 20,8%) preferred word-by-word captions compared to university students enrolled in courses of English as L2 ($\text{Students}_{\text{ENG\_L2}}$, N = 4, 22,2%). An equal number of participants from both groups ($\text{Students}_{\text{ENG\_L2}}$, N = 2, 11,1%; $\text{Students}_{\text{Uni}}$, N = 2, 8,3%) selected the option "captions + interlingual subtitles" (N = 4, 9,5% out the total),

preferring this solution to be displayed in class. Finally, three participants from the $Students_{ENG\_L2}$ group (16,7%) and five participants from the $Students_{Uni}$ group (20,8%) said that they would be distracted by captions on screen.

Figure 21. *Distribution of answers for Question n° 2 in the questionnaire on the use of automatic live captions in educational settings reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q3. *In your opinion, which characteristic(s) of the captions would be crucial to support comprehension of lectures [delivered in English at the University]?*

Overall, the most important characteristics selected by participants were "Speed of appearance of text" ($N = 34$, 80,9%) and "Accuracy of automatic transcription" ($N = 32$, 76,2%). "Font size" was the third most important characteristic ($N = 17$, 39,5%). "Font type" was selected as a crucial characteristic by sixteen participants (38,1%), as well as "Number of lines on which text is displayed". Lastly, "Speech-captions alignment" was chosen by fourteen participants (30,9%) to be a relevant characteristic that would support comprehension of content in class during lectures in English (*Table 17*).

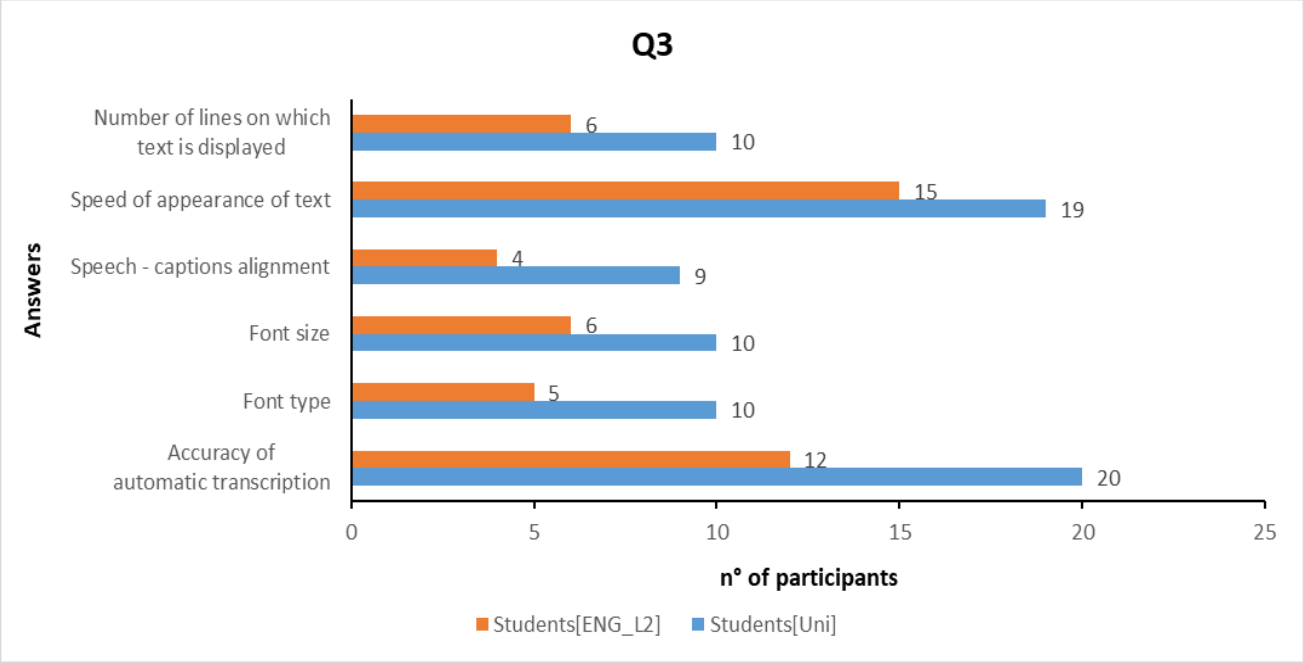
Table 17. *Number (count, percentage) of answers for Question n° 3 in the questionnaire on the use of automatic live captions in educational settings reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
N° of lines on which text is displayed	6 (33,3%)	10 (41,7%)	16 (38,1%)
Speed of appearance of text	15 (83,3%)	19 (79,2%)	34 (80,9%)
Speech-captions alignment	4 (22,2%)	9 (37,5%)	13 (30,9%)
Font size	6 (33,3%)	10 (41,7%)	16 (38,1%)
Font type	5 (27,8%)	10 (41,7%)	15 (35,7%)
Accuracy of automatic transcription	12 (66,7%)	20 (83,3%)	32 (76,2%)
I don't know	-	-	-

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Figure 22 shows the trend of answers by group: "Speed of appearance of text" ($Students_{ENG\_L2}$, N = 15, 83,3%; $Students_{Uni}$, N = 19, 79,2%) and "Accuracy of automatic transcription" ($Students_{ENG\_L2}$, N = 12, 66,7%; $Students_{Uni}$, N = 20, 83,3%) are the most chosen by participants in both groups. While "Number of lines on which text is displayed" ($Students_{ENG\_L2}$, N = 6, 33,3%; $Students_{Uni}$, N = 10, 41,7%) and "Font size" ($Students_{ENG\_L2}$, N = 6, 33,3%; $Students_{Uni}$, N = 10, 41,7%) showed the same trend, participants in the group of university students ($Students_{Uni}$) said that "speech-captions alignment" (N = 9, 37,5%) and "font type" (N = 10, 41,7%) were two important characteristics of live captions when it came to support comprehension of lectures, compared to university students enrolled in a *Foreign Languages* course ($Students_{ENG\_L2}$, "speech-captions alignment": N = 4, 22,2%; "font type": N = 5, 27,8%).

Figure 22. *Distribution of answers for Question n° 3 in the questionnaire on the use of automatic live captions in educational settings reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q4. *Do you think your knowledge of English would get better if live captions were provided? Why? [More than one answer per participant]*

Table 18 summarizes the answers for each group of participants to the fourth question. Most participants think that live captions would help them in recovering words (N = 29, 69,1%) and improve their vocabulary knowledge (N = 23, 54,8%). Nine participants (21,4%) said that their pronunciation would benefit from the presence of live captioning in class during lectures. Only four participants (9,5%) said that live captions would not help them improving their knowledge of English, that the written input would be redundant and that they prefer listening to the professor speaking (*Figure 23*).

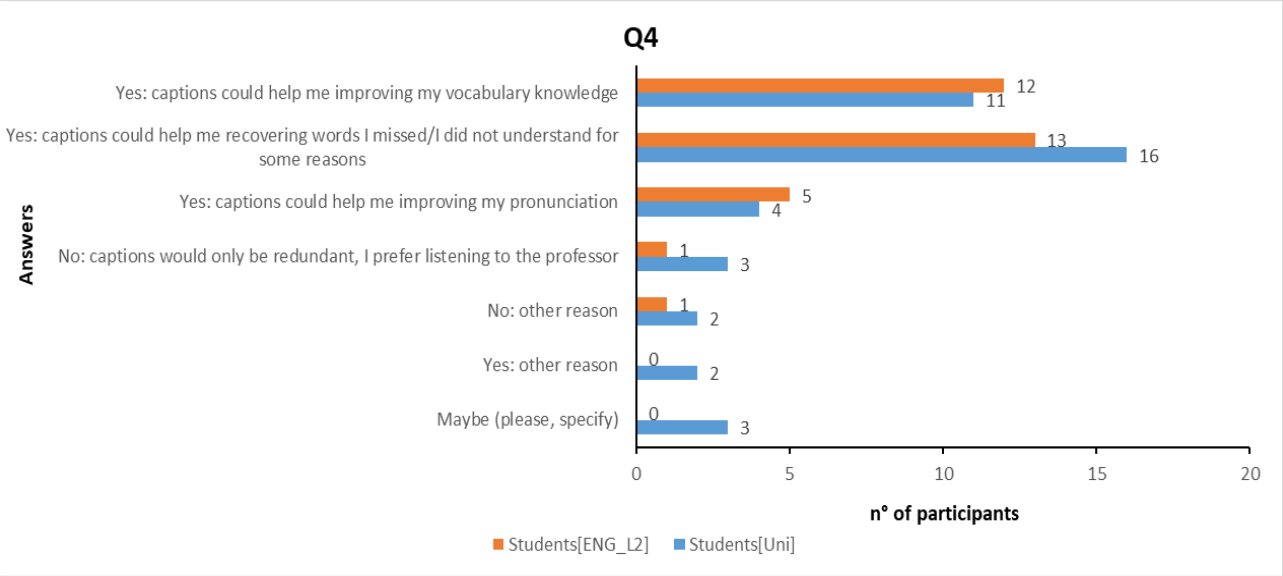
Table 18. *Number (count, percentage) of answers for Question n° 4 in the questionnaire on the use of automatic live captions in educational settings reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
Yes - Improvement of vocabulary knowledge	12 (66,7%)	11 (45,8%)	23 (54,8%)
Yes – Help in recovering words	13 (72,2%)	16 (66,7%)	29 (69,1%)
Yes – Improvement in pronunciation	5 (27,8%)	4 (16,7%)	9 (21,4%)
No – Redundancy of captions	1 (5,6%)	3 (12,5%)	4 (9,5%)
Yes – Other reason	-	2 (8,3%)	2 (4,8%)
No – Other reason	1 (5,6%)	2 (8,3%)	3 (7,1%)
Maybe	1 (5,6%)	3 (12,5%)	3 (7,1%)
Other	-	-	-

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Figure 23 shows the distributions of the answers by group, highlighting a similar trend between the two, with improvement of vocabulary knowledge ($Students_{ENG\_L2}$, N = 12, 66,7%; $Students_{Uni}$, N = 11, 45,8%) and help in recovering words ($Students_{ENG\_L2}$, N = 13, 72,2%; $Students_{Uni}$, N = 16, 66,7%) as major benefits from live captions. More participants from the $Students_{ENG\_L2}$ group think that live captions would help them with improving their pronunciation (N = 5, 27,8%) compared to their peers enrolled in various courses at university ($Students_{Uni}$, N = 4, 16,7%). Only one participant in the $Students_{ENG\_L2}$ group (5,6%) and three participants in the $Students_{Uni}$ group (12,5%) said that they would be distracted by captions because text would be redundant.

Figure 23. *Distribution of answers for Question n° 4 in the questionnaire on the use of automatic live captions in educational settings reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Two participants from the $Students_{Uni}$ group stated that live captions would help them due to other reasons. One participant stated that live captions

"*[…] would help me to be more concentrated to the speech*",

while the other motivated their answer writing the following statement:

"*There are times when accent of the speaker makes it difficult to understand or turns into distraction. In that case caption helps follow the lesson.*"

On the other hand, two other participants from the $Students_{Uni}$ group stated that live captions would not help them improving their knowledge of English. One participant said that

"*captions would distract them too much*",

while another stated that

"*I would feel the cognitive workload to be higher. Besides I suspect that the English of several non-native English speakers professors is probably not at a good enough level to be interpreted well by ASR*".

One participant stated that their knowledge of English would not benefit from the presence of live captions, but did not add any further motivation to their answer.

Lastly, three participants from the $Students_{Uni}$ group said that maybe they would benefit from live captions. One participant stated that

"*It would be a way to make sure I am understanding the class, but it may be a source of distraction, too*",

while the other said that "*Maybe because in a lecture there are maybe technical terms that I don't know or don't know how to write but I feel like I would be distracted from the lecturer and would only pay attention to the captions*."

One participant among those did not motivate their answer.

Q5. *ASR systems are not always accurate in their transcriptions due to various factors (e.g., environmental noise, low quality microphone, etc.). Would you find live captions useful even though there were some errors in them?*

Table 19 sums up the answers of participants by group. Thirty-one participants (73,8%) answered that they would consider the quantity and/or quality of errors before deciding the usefulness of automatically generated live captions. Only four participants (9,5%) would consider live captions useful despite errors, while seven participants (16,7%) would not find live captions useful if the transcription had some errors in it.

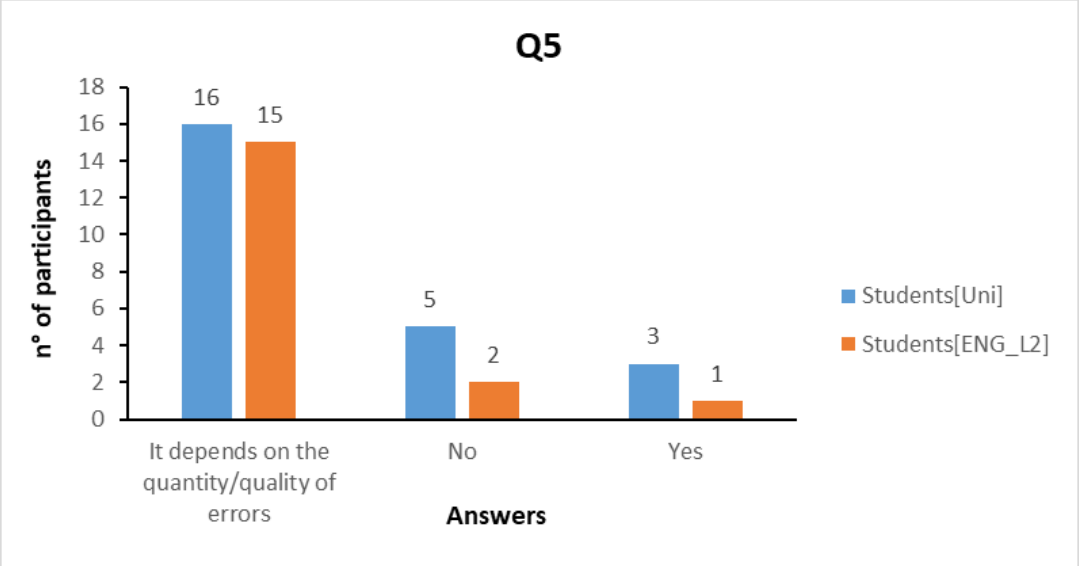
Table 19. *Number (count, percentage) of answers for Question n° 5 in the questionnaire on the use of automatic live captions in educational settings reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
Yes	1 (5,6%)	3 (12,5%)	4 (9,5%)
No	2 (11,1%)	5 (20,8%)	7 (16,7%)
It depends on the quantity/quality of errors	15 (83,3%)	16 (66,7%)	31 (73,8%)

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Figure 24 shows a similar trend for all answers for both groups of students, with most of them considering the overall quality of the transcription before deciding if live captioning would still be useful despite the errors ($Students_{ENG\_L2}$, N = 15, 83,3%; $Students_{Uni}$, N = 16, 66,7%). However, a higher number of participants from the $Students_{Uni}$ group (N = 5, 20,8%) compared to the number of participants from the $Students_{ENG\_L2}$ group (N = 2, 11,1%) don't think that captions would be useful if they contained errors. Similarly, a higher number of participants from the $Students_{Uni}$ group (N = 3, 12,5%) compared to the number of participants from the $Students_{ENG\_L2}$ group (N = 1) think that captions would be useful despite errors.

Figure 24. *Distribution of answers for Question n° 5 in the questionnaire on the use of automatic live captions in educational settings reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q6. *Why? Motivate your previous answer. [More than one answer per participant]*

Overall, participants stated that they would consider using live captions depending on the overall quality of the transcription because errors could confuse them (N = 23, 54,8%) and hinder their ability to understand the content of lectures (N = 22, 52,4%). Participants also think that errors could unnecessarily distract them from listening to the speakers (N = 19, 45,2%). Only four

participants (9,5%) said that they would ignore errors and just focus on the speaker's speech (see *Table 20*).

Table 20. *Number (count, percentage) of answers for Question n° 6 in the questionnaire on the use of automatic live captions in educational settings reported by group.*

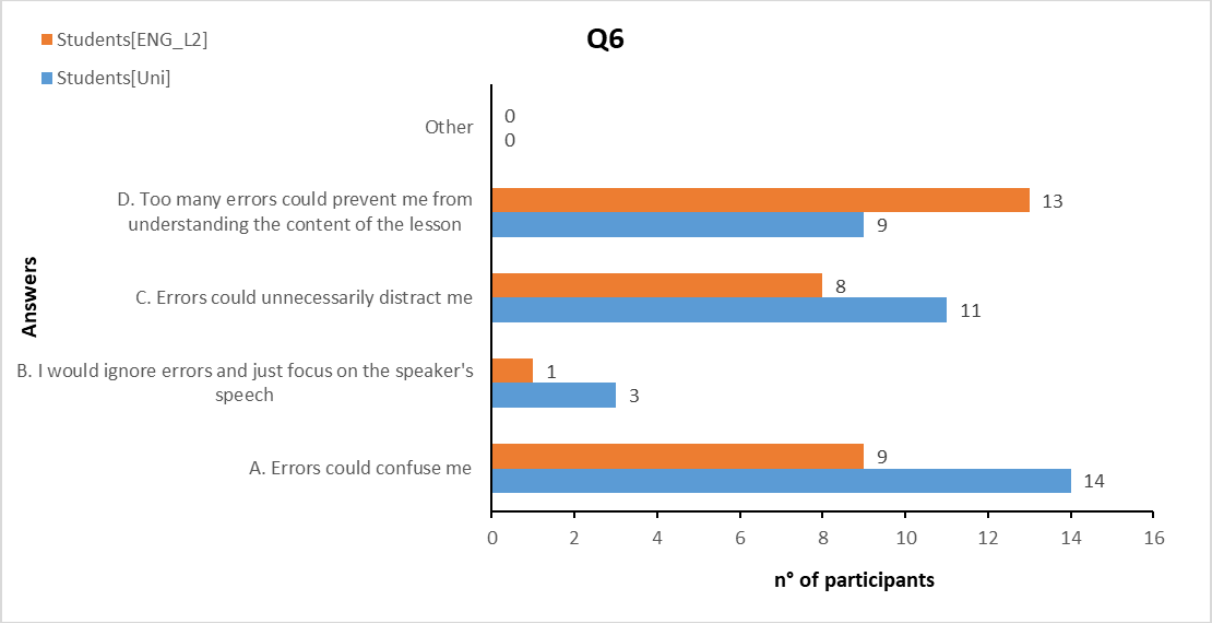
Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
A	9 (50%)	14 (58,3%)	23 (54,8%)
B	1 (5,6%)	3 (12,5%)	4 (9,5%)
C	8 (52,4%)	11 (45,8%)	19 (45,2%)
D	13 (72,2%)	9 (37,5%)	22 (52,4%)
Other	-	-	-

Note 1. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Note 2. Labels in the *Answer* column: *Option A*: Errors could confuse me; *Option B*: I would ignore errors and just focus on the speaker's speech; *Option C*: Errors could unnecessarily distract me; *Option D*: Too many errors could prevent me from understanding the content of the lesson; *Other*: other answer.

Figure 25 shows that most university participants in the $Students_{ENG\_L2}$ group said that their main concern was that too many errors could prevent them from understanding the content of the lesson (*Option D*: $N = 13$, 72,2%). Participants were also concerned that errors could confuse (*Option A*: $N = 9$, 50%) and distract them from listening (*Option C*: $N = 8$, 52,4%). On the other hand, participants from the $Students_{Uni}$ group stated that their main concern was that errors could confuse (*Option A*: $N = 14$, 58,3%) and distract them (*Option C*: $N = 11$, 45,8%). Nine participants (37,5%) stated that too many errors could hamper the comprehension of the content of the lecture (*Option D*). One participant from the $Students_{ENG\_L2}$ group (5,6%) and three participants from the $Students_{Uni}$ group (12,5%) stated that they would ignore the errors in the text and mostly focus on the speaker's speech (*Option B*).

Figure 25. *Distribution of answers for Question n° 6 in the questionnaire on the use of automatic live captions in educational settings reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q7. *In your opinion, would it be helpful if a specific display format signaled transcription errors/how confident the system was of its transcription?*

Overall (*Table 21*), the trend of answers is similar for both groups, despite the differences in the formulation of the question. Less than half of the participants (N = 16, 38,1%) were not sure this kind of implementation would be useful to them. Sixteen participants (38,1%) stated that they would not find this markup useful. Finally, only ten participants (23,8%) would find this type of markup useful. The graph in *Figure 26* shows the distribution of answers to Question n°7 for each group.

Table 21. *Number (count, percentage) of answers for Question n° 7 in the questionnaire on the use of automatic live captions in educational settings reported by group.*

Answer	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
Yes	4 (22,2%)	6 (25%)	10 (23,8%)
No	7 (38,9%)	9 (37,5%)	16 (38,1%)
I'm not sure	7 (38,9%)	9 (37,5%)	16 (38,1%)

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

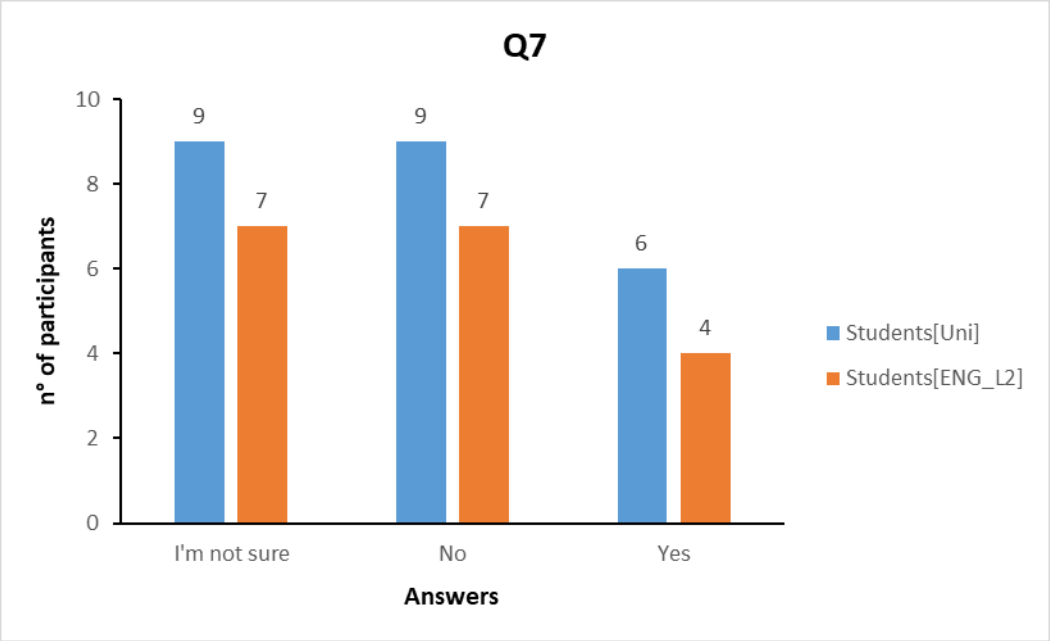
In your opinion, would it be helpful if a specific display format signaled transcription errors? [$Students_{ENG\_L2}$ group]

Fourteen participants from the *$Students_{ENG\_L2}$ group* were more inclined to answer to think that this markup would not be helpful (*N* = 7, 38,9%) or they were unsure of its usefulness (*N* = 7, 38,9%). Only four participants in this group said that they thought the system to be useful (*N* = 4, 22,2%).

In your opinion, would it be helpful if a specific display format signaled how confident the system was in its transcription? [$Students_{Uni}$]

Nine participants in the $Students_{Uni}$ group answered that they weren't sure of the usefulness of this type of markup (37,5%) and nine participants were more inclined to not find it helpful (37,5%). Only six participants (25%) thought that having a color-coded markup to signal the system's confidence could be useful (25%).

Figure 26. *Distribution of answers for Question n° 7 in the questionnaire on the use of automatic live captions in educational settings reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

Q8. *Why? Please, motivate your previous answer. [More than one answer per participant]*

Most participants were concerned that this color-coded markup may distract them (*N* = 28, 66,7%), while the remaining participants thought they would help them to know if the system is confident enough of its transcription (*N* = 11, 26,2%) (see *Table 22* for details).

Table 22. *Number (count, percentage) of answers for Question n° 8 in the questionnaire on the use of automatic live captions in educational settings reported by group.*

Answers	$Students_{ENG\_L2}$	$Students_{Uni}$	Total
It would help me	4 (22,2%)	7 (29,2%)	11 (26,2%)
I would find it distracting	14 (77,8%)	14 (58,3%)	28 (66,7%)
Other	-	4 (16,7%)	4 (9,5%)

Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs, $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program).

Analyzing data by group (*Figure 27*), we can see a similar trend in the answers. Participants from both groups mainly think that this kind of markup may be distracting ($Students_{ENG\_L2}$, $N = 14$, 77,8%; $Students_{Uni}$, $N = 14$, 58,3%); only a small percentage think that this system may be informative ($Students_{ENG\_L2}$, $N = 4$, 22,2%; $Students_{Uni}$, $N = 7$, 29,2%).

Four participants from the $Students_{Uni}$ group added considerations to their answers. One of them specified that they

"[…] don't need them. But if I did, at least at the beginning it would add an additional cognitive burden if I had issues with English",

while the other added that this markup

"[…] helps to know which words may present errors, but it is overall confusing and too distracting".

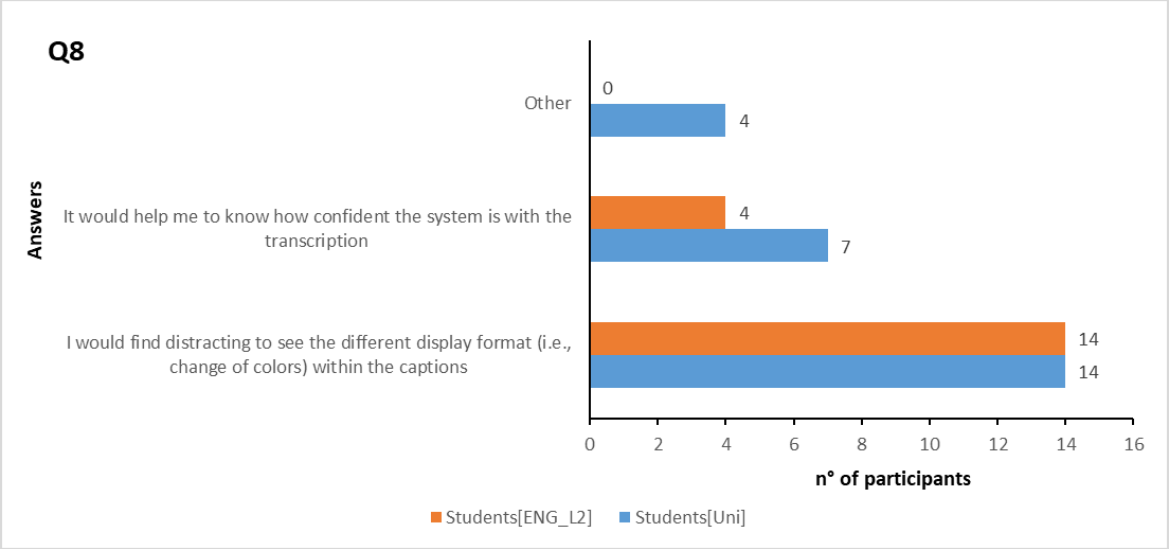
Another one said that

"It would be confusing, I'd prefer not having them if they're not correct",

and the last one concluded that

"*It would be a double distraction".*

Figure 27. *Distribution of answers for Question n° 8 in the questionnaire on the use of automatic live captions in educational settings reported by group.*



Note. Results are reported by group ($Students_{Uni}$: university students enrolled in various degree programs – blue bars; $Students_{ENG\_L2}$: university students enrolled in a Foreign Languages and Cultures degree program – orange bars).

3.5 Discussion

The goal of this study was twofold:

- To assess the day-to-day viewing habits of audiovisual products in English and the use of audiovisual translation products to aid speech processing in university students (native speakers of Italian, L2 speakers of English);
- To evaluate the use of live automatic captions during lectures in English at university and the implementation of a color-coded markup to display the confidence the ASR system has in its transcription.

We formulated three research questions, which we will address below.

RQ1. Do students regularly watch audiovisual products in English and use audiovisual translation products to aid speech processing and content comprehension of the L2 in their day-to-day lives?

The first questionnaire examined the viewing habits of Italian university students in watching audiovisual content in English, as well as their use of audiovisual translation products (or supporting written content - captions, intralingual, and interlingual subtitles) to aid speech processing and content comprehension as L2 speakers of English. Since speech processing can be particularly challenging for L2 speakers (Goh, 2000), we aimed to investigate whether this population uses audiovisual translation products and how frequently they rely on this resource to aid speech processing when exposed to authentic linguistic input.

Overall, university students - even though they differed in their educational backgrounds and degree programs they were enrolled in, but were comparable in terms of proficiency and listening skills - demonstrated similar habits and opinions regarding the use of audiovisual translation products to assist with speech processing and content comprehension when watching audiovisual products in English. Results show that most university students frequently watch audiovisual content in English, particularly on their portable devices. They primarily watch products related to news and entertainment, such as TV series, movies, vloggers, and short video clips across various web platforms. In contrast, educational content is watched less frequently. These results align with the latest trends (Preply 2023, 2024), where younger generations watch audiovisual products daily on streaming services and social media and are exposed to authentic linguistic input (Montero Perez, 2022).

Interestingly, the first difference between the two groups arises from their preference for the use of specific audiovisual translation products over others. Most participants stated that they regularly use audiovisual translation products, but while students enrolled in the *Foreign Languages and Cultures* degree program prefer captions over intralingual and interlingual subtitles, university students enrolled in other degree programs prefer intralingual subtitles over captions and interlingual subtitles. In both cases, these students prefer using audiovisual translation products that match the language spoken in the aural input - that is, English. The reasons behind their preference for these audiovisual translation products were the same: the trend showed that students use these tools to understand better the content they are watching, learn new words, and improve their pronunciation (even though their language proficiency and listening skills are at the high end of the spectrum). Again, these results are comparable to those summarized in the Preply web platform (2023, 2024) and with the main reason why L2 speakers use audiovisual translation products to improve their linguistic knowledge and support content comprehension (e.g., Gernsbacher, 2015; Montero Perez, 2022). Only a few participants (mainly students enrolled in various degree programs, $Students_{Uni}$ group) stated that they do not use audiovisual translation products because they find them detrimental in some way - especially at the level of the distribution of attention between the oral speech stream and the text - or useless, preferring to only listen to speakers to support speech processing and improve their skills. These results seem to highlight that students may be aware of the benefits the use of captions and subtitles during the vision of audiovisual content in their L2 can bring (e.g., Danan, 2016; Alm, 2019). Some additional responses provided by students also highlight *when* they utilize supporting written content the most to enhance their speech processing, that is, in those specific situations where mapping sounds to meaning can be difficult - for example, when listening to speakers talking with unfamiliar accents or when the speech is delivered at a high rate (Mitterer & McQueen, 2009; Birulés-Muntané & Soto-Faraco, 2016). In these cases, audiovisual translation products (especially captions and intralingual subtitles for students with high proficiency in English) are also found to be beneficial by students who rarely use captions and intralingual subtitles to aid speech processing.

Results also highlighted that participants (especially from the $Students_{ENG\_L2}$ group) consider it important to read the exact words they hear while watching audiovisual products in their second language. This preference seems to support the benefits of redundant, bimodal input: in fact, existing literature on the topic suggests that the presence of textual information matching the aural input helps learners at the lower levels of language processing (word identification and sentence parsing) to devote more cognitive resources for the higher-level processes, such as content comprehension (e.g., Paivio, 1991; Bird & Williams, 2002; Hulstijn, 2003; Danan, 2016). In the same way, incongruities (namely, the mismatch between the phonological form coming from speech and the graphical form

coming from the text) may lead to an increase in frustration and less enjoyment of the audiovisual product, as supported by previous research (Szarkowska *et al*., 2024). For the same reasons, the temporal alignment of speech and text seems to be an important graphical feature for many students from both groups. However, some students from the StudentsUni group do not deem this feature important, and this may be related to their preference for using intralingual subtitles when watching audiovisual products in English: in fact, this type of audiovisual translation products generally display text on two lines (42 characters per line in subtitles, 32 characters per line in captions) for maximum 6/7 seconds (e.g., 3PlayMedia, 2023; Netflix, 2024; Nettelbeck, 2024). This textual "preview" may help listeners anticipate upcoming speech, help manage the listening effort, and aid speech processing (Danan, 2016). However, further research is needed to answer these questions more related to the characteristics of the audiovisual translation products (see Nettelbeck, 2024, for a literature review). Finally, even though the proficiency level and listening skills of participants are at the high end of the spectrum, some students from the $\text{Students}_{\text{Uni}}$ group still prefer using interlingual subtitles. The primary reason for their choice may be related to the fact that reading the text in their first language may be less cognitively taxing than reading-while-listening to it in their second language, since they may be attempting to match sounds to letters while engaged in processing information in a multimodal setting (Baranowska, 2020).

RQ2. Would university students use live automatic captions in class during lectures in English?

The second questionnaire aimed to investigate whether university students find automatic live captions helpful and their interest in using this tool during lectures delivered in English. Additionally, it sought to determine if a color-coded markup of the text would be beneficial to students in indicating the confidence level of the system in its transcriptions.

A significant portion of students - approximately 70% of the participants in our sample - would be open to using automatic live captions in class during lectures if their university offered this tool. However, a high percentage of university students enrolled in various degree programs (N = 9) compared to students from the $\text{Students}_{\text{ENG\_L2}}$ group (N = 5) stated that they would not use live automatic captions if provided by their university. Most students had a positive attitude toward the potential implementation of automatic live captions during English lectures. However, when asked about the usefulness of this tool despite its inaccuracies, many expressed uncertainty regarding its effectiveness in aiding speech processing and content comprehension, which they believed depended on the number and quality of errors produced by the automatic speech recognition (ASR) system.

Their main concern was the overall quality of the transcription. They understood that the poor quality of automatic live captions could impact their ability to process speech and hinder comprehension of the content of lectures. This could leave students feeling confused and distracted, as errors in the captions might divert their attention away from what was being said. Therefore, the simultaneous presence of written and spoken input could be counterproductive, as previous research has already highlighted (e.g., Chan *et al*., 2019). In this context, only a few students indicated that they would find automatic live captions helpful, even with transcription errors, as they claimed they could ignore those errors to focus entirely on the speech.

When asked whether they believed that live automatic captions would help them improve their knowledge of the L2, most participants from both groups generally agreed that this tool would assist them in recovering words they missed while listening and improving their vocabulary size. Students indicated they would use live automatic captions also to improve their pronunciation, focus more on the listening task, and effectively manage speech processing in challenging situations — such as trying to comprehend speech from someone with an unfamiliar accent. A notable point from these responses is that, regardless of their educational backgrounds and current enrollment in various degree programs, students from both groups consistently expressed similar opinions about the potential benefits of live automatic captions for improving their proficiency in English. However, while most students expressed a positive attitude towards the use of automatic live captions in educational settings, nearly 35% of students in our sample (mostly students enrolled in various degree programs) indicated that they would not consider using this type of captions during lectures delivered in English. Their concerns were mainly related to the belief that this tool would be somewhat distracting, and they felt that having captions was unnecessary; they preferred simply listening to the professor. In summary, these students worried that the presence of both captions and slide text could negatively affect their attention, and that having to read a large amount of text (captions + the text in the slides) while processing spoken information in their second language - especially while taking notes - might be overwhelming for them (Kruger, 2013).

In this questionnaire, we were also interested in exploring students' preferences regarding the characteristics of live automatic captions. Our goal was to gather their opinions on the display format and overall features live captions should have in educational contexts to aid speech processing and content comprehension. In this regard, the majority of students in both groups stated that they would prefer automatic live captions to be displayed on two lines rather than in other formats, such as an incremental, word-by-word display format or in a “captions + interlingual subtitles” (where text in both L2 and its translation in the L1 was to be shown simultaneously, preferred by 10% of the students

in our sample). The two-line subtitles are the most common display format for audiovisual translation products (Nettelbeck, 2024), and, in general, many viewers' preferred type of format, as the existing literature highlights (e.g., Szarkowska & Gerber-Morón, 2019). Interestingly, a comparable number of students from both samples (about 20% of students for each option) stated that they preferred the incremental word-by-word display format or not seeing automatic live captions at all, since they would be distracted by the text. Among the other characteristics automatic live captions should have to aid speech processing and content comprehension during lectures, students considered the speed at which the text appears on the screen and the accuracy of the transcriptions to be the most important features. These results align with previous research on the importance of the timing of text appearance (Kruger *et al*., 2022) and the quality of transcriptions (Shimogori *et al*., 2010; Chan *et al*., 2019). If the automatic live captions produced by the ASR system were to be generated too slowly, it could result in a significant misalignment between the spoken words of the professor and the text displayed as captions. This misalignment could lead students to struggle with managing the different modalities of information and find it challenging to focus on all the elements in the class (the professor talking, the slides and their graphical and textual content, and the captions) (Szarkowska *et al*., 2011). Additionally, if the text of automatic live captions contained errors, it could again affect speech processing, not helping listeners and hindering content comprehension for good (Chan *et al*., 2019). Indeed, a participant in our study pointed out an important concern: the ASR system may be more likely to make errors when recognizing the speech of professors who are not native English speakers. This concern is valid, especially when considering the various factors that can influence the performance of speaker-independent ASR systems (see Chapter 1 for a discussion on the topic). In sum, these results highlight that accuracy needs to be taken seriously to guarantee access to information and content of the lectures, which aligns with previous research (Shimogori *et al*., 2010; Romero-Fresco & Fresno, 2023).

RQ3. Would university students find it helpful to have the text display transcription errors/the confidence level via a color-coded markup?

When asked about the perceived usefulness of the implementation of a color-coded markup in automatic live captions to display transcriptions errors (students enrolled in the *Foreign Languages and Cultures* degree program, $\text{Students}_{\text{ENG\_L2}}$ group) or the confidence level of the ASR system in its transcription (university students enrolled in various degree programs, $\text{Students}_{\text{Uni}}$ group), most participants were unsure or not inclined to say that this colored markup would be useful in the educational context. These findings are consistent with those of the study on the topic (e.g., Shiver &

Wolfe, 2015; Berke *et al*., 2017). These students argued that the changing colors in the text during lectures, while they were looking at slides and listening to the professor, could be distracting. Students also mentioned that the color-coded markup in the automatic live captions could be too overwhelming or distracting and, therefore, not helpful. During a lecture, students would be exposed to many different inputs (spoken words, text, and images) while engaged in multiple tasks simultaneously, such as listening to the professor, processing speech, taking notes, reading the text in the slides and the captions. Incorporating a color-coded markup into the captions would introduce an additional layer of meaning to the text. This would require students to interpret the significance of each color, which could distract them from other tasks and potentially hinder their understanding of the speech and the overall content. However, almost 25% of participants in our sample stated that they would find this markup useful to signal transcription errors and the confidence levels of the ASR system, denoting a certain trust in the system's performance, similar to what Vertanen and Kristensson found in their study (2018).

3.6 Limitations and future directions

This study has three major limitations:

1. The first limitation is related to the fact that data was collected using two slightly different questionnaires (as previously declared in the *Methods* section). The first version of the questionnaire was used to collect data from university students enrolled in the *Foreign Languages and Cultures* degree program. These questions were included as a part of a larger questionnaire on the effects of different display formats of captions (word-by-word versus two-line captions) shown during lectures in English (Ivanchenko, unpublished MA thesis, 2023) (*Appendix A.I* and *B.I*). The second version of the questionnaire was used in two subsequent studies to collect data from students enrolled in various degree programs at the University (*Appendix A.II* and *B.II*). Due to these discrepancies in the formulation of the questions, participants' profiles regarding viewing habits and supporting written content use could be partially compared. While some of these differences are minor (e.g., the inclusion of options related to interlingual subtitles in Question n° 7), they prevented the generalization of results from the data collected across three different experiments. This is also true for the second questionnaire, where a crucial question – Question ° 7 – had a notable difference in the text of the question itself (see *Procedure*, §3.3.2). The differences between the first and

second versions of the questionnaires were linked to the improvements in the formulation of the questions and options after the first study (Ivanchenko, unpublished master's thesis, 2023). Moreover, future research should consider using a questionnaire which aims to investigate more in depth participants' linguistic profiles and learning strategies, so as to investigate more in depth (and potentially correlate) their individual characteristics and preferences for certain audiovisual translation products.

2. Another limitation of our study is that the sample groups lacked diversity regarding proficiency levels and listening abilities. In fact, most participants exhibited a high level of proficiency in English and possessed strong listening skills (see *Participants*, §3.3.1). Due to the variability in these factors, students' preferences for different types and characteristics of written support content may differ. Future studies should aim to test a larger number of participants at the lower end of the proficiency and listening ability spectrum (below the B2 level). This approach would help further investigate their preferences for caption usage and viewing habits and the potential implementation of live captions in educational settings, which may vary based on students' proficiency in English (Venturini *et al*., 2022).

3. The third and final limitation relates to the example of the color-coded markup we presented to participants. Since the second study of this project (see following chapter) had not yet begun when we started collecting data, we presented participants with a frame from Berke (2017) as an example of color-coded markup (see *Appendix B.I* and *B.II*) instead of providing an example created *ad hoc* for our follow-up study. Future researchers should consider incorporating customized demonstrations of the markups or display formats they intend to implement. This way, they can present a completed demo and solicit feedback, similar to Berke's approach in his 2017 project.

Lastly, this study provided an insight into how L2 speakers of English use audiovisual translation products to aid speech processing and content comprehension. Their responses offered a glimpse into the strategies they employ when they face challenges during listening comprehension. However, to validate our hypotheses, a follow-up study is needed. Specifically, regarding the effects of display formats, we conducted an eye-tracking experiment to measure the impact of two-line versus word-by-word human- and automatic-generated captions on speech processing and content comprehension among L2 speakers of English with varying levels of proficiency (Pucci & Bencini, *in prep.*).

3.7 Conclusion

This study was conducted to investigate the viewing habits of Italian university students of audiovisual products in English, as well as their use of supporting written content (captions, intralingual, and interlingual subtitles) to aid speech processing of their L2. Previous literature on the topic has highlighted that intralingual subtitles aid speech processing and comprehension (e.g., Gernsbacher, 2015; Montero Perez, 2022); however, only a few studies have been carried out to assess the benefits of automatic captions in educational contexts (e.g., Wald & Bain, 2008; Chan *et al*., 2019). One of the main focuses related to captions generated by Automatic Speech Recognition (ASR) systems is *accuracy*: due to various factors (speaker-, acoustic-, and environment-related), transcriptions generated by these systems may be inaccurate, compromising speech processing and content comprehension (Chan *et al*., 2019). Therefore, it may be beneficial for users to know when and how the system is confident with its transcription. For instance, users may be informed by the system about how confident it is in transcribing a word in the captions by showing it in a color-coded markup. This display format could help users assess the reliability of transcriptions and potentially detect errors. Previous research on the topic has found mixed results across different populations (e.g., Berke *et al*., 2017; Vertanen & Kristensson, 2018): for this reason, we also investigated if university students would use automatic live captions in lectures delivered in English and if they would find the implementation of a color-coded markup in the captions helpful.

Our study found that most university students rely on audiovisual translation products, particularly captions and intralingual subtitles, to help them process speech, improve their vocabulary, and enhance their understanding of audiovisual content in English. This is true not only for students with high proficiency and strong listening skills in their second language, but also for those who rarely use captions or subtitles (especially when listening effort may increase due to adverse acoustic conditions - for example, when listening to someone with a heavy, unfamiliar accent). Processing L2 sounds is more challenging than processing L1 sounds; therefore, written text can provide a stable input compared to speech and help learners map sounds to concepts through the text of the captions or subtitles (Bird & Williams, 2002; Mitterer & McQueen, 2009). Most of these students would use automatic live captions in educational settings - specifically, during lectures delivered in English. However, they are concerned about the accuracy of the transcriptions; if they are not accurate enough, errors might distract students and negatively impact speech processing and comprehension (Chan *et al*., 2019). When presented with the possibility of the implementation of a color-coded markup in the text of the captions to signal errors or the reliability of the ASR system in its transcription, students expressed that they did not find this feature useful or were uncertain about its potential benefits (Berke

et al., 2017). Their concern was due to the dynamic color changes, which may distract them from following the lecture.

In sum, this study provided an overview of the viewing habits and preferences for audiovisual translation products among L2 speakers of English, as well as the strategies of use of audiovisual translation products to support speech processing and comprehension. It allowed us the opportunity to conduct a preliminary evaluation on the helpfulness of color-coded markups in improving the reliability of automatic live captions. Additionally, we gathered opinions, concerns, and insights from L2 speakers of English regarding the essential features that display formats of live automatic captions should have to enhance speech processing and comprehension in educational contexts, particularly during academic lectures delivered in English.

These results served as a starting point for the subsequent two studies we conducted. We took L2 speakers of English's answers into account while developing the markups presented in our second and third study (see Chapters 4 and 5). In fact, the results and students' feedback contributed to the creation of two of the four markups we presented and tested in our third study (V3, see Chapter 4, §4.5.5).

4 Testing the reliability of an ASR system in real-world contexts

An analysis of accuracy and confidence of transcription aimed at creating a series of color-coded markups to display confidence in automatic captioning

4.1 Introduction

The accuracy of transcriptions in automatic live subtitles and captions is essential for aiding speech comprehension and access to information for many populations, L2 speakers included (Wald & Bain, 2008; Shimogori *et al*., 2010; Gernsbacher, 2015; Butler *et al*., 2019; Chan *et al*., 2019; Romero-Fresco & Fresno, 2023; Kuhn *et al*., 2024). High-quality transcriptions result from the interplay of a robust infrastructure of the ASR system and the quality of the audio tracks the system needs to process. While much effort is being put into improving the architecture of ASR systems by engineers (O'Shaughnessy, 2024), previous research has also highlighted the equally important role of the acoustic qualities of the speech signal in ensuring high transcription accuracy (Alharbi *et al*., 2021). Specifically, acoustic and environmental factors (for instance, the type of microphone used to collect the speech samples or the acoustic properties of rooms and halls) (Van Den Heuij *et al*., 2018; Del Rosso & Brambilla, 2022; Dua *et al.*, 2023) and speaker-related characteristics (Benzeghiba *et al*., 2007; Feng *et al*., 2021, 2024) are listed as the factors that most affect the recognition process and accuracy of ASR systems. While some factors can be controlled to ensure a high-quality signal, others are inherently linked to real-world contexts and the natural variability that exists among speakers. For example, a person could be a non-native speaker of a specific language, and the pronunciation of words and prosody while producing speech in the L2 may be affected by their L1 (Mattys *et al*., 2012; Emara & Shaker, 2024). Recognizing speech from non-native speakers may pose a challenge for ASR

systems, because the sounds produced by non-native speakers may differ from the sound set on which the model was initially trained. This is one of the most relevant differences with the humans' speech processing system: while listeners can flexibly adapt to natural variability, ASR systems struggle to transcribe words accurately (Birulés-Muntané & Soto-Faraco, 2016; O'Shaughnessy, 2024; Jurafsky & Martin, 2025). Additionally, speech from speakers may be collected using omnidirectional microphones[36] inside rooms with poor acoustics (e.g., background noise and/or reverberation), further degrading the quality of transcriptions.

These factors may also have an impact on the degree of confidence the ASR system has in its transcription. *Confidence scores* - numerical values ranging from 0 to 1 that indicate how sure the system is that the transcribed word matches what the speaker actually pronounced (Jiang, 2005; Vertanen & Kristensson, 2008) - are typically not visible in the textual output delivered to users, but they are helpful to engineers to spot some of the weak points of the ASR systems they build (Li, 2018). However, displaying confidence through specific fonts or colors may also enhance diverse users' reliability on the transcriptions, making it easier to decide whether to trust the output of the ASR system (Wald & Bain, 2008). Previous research on the usefulness of showing the ASR system's confidence to users has yielded mixed results. For example, while deaf and hard-of-hearing participants in Shiver and Wolfe's study (2016) benefitted from the color-coded markup in the ASR-generated captions (measured in higher comprehension rate compared to the 'no captions' baseline condition), in 2017 Berke and colleagues (who developed a series of color-coded markup schemes to signal to deaf and hard-of-hearing persons the confidence the ASR system had in its transcriptions in one-on-one meetings) found no statistically significant effect of markup across different conditions.

In this chapter, we present the second study of the project. We will discuss the results of a corpus analysis aimed at assessing 1) the performance of an ASR system in real-world settings with a focus on academic lectures delivered in English, 2) whether confidence scores can be a reliable metric to build display formats to signal users how sure the system is of its transcriptions, and 3) which speaker- or environmental-related factors affect the most accuracy of transcription and confidence values. The analysis results led us to develop two experimental markups that were later presented to participants in the third experiment of this research project (see Chapter 5). To the best of our knowledge, this study represents the first attempt to analyze the reliability of the confidence score metric with the aim

[36] This type of microphones can collect sounds from whichever direction and don't need to be pointed in a particular direction. On the contrary, (uni)directional microphones are sensitive to sounds coming only from the front, not picking up ambient noise. Source: Shure (2025). Microphone directionality and polar pattern basics. https://www.shure.com/it-IT/performance-produzione/louder/microphone-directionality-polar-pattern-basics .

to develop experimental display formats for live captions to be used in real-life settings - specifically, during academic lectures delivered in English.

4.2 Objectives and research questions

The study aimed to investigate

- The performance of an ASR system by evaluating 1) its accuracy and 2) the confidence the system has in its output when transcribing speech collected in real-world settings.
- The reliability of confidence scores as a measure for developing a set of graphical features (color-coded markups). These markups would visually represent the confidence level of the system in its transcriptions within the text of the automatic captions.
- The influence of specific factors (e.g., the characteristics of individual speaker(s) and/or environments) on both the accuracy of the transcriptions and confidence values.

We therefore aimed to address the following research questions:

RQ1. How well does a traditional, speaker-independent ASR system perform when dealing with the transcription of audio recordings collected in various real-world contexts?

RQ2. Which speaker- and environment-related factors affect accuracy of transcriptions and confidence values the most?

RQ3. Is there a correlation between confidence scores and the type of transcription (correct/erroneous)?

RQ4. Is there a threshold value determining that a certain correlation is always true (e.g., a specific low value corresponds 100% with an erroneous transcription)?

Engineers are constantly working to improve the robustness of the architecture of ASR systems to enhance their performance (e.g., O'Shaughnessy, 2024). However, the effectiveness of these systems depends not just on how their architecture is designed, but also on various factors related to the speakers and their surrounding environment (Pucci, 2023). Previous research has shown that several

factors significantly affect the performance of ASR systems, with *spontaneous speech* being one of the most critical contributors to transcription accuracy. In fact, spontaneous speech is characterized by several phenomena, including connected speech, variations in speech rate, false starts, and hesitations. These linguistic phenomena can significantly impact the structure of a speech signal, making it challenging for ASR systems to analyze the signal itself, recognize the lexical elements in it, and produce an accurate transcription (e.g., Benzeghiba *et al*., 2007). As a result, these phenomena also influence confidence values, which indicate the reliability of transcriptions (or the probability of correctness for a specific lexical item or sentence - Jiang, 2005). However, when presented with read speech, ASR systems tend to perform more effectively, resulting in a higher rate of accuracy, higher confidence scores, and lower word error rate (WER) scores (Nakamura *et al*., 2008). But why is this the case? The difference lies in the distinctive characteristics of the two types of speech. Read speech often displays fewer irregularities that can complicate the recognition process, making it acoustically and linguistically easier for the ASR system to match the elements in the signal with those in its architecture (Furui, 2003; Nakamura *et al*., 2008).

In a context such as the educational one, teachers and professors mainly engage in semi-spontaneous speech during their lectures (since they may read the text included in their slides while speaking). Naturally, this type of speech retains the characteristics commonly found in spontaneous speech. As a result, speakers can make speech errors, experience false starts, and hesitate while speaking, even though they have a written script to refer to. Similar to spontaneous speech, these factors can affect the accuracy of transcription, ultimately impacting the performance of ASR systems (Benzeghiba *et al*., 2007). It is therefore essential to evaluate the performance of ASR systems in these settings if institutions want to make sure to aid language processing and guarantee access to information for diverse populations. For this reason, we analyzed a corpus of transcriptions generated by the traditional, independent-speaker ASR system provided by *Cedat85*[37] - the partner company for this doctoral research project. To assess the performance of this ASR system (both in terms of accuracy and reliability of its written output), we compared the transcription of audio recordings from two distinct domains (or *general topics*): *education* (specifically, academic lectures) and *politics* (political interventions and debates). While academic lectures are characterized by the use of semi-spontaneous speech, political interventions and debates are examples of read speech, as politicians frequently read from prepared scripts. Therefore, the ASR system may perform better in transcribing audio recordings from politics (read speech) compared to when transcribing academic lectures (semi-spontaneous

[37] The partner company is a leader in the ASR sector, and it provides automatic transcriptions and captions services to private companies, public institutions (like the Italian, British, and European Union Parliaments), and universities for various purposes (e.g., to generate summaries of meetings or increase accessibility of information during live events).

speech). On the contrary, if the ASR is sufficiently robust, it should demonstrate comparable performance in terms of accuracy and confidence scores across both domains.

To further investigate the performance of the ASR system in real-world settings, we compared the set of transcriptions by *language* (English, Italian). Previous research on accessibility has emphasized that accuracy is essential for effectively aiding speech processing, content comprehension, and access to information for diverse populations (e.g., Gernsbacher, 2015; Butler *et al*., 2019; Chan *et al*., 2019; Romero-Fresco & Fresno, 2023). Among many others, for instance, a study conducted by Butler and colleagues in 2019 explored the advantages of providing live automatic captions to deaf and hard-of-hearing students in higher education classes. Findings revealed that participants considered it extremely important that this audiovisual translation product should be highly accurate and display easily readable text (in terms of graphical features) to have access to the information delivered in class by professors. Another study on the use of automatic captions to assist L2 speakers' speech processing and content comprehension indicated that the most acceptable maximum error threshold to aid speech comprehension is attested at a 20%-word error rate (Shimogori *et al*., 2010). Similarly, Chan and colleagues (2019) pointed out that an accuracy rate of approximately 70% is too low for automatic captions to be considered useful (p. 257). They observed that the transcription errors significantly hindered content comprehension, which was evident in the comprehension task results. Therefore, to ensure access to information for diverse populations, it is essential to evaluate this ASR system by assessing its performance when transcribing speech delivered in English and comparing these results with those in Italian (the primary language of instruction at Ca' Foscari University of Venice). Again, if the system is robust enough for use in these different settings, it should demonstrate similar performance when transcribing audio recordings in both languages, with a minimum threshold of acceptable accuracy rate above 80% to effectively support speech processing and access to information.

It's important to acknowledge that real-world contexts are often imperfect. For instance, unlike other instructional settings, academic lectures can take place both in smaller classrooms and larger lecture halls. Unfortunately, these spaces may not have been accurately designed acoustically, or they may have been spaces that were not primarily intended to host academic lectures. As previously mentioned, various acoustic factors can intervene and alter speech signals, which can impact individuals' ability to perceive speech (e.g., Van Den Heuij *et al*., 2018). Additionally, these factors can make it difficult for ASR systems to perform the recognition and transcription tasks accurately enough by altering the speech signals, therefore lowering the accuracy and reliability rates of ASR-generated captions/transcriptions (Benzeghiba *et al*., 2007). For example, the number of speakers in

a room and the overlap of their voices could affect the speech signals, ultimately impacting the recognition and transcription processes of the ASR system. Additionally, speakers' characteristics associated with natural variability - such as gender, age, and linguistic profile - could influence the performance of ASR systems, regardless of the language and topic being discussed.

In sum, we compiled a list of factors that could potentially affect the performance of the ASR system. These factors were then included in an analysis which aimed to determine which one had the greatest impact on both transcription accuracy and probability of correctness - that is, confidence scores. Our main factors of interest were the following:

Environment-related factors

- Number of speakers
- Overlapping voices (two or more speakers talking at the same time)

Speaker-related factors

- Gender
- Native speaker of a language
- Speaking/reading rate

Finally, this overall analysis aimed to assess if the potential correlation between the types of transcription (correct, incorrect) and confidence scores (see RQ3 and RQ4). In this case, we expect that higher confidence values would correspond to a greater probability of correct transcriptions, and conversely, lower confidence values would indicate a lower probability of accuracy. However, there are instances where unigrams and bigrams with high confidence scores may be transcribed incorrectly, while those with low confidence scores may be transcribed correctly. Therefore, we do not anticipate a specific threshold value that guarantees consistent correlation.

The outcome of this study was meant to lead to the creation of two experimental graphical markups to provide users with information about the confidence level of the ASR-generated transcription. The potential effects of the different markups on content comprehension and attention were investigated in a study with L2 speakers of English during the third phase of this project (see Chapter 5).

4.3 Methods

4.3.1 Materials

4.3.1.1 Overall test set

In this section we describe the main characteristics of the corpus of audio recordings (henceforth, *test set*) we used in this study to answer our research questions.

Overall, the test set contained four hours of spoken input[38] with different characteristics, for a total of fifteen audio recordings with different lengths. All audio files were recorded using a unidirectional microphone to ensure high quality of the speech signal and minimize ambient noise.

The main independent variables were *topic* (*Education*, *Politics*) and *language* (*English*, *Italian*). To enhance comparability, we counterbalanced the "size" of each group according to the cumulative duration of the audio recordings per variable, dividing the test set into four subsets (two for each variable).

For both *languages*, roughly two hours of speech in English and two hours of speech in Italian was included in the test set. *Figure 28* shows the distribution of tokens for each language in the test set. The histogram shows that the English subset was made of 17396 tokens - accounting for 52,2% of the entire test set – while the Italian subset contained 15946 tokens – 47,8% of the entire test set.

[38] The partner company typically uses a four-hour training set to test its ASR system before officially implementing and releasing new languages to the public.

Figure 28. *Distributions of tokens by language spoken by the speaker(s) in the audio recording (English, Italian).*

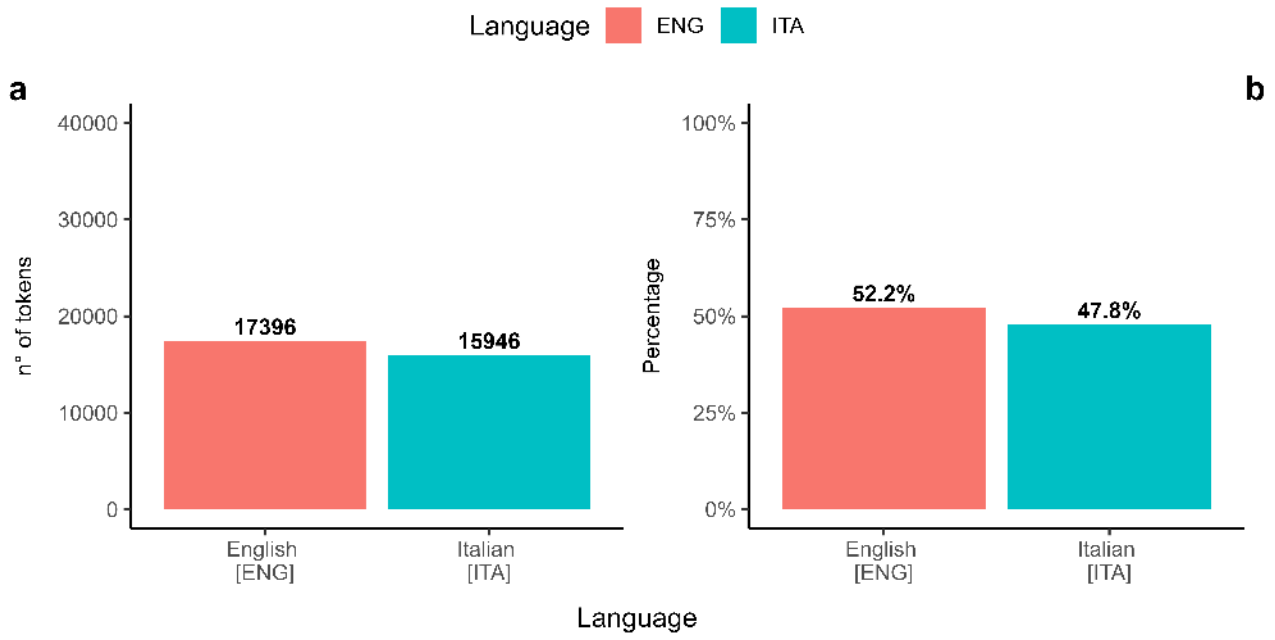
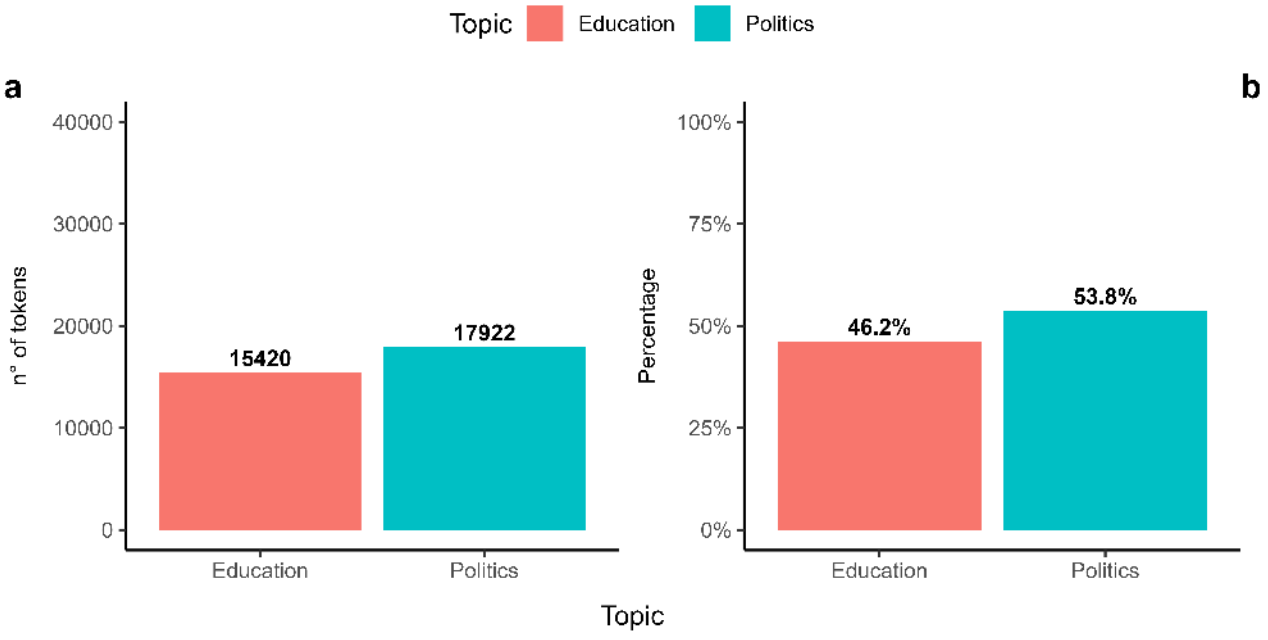


Figure 29 shows the distributions of tokens for the *topic* variable. The audio recordings for the *Education* topic have a total of 15420 tokens (46,2% of the entire test set), while the audio recordings for the *Politics* topic have a total of 17922 tokens, representing 53,8% of the tokens in the entire test set.

Figure 29. *Distributions of tokens by topic (education, politics).*



Overall, the test set has a slightly higher number of tokens in the *English* and *Politics* subsets compared to the number of tokens in the *Italian* and *Education* subsets.

In addition to the main independent variables, the audio recordings were categorized based on specific characteristics related to the speakers and the environment. We selected those characteristics that we believed could potentially influence the performance of the ASR system. Here we list the relevant factors:

- Number of speakers talking in the audio recordings.
- Overlapping voices (if two or more speakers talked simultaneously).
- Overall quality of the speech (if the audio recording contained environmental noise, overlapping voices, and/or distance from the recording device).
- Speaker(s)' gender.
- Native speaker (if the talker/s were native speakers of English/Italian).
- Speaking/Reading rate (measured in word per minute, wpm).

These variables of interest were used to analyze our test set and answer RQ3.

The following section reports a detailed description of each audio recording included in the test set.

4.3.1.2 Characteristics of each audio recording

Overall, fifteen audio recordings were included in the test set. The test set contained:

1) Eleven audio recordings where speakers of *English* talked about *politics* (from REC_001 to REC_011). The audio files contained some excerpts (range: 1.01 – 27.51) of political interventions and debates at the European Union and British Parliaments. The main topics of this group of recordings were about welfare, domestic and international politics. Five recordings (REC_001, REC_008 – REC_011) had only female speakers, four recordings (REC_002 – REC_005) had only male speakers, and two recordings had multiple speakers of different genders (REC_006, REC_007). Reading speed was attested in a range comprised between 122 and 207 (average reading aloud rate for native speakers of English is attested at 183 wpm – Brysbaert, 2019). The majority of individuals talking in the recordings where

native speakers of English (N = 8). Lastly, there were occurrences during which two or more speakers talked at the same time (two out of eleven recordings).

2) Two audio recordings where speakers of *Italian* talked about *politics* (REC_012 and REC_013). In these audio files, speakers read their interventions on welfare and domestic politics during debates at the Italian Parliament. In both recordings, both female and male native speakers of Italian took turns to read their interventions (reading rate range: 141 – 144).

3) One audio recording where a speaker delivered a *lecture* in *English* (REC_014). In this audio recording, a female professor (native speaker of English) delivered a lecture on English Phonetics and Phonology at the Ca' Foscari University of Venice (Italy). She had a slow pace while delivering the lecture (wpm = 105) and occasionally read the text in some slides. At some point, a female student asked a question of clarification: however, she was very far from the microphone. For this reason, only a few words were recorded and subsequently transcribed by the ASR system.

4) One audio recording where a speaker delivered a *lecture* in *Italian*[39] (REC_015). Similarly to the previous recording, a female professor (native speaker of Italian) delivered in Italian an introductory lecture to General Linguistics at the Ca' Foscari University of Venice (Italy). She delivered the lectures at a slightly higher pace compared to the professor in REC_014; however, it is not clear from the video if she read some text or if slides were provided during the lecture.

Lastly, two of the audio files – namely REC_007 (*Politics*, *English*) and REC_014 (*English*, *Education*) - contained speech that potentially was affected by a combination of factors (environmental noise, overlapping voices, and/or distance from the recording device). These issues may have contributed to degrading the speech signal during the recording. As a result, these two recordings were classified in the "degraded audio quality" subset, while the remaining audio files were placed in the "clean audio quality" subset.

All audio recordings concerning *Politics* (REC_001 – REC_013) and the audio recording of the academic lecture delivered in Italian (REC_015) were sourced on the Web by the research team of the partner company, while the academic lecture delivered in English (REC_014) was recorded during a testing period in class of the ASR system of the partner company (February – May 2022).

[39] youcafoscari (YouTube channel). (2020, October 13). *La prima lezione di Linguistica generale — Alessandra Giorgi* [Video recording]. https://www.youtube.com/watch?v=z5LYxtVIg4I

Table 23 includes the main characteristics for each audio recording included in the test set, while *Table 24* reports details regarding the specific topic discussed in each audio recording. Lastly, *Figure 30* below shows the distribution of tokens in each audio recording of the test set.

Table 23. *Main characteristics of the tracks in the audio files included in the test set.*

Recording ID	Recording duration	N° Tokens (REF)	Total n° speakers	Speaker(s)' gender	Speakers' language	NS?	Words per minute (WPM)[40]	Voices	Audio Quality
REC_001	15.41	1883	1	F	English	N	122	N	Clean
REC_002	1.28	222	2	M (2)	English	Y	173	Y	Clean
REC_003	1.30	209	1	M	English	Y	161	N	Clean
REC_004	1.01	164	1	M	English	Y	162	N	Clean
REC_005	1.19	200	1	M	English	Y	168	N	Clean
REC_006	1.21	184	3	F (1), M (2)	English	N	152	N	Clean
REC_007	27.51	4983	3+	F (15), M (26)	English	Y	181	Y	Degraded
REC_008	1.27	296	1	F	English	Y	233	N	Clean
REC_009	4.15	587	1	F	English	Y	141	N	Clean
REC_010	2.22	387	2	F (2)	English	N	174	N	Clean
REC_011	1.52	315	2	F (2)	English	N	207	N	Clean
REC_012	29.39	4152	3+	F (1), M (5)	Italian	Y	141	N	Clean
REC_013	30.04	4340	3+	F (2), M (3)	Italian	Y	144	N	Clean

[40] The reading and speaking rate (words per minute, wpm) is defined as "the number of spoken units per unit time" (Dowding *et al*., 2024: p.) was calculated using the following formula: $\frac{n° \, Tokens \, REF}{Recording \, duration}$

On the one hand, in his review and meta-analysis, Brysbaert (2019) estimated that the oral reading rate for native speakers of English is 183 wpm, while it is lower for readers with English as second language. On the other hand, the average speaking rate during lectures delivered in British English is comprised in the range 125-160 wpm (Tauroza & Allison, 1990). In general, both average speaking and reading rates are variable and depend on many factors – including linguistic, sociolinguistic, psychological, and physiological ones - related to the speakers (Yuan *et al*., 2006; Brysbaert, 2019).

REC_014	76.05	7966	2	F (2)	English	Y	105	N	Degraded
REC_015	60.18	7454	1	F	Italian	Y	124	N	Clean

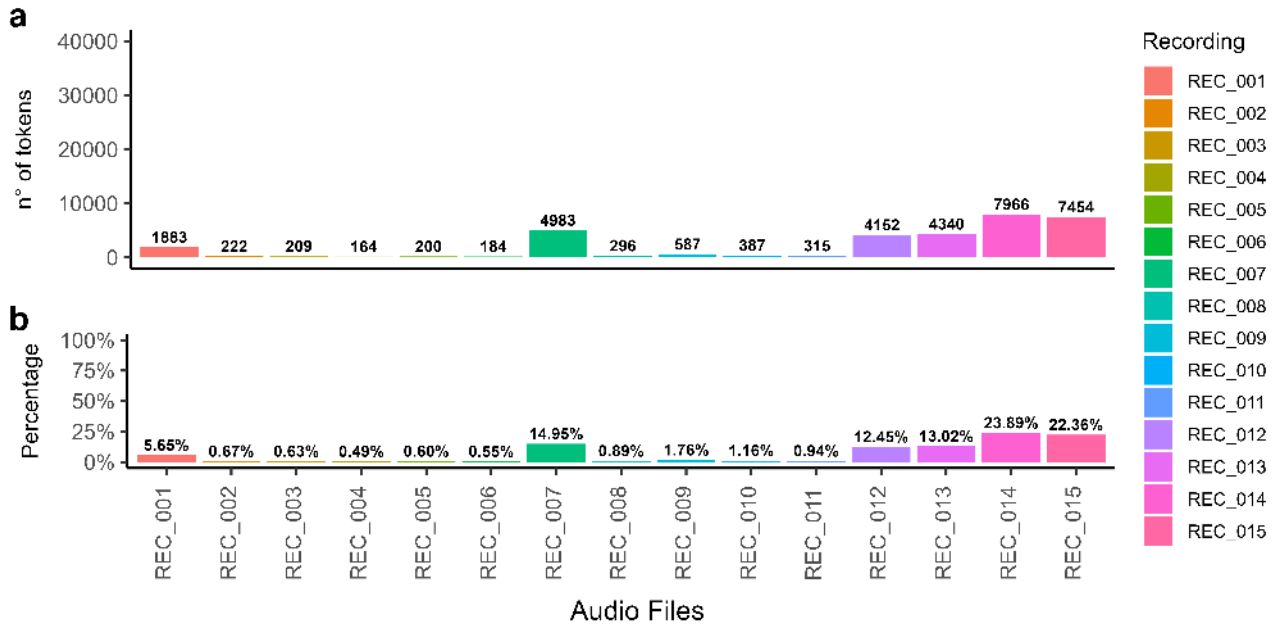
Note. For each audio file, in order, we report their length (in minutes), the number of tokens in the reference (REF) text, the number of speakers in the audio recording (1, 2, 3, or more than 3 – 3+), the speaker(s)' gender (F, M or F,M) and the precise number of persons speaking in the audio recording grouped by gender, the language spoken by the speakers in the recording (English/Italian), if those who talked in the audio recording was a native speaker or the language s/he spoke (*NS?* column: Yes, No), the speech/reading rate (WPM), if people talked over each other (*Voices* column: Yes, No), the quality of the audio based on environmental and acoustic characteristics of the audio file (Audio quality column: Clean, Degraded).

Table 24. *General topic, genre, and topic of each recording in the test set.*

Recording ID	General topic	Genre	Topic of the recording
REC_001	Politics	Intervention	State of the Union 2021
REC_002	Politics	Intervention	Discussion on fishing vessels (EU-UK)
REC_003	Politics	Intervention	Discussion on fishing vessels (EU-UK)
REC_004	Politics	Intervention	Right to vote for every EU citizen - European elections
REC_005	Politics	Intervention	Discussion on fishing vessels (EU-UK)
REC_006	Politics	Intervention	Technology and data protection in the EU
REC_007	Politics	Debate	Various topics concerning welfare in UK
REC_008	Politics	Intervention	Discussion on policies concerning farmers and new technologies (EU)
REC_009	Politics	Intervention	Gender equality
REC_010	Politics	Intervention	EU policies after the COVID-19 pandemic
REC_011	Politics	Intervention	International Women's Day
REC_012	Politics	Debate	Various topics concerning welfare and domestic politics in Italy

REC_013	Politics	Debate	Various topics concerning welfare and domestic politics in Italy
REC_014	Education	Lecture	Lecture on English Phonetics and Phonology
REC_015	Education	Lecture	Introductory lecture to General Linguistics

Figure 30. *Distribution of tokens of the test set (Figure a: number of occurrences; Figure b: percentage of occurrences), divided by recording.*



Note. The percentage of occurrences for each audio recording was calculated out of the total number of tokens in the test set.

4.3.2 Procedure and analysis

This section details the pre-processing steps done to prepare the data file and a brief description of the preliminary analyses conducted on the test set.

4.3.2.1 Pre-processing steps: data file creation and scoring

Pre-processing of the data began with the creation of reference (REF) and hypothesis (HYP) texts. I created REF texts for audio files in the *Education* subset, while the REF texts in the *Politics* subset were manually created by professional transcribers working with *Cedat85*, the partner

company. The *hypothesis* (HYP) texts were generated using the ASR system of the partner company, which is a speaker-independent engine based on a traditional hidden Markov statistical model (Palmerini & Savy, 2014). The company also provided us with the *html* files containing the confidence score values for each token (unigram or bigram – see below for the definition) transcribed by the system for each audio recording of the test set. Each token in the HYP texts was printed in different colors based on the confidence score the ASR system assigned to it. An IT technician from the research team at the partner company arbitrarily assigned a color to a range of values (*Table 25*): ~~red~~ for tokens printed with a confidence score lower than 69,9%, red for tokens printed with a confidence score between 70% and 89,9%, grey for tokens printed with a confidence score between 90% and 99,9%, and **black** for tokens printed with a confidence score of 100%. Each range was then labeled 'high', 'mid', 'low' based on how close the confidence score values were to the maximum value of the scale (100%). For this reason, both **black** and grey ranges were assigned the 'high' label, while the red and ~~red~~ ranges were respectively assigned the 'mid' and 'low' labels. This labeling allowed us to later assess the potential correlation between the transcription type (correct, incorrect) and the values within these ranges (see section 4.4.3). In agreement with the research team at the partner company, we established the threshold values for each range, along with the corresponding labels ('high', 'mid', 'low'), based on their experience with the ASR technology. This decision was made because there is currently no existing standard for these ranges or labels.

Table 25. *First version of the color-coded markup.*

Confidence Score Range	Color-coded Label	Range Labels
100%	**Black**	High
99,9% - 90%	Grey	
89,9% - 70%	Red	Mid
≤ 69,9%	~~Red~~	Low

Note. Each confidence score range was graphically represented with a color (**black**, grey, red, ~~red~~) and a label (high, mid, low). Both elements were defined arbitrarily by the research team from the partner company based on their experience in the field.

Texts were then aligned to perform the comparative analysis using the *sclite* tool (a component of the package *Speech Recognition Scoring Toolkit*[41] built by the *National Institute of Standards and Technology*[42]). The *sclite* tool was also used to detect the differences between the REF and the HYP texts and count errors in the transcriptions - namely substitutions, deletions and insertions. Based on this counting, the tool computed the *word error rate* (WER) for each HYP text associated with each audio recording of the test set.

After aligning the REF and HYP texts of each audio recording, we created an Excel file where each cell contained a token. For the sake of this analysis, we define a token as a linguistic element (single word – unigram; two words - bigram)[43] that was assigned a confidence score value by the ASR system and could be retrieved from the *html* file. For each token, we reported the following information (*Table 26*):

Table 26. *Structure of the Excel file used to conduct the corpus analysis.*

Label	Value	Description
Filename_Coded	REC_XXX	Acronyms used to identify the single audio files during the analysis of the test set. The acronym contained the word REC (that stands for Recording) and a sequential number (from 1 to 15).
NoOfSpeakers	1, 2, 3, 3+	Number of speakers in each audio recording.
SpeechType	Speaking, Reading	The term specifies if the speaker(s) in the audio recordings were engaged in semi-spontaneous speech or if they read a script.
OverlappingVoices	Yes, No	Value that signals the presence of speakers talking simultaneously in the audio recordings.
SpeakerGender	Female, Male, Mixed	Gender of the speaker(s).
NativeSpeaker	Yes, No	This value signals if the speakers(s) in the audio recordings are native speakers of the language in which they are talking.
SpeakingRate	Various values	Exact value of the speaking/reading rate (wpm, see *Table 23* for details).

[41] https://github.com/usnistgov/SCTK
[42] https://www.nist.gov/
[43] Examples of bigrams: Italian, “l’impegno” (DET + NOUN); English, “he’s” (PRON + VERB).

SpeakingRateCoding	Very slow, slow, average, fast, very fast	Speaking/Reading rate categorization on the basis of Tauroza and Allisson (1990) table.
AudioQuality	Clean, Degraded	The quality of the signal is determined by environmental factors, such as distance from the microphone or overlapping voices of multiple speakers.
Topic	Education, Politics	[**Independent variable**] Main topic of the speech contained in the audio recordings.
Genre	Intervention, Debate, Lecture	Genre of the audio recording.
Language	English, Italian	[**Independent variable**] Language spoken by the speakers.
REF	Different uni/bigrams	Uni/bigrams contained in the REF text file/aural speech.
REF_POS	Part-of-speech tags	Classification by universal part-of-speech tags for each token in the REF.
REF_POS_Palm	Part-of-speech tags	Classification by part of speech for each token in the REF based on the part-of-speech tagging process done by Palmerini & Savy (2014).
HYP	Different uni/bigrams	Uni/bigrams contained in the HYP text file/aural speech.
ConfScore	Value in the range 0 - 1	Confidence score value for each token(s) in the HYP files.
ConfColorCoding	**black**, grey, red (red_underlined), ~~red~~ (red_bar)	Color-coded markup labels in which the token(s) was displayed in the output text.
ConfCoding	High, mid, low	Labels assigned to the ranges of the confidence scores. Used to conduct the correlation analysis (see section 4.4.3).
ErrorWER	Correct, Incorrect	Transcription type based on the WER labeling carried out automatically by the *sclite* tool.
ErrorWERCoding	0, 1	Scoring of transcription type (0 for erroneous transcription/1 for correct transcription) carried out manually by the researcher based on the WER labeling.
ErrorTypeWER	None, SUBstitution, INSertion, DELetion	Error type (None: tokens transcribed correctly, SUB, INS, DEL) based on the WER labeling carried out automatically by the *sclite* tool.
TrueError	True, False, Not Counted	Assessment of true errors vs. false errors of transcription due to discrepancies in the formatting of the text (see section 4.4.1.1).

Note. The column "*Label*" contains the name of the columns in the Excel file used to conduct the corpus analysis, the column "*Value*" contains the data for each label inserted in the cells of the data file, and the column "*Description*" contains a brief description of the elements of each column.

After filling out each cell in the data file, the scoring for the accuracy of transcription was carried out. On the one hand, we assigned the score 0 to each token that was transcribed erroneously; on the other hand, we assigned the score 1 to each token that was transcribed correctly. Values were inserted in the *ErrorWERCoding* column.

Lastly, we performed two types of part-of-speech (PoS) tagging on the REF texts to qualitatively analyze the data from a linguistic perspective. The first type of PoS tagging was performed automatically using the *spaCy* package (Honnibal *et al*., 2020) on *Google Colab*[44] (see *Appendix E* for the Python syntax). Each token in the test set was assigned a tag, which was recorded in the *REF_POS* column (see *Table 26* for details) of the Excel file. After this first stage, the tags in the *REF_POS* column were manually revised to ensure the assignment of the correct labels, particularly for bigrams. This initial type of PoS tagging was carried out automatically using *spaCy*[45] to facilitate the grouping of the tokens in the subsequent manual PoS tagging. In this second phase, tokens were manually grouped by the tags reported in the study conducted by Palmerini and Savy in 2014 (which included the following categories: *noun*, *verb*, *adjective*, *adverb*, *function word*, and *other*[46]). In addition to the existing tags, we included the category *Composite*, which included all those tokens that the ASR system printed out as bigrams (English - e.g., VERB + NEG, "don't"; Italian – e.g., DET + NOUN - "l'articolo", translation: "the article"). The linguistic analysis of the tokens in the current test set was conducted using this second PoS tagging method. The analysis of accuracy and the probability of correctness for specific grammatical categories aimed to determine which categories should be included in or excluded from the new color-coding schemes. However, since we only adjusted the ranges of confidence scores and the color schemes, and then tested these two new markups in the third study (see Chapter 5), we chose not to include the results of this analysis in this chapter.

4.3.2.2 Preliminary Analyses

After the creation of the Excel file with the data relevant to our analysis and before running the inferential statistical analyses, we conducted a descriptive analysis of the performance of the ASR

[44] Google. (2024). Google Colaboratory. Retrieved December 5, 2024, from https://colab.research.google.com/.
[45] *SpaCy* automatically assigned a label from the universal PoS tags list to each token (Source: Universal POS tags. https://universaldependencies.org/u/pos/).
[46] Following Palmerini & Savy (2014), this category comprised disfluent speech and false starts that resulted in nonwords or incomplete words.

system in the transcription of the audio recordings in the test set. Before moving to the discussion of these results, we briefly present the method we used to calculate the word error rate (WER).

4.3.2.2.1 Word Error Rate

We calculated the *word error rate* (WER) using the Python syntax via the platform HuggingFace[47]. The syntax compared the text in the REF and in the HYP and calculated automatically the WER using the standard formula:

$$\text{WER} = \frac{SUB+DEL+INS}{N}$$

The sum of substitutions, deletions and insertions divided by the total number of words in the REF (N) gives back a number comprised between zero and one: the lower the result, the higher the accuracy of the transcription generated by the ASR system (Jurafsky & Martin, 2025).

The syntax used to calculate the WER can be found in *Appendix D*.

4.4 Results

Here we report the results of the analysis of the accuracy of the transcriptions of the audio files and the confidence scores signaled by the ASR system. We also assessed how confident the ASR system was of its transcriptions by analyzing the confidence scores associated with each token of the test set. This section also contains the results of the influence of some factors – namely – *topic* and *language* – had on accuracy and the confidence scores. Finally, I will discuss how the range of values were modified to create the second version of the color-coded markup and other markups.

[47] https://huggingface.co/learn/audio-course/chapter5/evaluation

4.4.1 Complete test set: overall accuracy of transcription and confidence

4.4.1.1 Accuracy

We analyzed the overall number of tokens correctly and incorrectly transcribed by the system as a measure of accuracy.

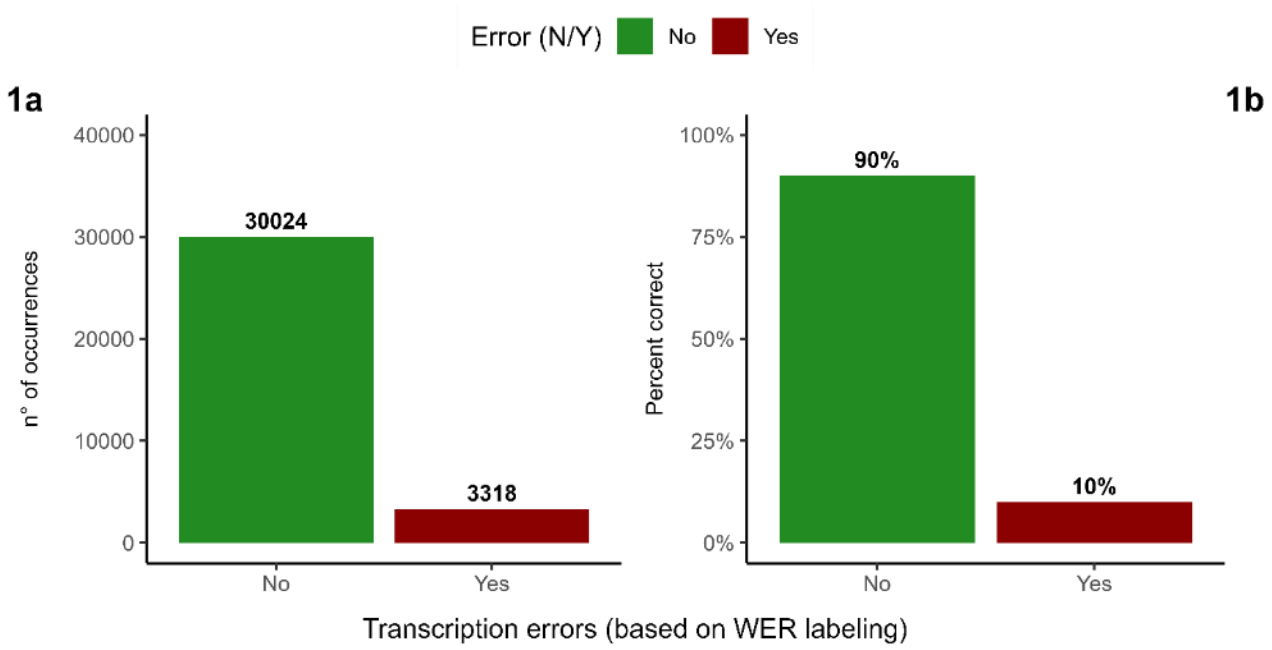
Table 27. *Count and percentage (exact value) of tokens correctly (No) and erroneously (Yes) transcribed across the test set.*

	Count (n°)	**Count (%)**
Tokens correctly transcribed	30024	90,05%
Tokens incorrectly transcribed	3318	9,95%
Total	33342	100%

Note. Scoring was assessed manually by the researcher based on the Word Error Rate (WER) labeling generated using *sclite*.

Table 27 reports the overall performance of the ASR system in transcribing the audio recordings of the test set. Out the total of 33342 transcribed tokens, the ASR system erroneously transcribed less than 10% of the tokens of entire test set ($N = 3318$). The system correctly transcribed 90,05% of the tokens ($N = 30024$) of the entire test set. *Figure 31* shows the distribution of data in two bar graphs.

Figure 31. *a) Count and b) percentage (rounded value) of tokens correctly (green bars = No – correct transcription) and erroneously (dark red bar = Yes – erroneous transcription) transcribed across the test set.*



Note. Scoring was assessed manually by the researcher based on the Word Error Rate (WER) labeling generated using *sclite*.

In the following sections, we will continue analyzing the data concerning correctly and incorrectly transcribed tokens by referring to the results given by *sclite*. However, a deeper analysis of the transcription errors revealed that some tokens in the HYP texts that were counted as errors by the *sclite* analysis *were not errors* (or "real errors", as reported in *Table 28*). In other words, the choice made by some transcribers to format the REF text differently from the ASR system led to discrepancies between the REF and the HYP texts. As a result, these inconsistencies were scored as errors by *sclite*. For instance, one of the human transcribers adopted the British English spelling system to format the REF text for one of the audio recordings in the test set, while the ASR system printed out its transcriptions using the American English spelling system[48]. The difference in spelling was counted as an erroneous transcription by *sclite*, which categorized all the tokens with this kind of formatting as substitution errors. In *Table 28,* we report the number and percentage of the tokens involved: almost 1% of the tokens counted as errors by *sclite* were false errors (or formatting

[48] E.g., the human transcribers' choice to use the British English orthographic conventions compared to the American English ones used by the ASR system ('neighbourhood' vs 'neighborhood') or the formatting choices on which the system was built on (e.g., the transcription of numbers - '3' in the human transcription versus 'three' in the ASR-generated transcription).

discrepancies - *N* = 285), while true errors were 2913 (out of 3318 total errors – 8,74%). We also included the number and percentage of "errors not counted", which refers to the tokens that human transcribers included in the REF text even though they represented disfluencies or false starts in the speech. Since these tokens were accurately or inaccurately transcribed by the ASR system probably due to the sufficient or insufficient acoustic cues that lead to a correct or incorrect prediction, we decided to exclude these instances from the "true error" count and create a separate group to report this data. In this category we counted 120 occurrences (0,36% out of the total number of tokens in the test set).

Table 28. *Number and percentage of true and false errors across the test set.*

	Count (n°)	**Count (%)**
True errors	2913	8,74%
False errors	285	0,85%
Errors not counted	120	0,36%
Tokens correctly transcribed	30024	90,05%
Total	33342	100%

Note 1. *True errors*: transcription errors committed by the ASR system and printed in the HYP; *False errors*: tokens counted by errors by *sclite*, but only differed from those in the REF in the formatting of the text; *Errors not counted* are disfluencies in the oral speech transcribed by the ASR system, scored as errors by the WER analysis. Scoring was assessed manually by the researcher based on WER labeling; *Tokens correctly transcribed*: tokens that sclite counted as tokens correctly transcribed after the comparison between REF and HYP texts.

Note 2. Percentages were calculated out of the total number of tokens in the test set.

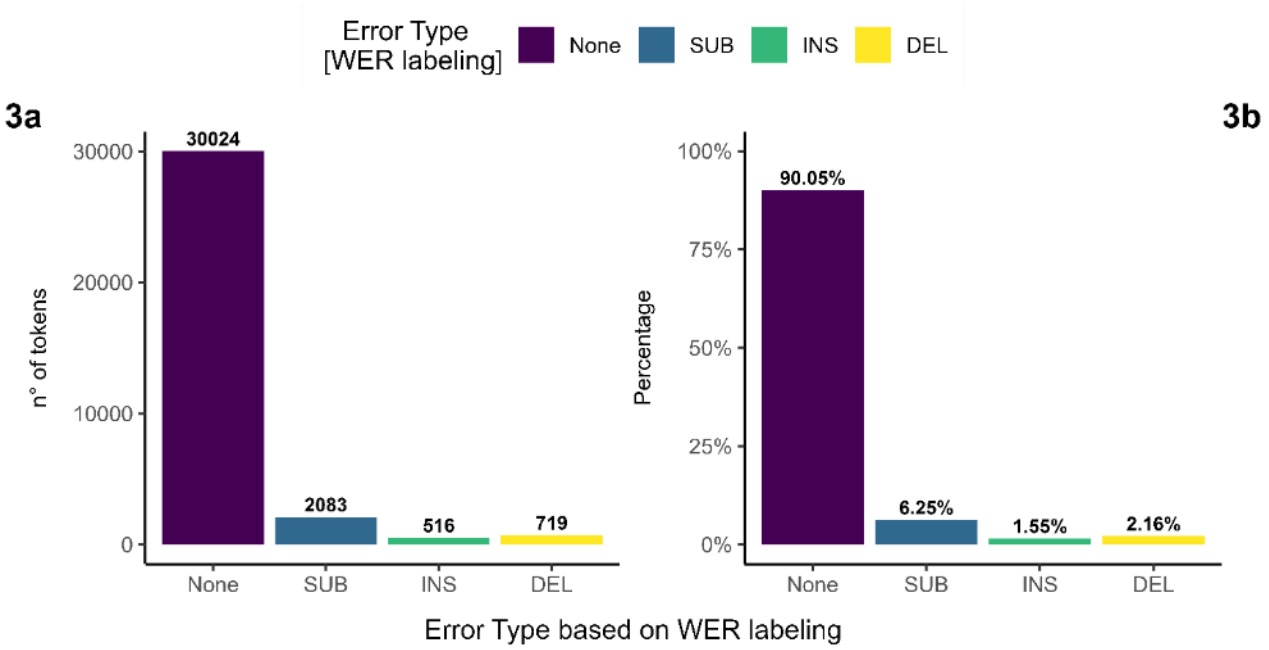
4.4.1.2 Error Type

We calculated how many deletions, insertions and substitutions were done by the ASR system when generating the transcriptions of the audio recordings. The WER analysis highlighted a high percentage of substitutions (6,25%, *N* = 2083), followed by deletions (2,15%, *N* = 719) and insertions (1,55%, *N* = 516) (*Table 29* and *Figure 32*).

Table 29. *Number and percentage of tokens correctly and erroneously transcribed across the test set.*

	Count (n°)	Count (%)
Deletions	719	2,15%
Insertions	516	1,55%
Substitution	2083	6,25%
Tokens correctly transcribed	30024	90,05%
Total	33342	100%

Figure 32. *a) Number and b) percentage of substitutions (SUB), insertions (INS), deletions (DEL) and correctly transcribed tokens (None) across the test set.*



Note 1. Percentages were calculated out of the total number of occurrences in the test set.

Note 2. None: this label refers to the tokens correctly transcribed.

4.4.1.3 Word Error Rate (WER)

Global WER score
0.1

Table 30. *WER score for the entire test set.*

The global WER score was computed using a Python script (see *Appendix D*) on *Google Colab*[49]. The global WER score (that is, the word error rate for the whole test set) is 0.1 (*Table 30*).

4.4.2 Complete test set: analysis of the confidence score values and potential correlation between the values and the accuracy of transcription

4.4.2.1 Distribution of Confidence Score values

Table 31 reports the descriptive analysis of confidence score value ($Mean_{ConfScore\_Value}$: 0,949; $SD_{ConfScore\_Value}$ = 0,11). The mean value of confidence scores indicates that overall the system is sure about its transcription.

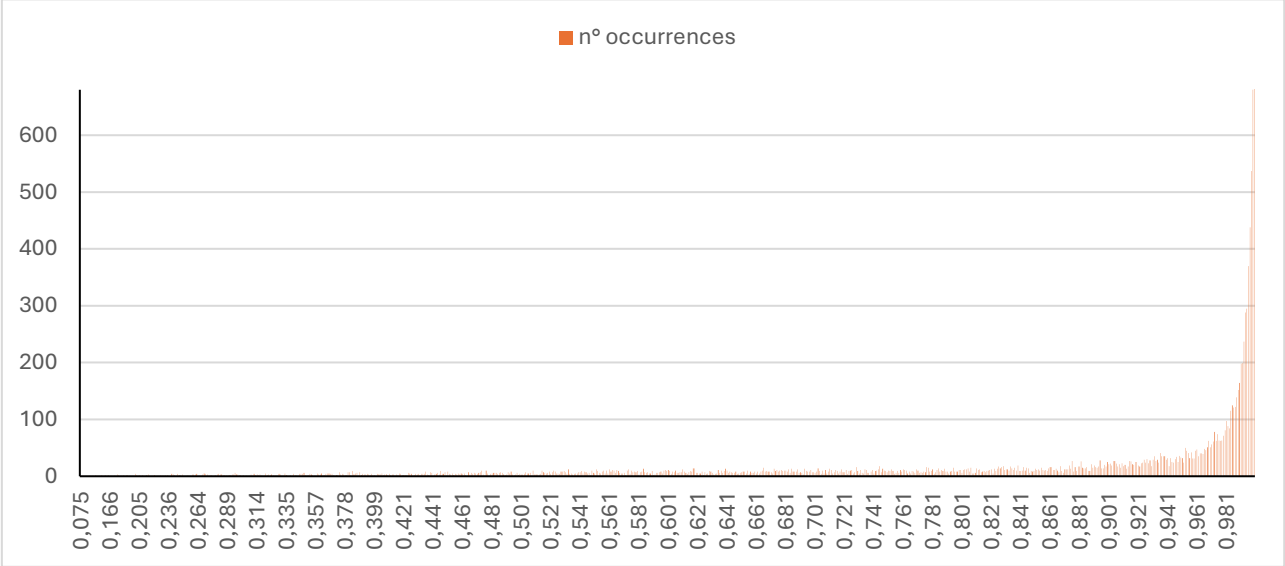
Table 31. *Descriptive statistics for the confidence scores in the test set.*

$Mean_{ConfScore\_Value}$	$SD_{ConfScore\_Value}$	$Range_{ConfScore\_Value}$
0,949	0,11	0,075 - 1

Figure 33 shows the number of tokens for each confidence score value in the test set. Data is skewed towards the higher end of scores ($Range_{ConfScore\_Value}$: 0,075-1), once again indicating that the system is sure about its transcription.

[49] Google. (2024). Google Colaboratory. Retrieved December 5, 2024, from https://colab.research.google.com/.

Figure 33. *Number of occurrences for each confidence score value across the test set.*



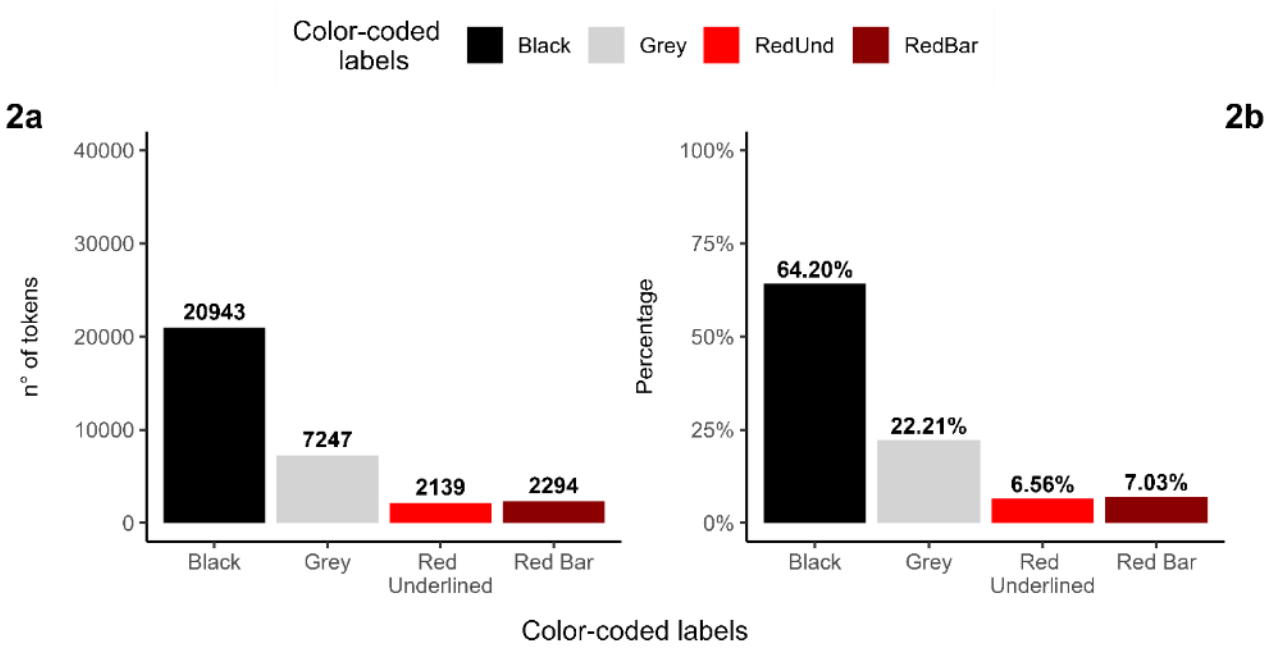
4.4.2.2 Distribution of Confidence Score values for each range

Table 32 and *Figure 34* show the distribution of tokens based on the original color-coded markup. The majority of the tokens in the dataset were transcribed with a confidence score of 1 (64%, $N = 20943$) or with a confidence score comprised in the range 99,9-90% (22,21%, $N = 7245$). Tokens in the last two ranges (89,9-70%, red and 69,9-0%, ~~red~~) were respectively 6,56% ($N = 2139$) and 7,03% ($N = 2294$) of the total. The majority of tokens have a confidence score in the range 90%-100% ($N = 28189$, 86,4% of total number of tokens) (see *Table 32*).

Table 32. *Number and percentage (exact value) of tokens in the test set in each color-coded confidence score range (**black**, grey, red, and ~~red~~).*

Confidence Score Range	Color-coded Label	Count (n°)	Count (%)
100%	**Black**	20943	64%
99,9% - 90%	Grey	7247	22,21%
89,9% - 70%	Red	2139	6,56%
≤ 69,9%	~~Red~~	2294	7,03%
Total n° tokens		32623	

Figure 34. *a) Count and b) percentage (exact value) of tokens in each color-coded confidence score range (**black**, grey, red, and ~~red~~).*



4.4.2.3 Accuracy of transcription in each confidence score range

Table 33 reports the number (and percentage) of correctly and incorrectly transcribed tokens for each color-coded confidence score range (as well as *Figure 35*). For this analysis, percentages were calculated out of the total number of tokens in each color-coded range.

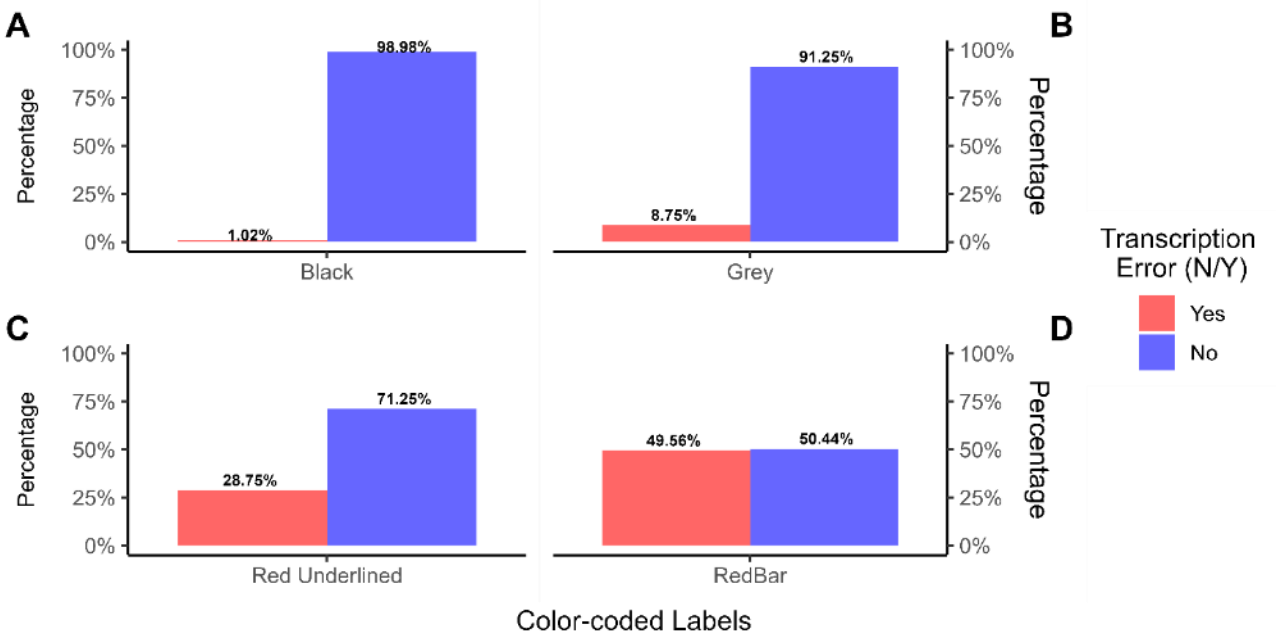
Table 33. *Number and percentage (exact value) of correct tokens and errors in each color-coded confidence score range (**black**, grey, red, and ~~red~~).*

Confidence Score Range	Color-coded Label	Error	
		Yes (n, percent)	No (n, percent)
100%	**Black**	213 (1,02%)	20730 (98,98%)
99,9% - 90%	Grey	634 (8,75%)	6613 (91,25%)
89,9% - 70%	Red	615 (28,75%)	1524 (71,25%)
≤ 69,9%	~~Red~~	1137 (49,56%)	1157 (50,44%)
Total		2599	30024

Note. Percentages were calculated out of the total number of tokens in each color-coded range.

As discussed in section 4.2, this analysis enables us to assess the reliability of each color-coded range and suggests a potential correlation between transcription type (correct, incorrect) and range confidence score. Ideally, at the highest confidence values (**black** and grey ranges), the number of errors should be very low, allowing users to trust the transcription of the ASR system almost completely. Conversely, at the lowest confidence values, the likelihood of errors should be higher: the red and ~~red~~ ranges, in fact, should warn users of the potential presence of errors in the transcription. In line with this hypothesis, *Figure 35* visually represents the distribution of errors and correct transcriptions across each color-coded range. Notably, at the highest value of confidence (indicated in **black**), the percentage of incorrect transcriptions is extremely low, with only 1% (N = 213) of the total number of tokens with a confidence score of 1 being transcribed incorrectly. In line with our hypothesis, it is clear from the graph that as confidence scores decrease, the number of errors increases. In fact, in the grey range, almost 9% of the tokens were scored as errors (N = 634). This percentage of incorrectly transcribed tokens increases significantly in the red range, where it reaches 28,75% (N = 615). The error rate peaks in the ~~red~~ range, which has the lowest values, with almost half of the tokens being transcribed incorrectly (49,61%, N = 1,137).

Figure 35. *a) Count and b) percentage (exact value) of correct tokens and errors in each color-coded confidence score range (**black**, grey, red, and ~~red~~).*



Note. Percentages were calculated out of the total number of tokens in each color-coded range.

4.4.2.4 Error Type for each confidence score range

In this section, we will focus solely on the distribution of insertion and substitution errors within the color-coded ranges. We will not include deletion errors in this analysis, as these tokens appear in the REF texts, but are missing in the HYP texts. Deletion errors occur because the ASR system fails to recognize or transcribe some lexical items that are actually spoken, and as a result, they are not printed in the HYP (and consequently, lack an associated confidence score).

Table 34 and *Figure 36* report the distribution of error types within each color-coded range. While some errors are present in the highest color-coded markup (**black**, INS: $N = 65$, 0,3% and SUB: $N = 148$, 0,7%), the majority of tokens were correctly transcribed (None – tokens correctly transcribed, $N = 20730$, 99%). *Table 34* also highlights that substitutions are the most frequent are the most frequent type of error in all color-coded ranges.

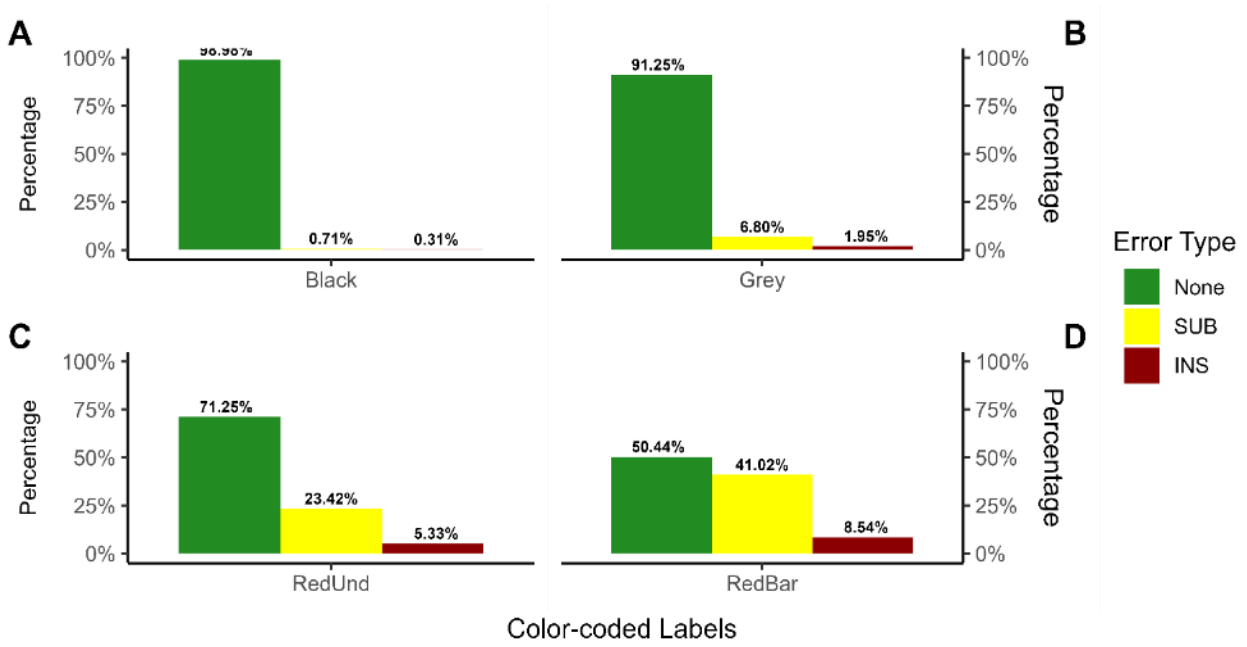
Table 34. *Number and percentage (exact value) of error type in each color-coded confidence score range (**black**, grey, red, and ~~red~~).*

Confidence Score Range	Color-coded Label	Error Type		
		Insertions	Substitutions	None
100%	**Black**	65 (0,31%)	148 (0,71%)	20730 (98,98%)
99,9% - 90%	Grey	141 (1,95%)	493 (6,80%)	6613 (91,25%)
89,9% - 70%	Red	114 (5,33%)	501 (23,42%)	1524 (71,25%)
≤ 69,9%	~~Red~~	196 (8,54%)	941 (41,02%)	1157 (50,44%)
Total		516	2083	30024

Note 1. Percentages were calculated out of the total number of occurrences in each color-coded range.

Note 2. None: this label refers to the tokens correctly transcribed.

Figure 36. *a) Count and b) percentage (exact value) of error type in each color-coded confidence score range (**black**, grey, red, and ~~red~~).*



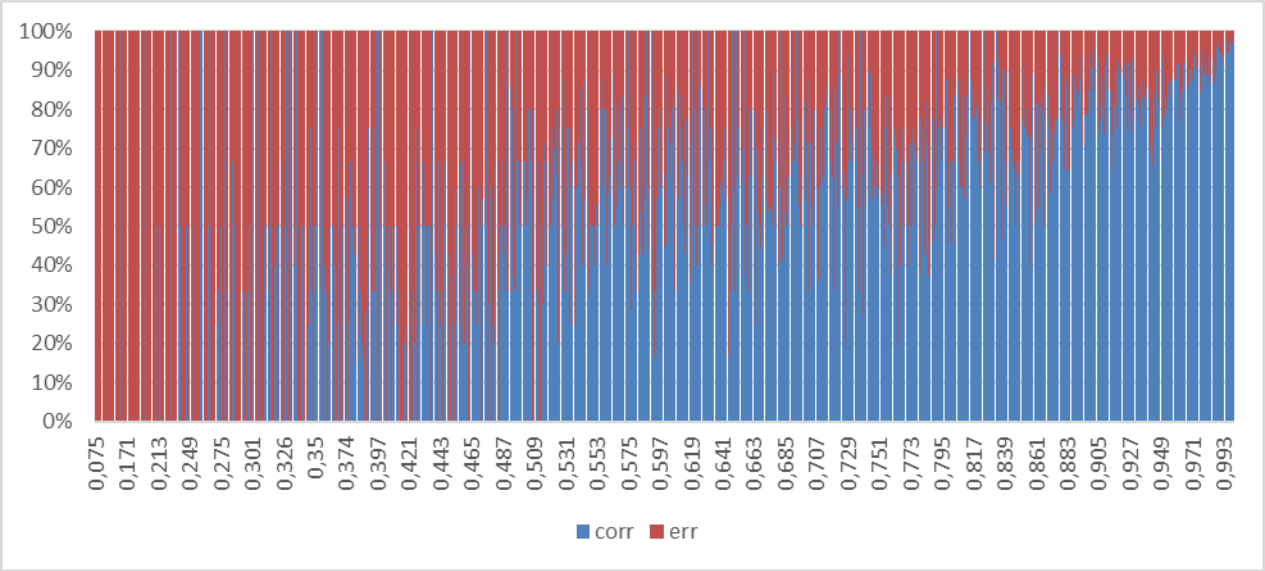
Note 1. Percentages were calculated out of the total number of occurrences in each color-coded range.

Note 2. None: this label refers to the tokens correctly transcribed.

4.4.3 Correlation between error and confidence score value

Figure 37 shows the percentage of errors (red bars) and correct transcriptions (blue bars) for each confidence score value in the test set. The graph illustrates a clear trend: there are more errors associated with lower confidence levels, as indicated by the greater number of red bars on the left side of the spectrum (closer to 0). Conversely, there is also a noticeable increase in the concentration of blue bars on the right side of the spectrum (closer to 1), representing higher confidence levels. It is evident that there remains a small percentage of errors and accurate transcriptions on both ends. This trend also suggests that it may not be possible to determine a correlation between a specific confidence value and the transcription type (correct/incorrect).

Figure 37. *Distribution of errors (red bars) and correct tokens (blue bars) for each confidence score value in the test set.*



After this first visual assessment, we investigated the potential correlation of transcription type and confidence values using the *Point Biserial Correlation*[50] method with the software Jasp[51] (v. 0.19.1). First, we calculated the correlation by analyzing the data points from the entire test set. *Table 35* shows a statistically significant, medium, positive correlation between confidence score values and the type of transcription (where '0' stands for 'incorrect' and '1' stands for 'correct'). We can interpret these results by saying that a higher confidence score indicates a greater probability that the corresponding token was transcribed correctly. *Figure 38* shows the positive correlation between the binary variable and the confidence scores.

Pearson's Correlations

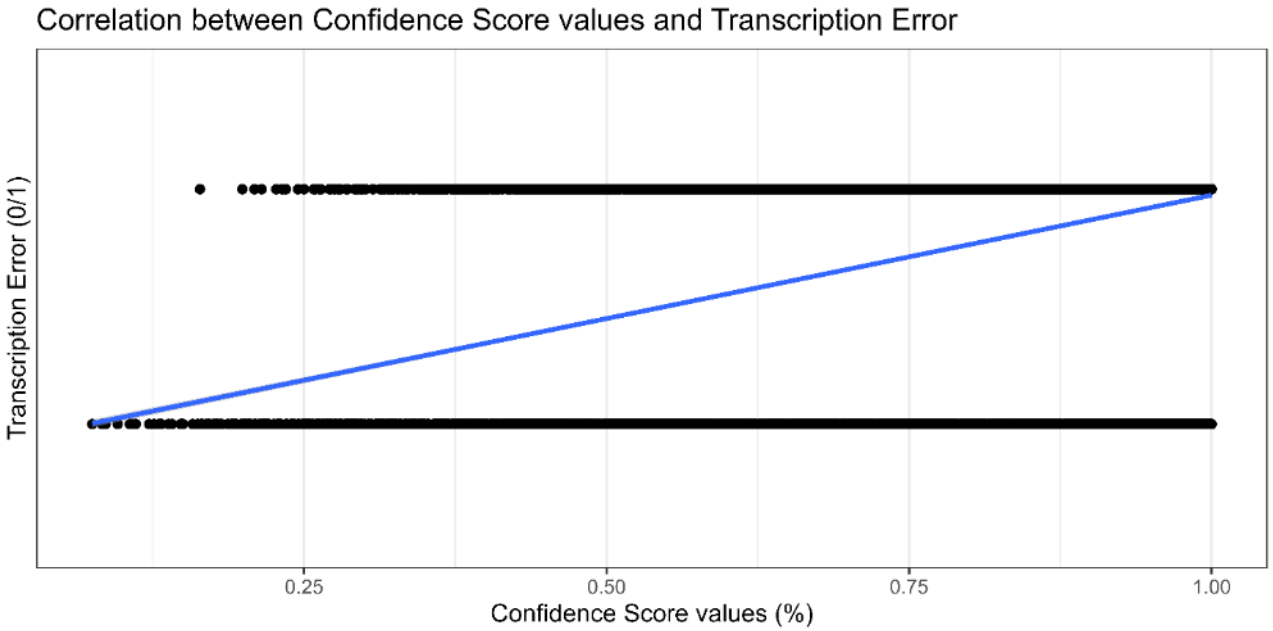
	Pearson's r	p
Confidence Score - ErrorWER Coding	0.516***	< .001

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

Table 35. *Results of the point biserial correlation analysis to assess the potential correlation of confidence values and correctness of the transcription (overall test set).*

[50] The Point Biserial correlation analysis is a variant of the *Pearson's Correlation*, and it is used when one of the two variables is dichotomous – in our case, we scored the Errors using a Yes/No scoring (Kornbrot, 2005).
[51] JASP Team (2024). JASP (Version 0.19.1) [Computer software].

Figure 38. *Correlation between confidence scores and transcription type (correct/erroneous).*



Note. Labels in the y axis (‘*Transcription Error’*): 0 stands for ‘incorrect transcription’, 1 stands for ‘correct transcription’.

We then analyzed the potential correlation between transcription type and the confidence values included within each range. This analysis aimed to evaluate the strength of the correlation: the higher the Pearson's r value, the stronger the correlation. Ultimately, this assessment would guide our decision on whether to adjust the ranges for each color-coded category in the new version of the markups or leave them unchanged (see section 4.4.5).

Table 36, *37* and *38* report the results of the analyses for the **black** + grey, red, and ~~red~~ ranges (which are subsets of the overall test set). As mentioned in section 4.3.2.1, we grouped the values in the **black** + grey ranges under the 'high' label and then performed the correlation analysis. Pearson’s r in all ranges indicates a very weak and weak statistically significant correlation between the two variables (respectively 0.233, 0.114, and 0.305).

Pearson's Correlations

	Pearson's r	p
Confidence Score - ErrorWER Coding	0.233***	< .001

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

Table 36. *Results of the point biserial correlation analysis to assess the potential correlation of confidence values and correctness of the transcription (black + grey ranges).*

Pearson's Correlations

	Pearson's r	p
Confidence Score - ErrorWER Coding	0.114***	< .001

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

Table 37. *Results of the point biserial correlation analysis to assess the potential correlation of confidence values and correctness of the transcription (red range).*

Pearson's Correlations

	Pearson's r	p
Confidence Score - ErrorWER Coding	0.305***	< .001

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

Table 38. *Results of the point biserial correlation analysis to assess the potential correlation of confidence values and correctness of the transcription (red range).*

4.4.4 Factors affecting transcription accuracy and confidence scores

In this section, we provide a descriptive analysis for each audio file to further evaluate the performance of the ASR system. This analysis also allowed us to conduct a preliminary assessment of the factors that may influence the ASR system's performance - namely, its accuracy and confidence scores (see section 4.3.1.2). Additionally, we present the results of the descriptive analysis for each variable of interest (topic, language, parts of speech), with the primary aim of assessing the performance of the ASR system.

4.4.4.1 Errors for each audio file

Overall, the accuracy rate for the transcription of each recording is notably high across all files. Most audio recordings were transcribed with an accuracy exceeding 90%, with rates ranging from 88% to 98%, irrespective of the language spoken or the topic discussed. However, three recordings fell below this threshold: REC_006 (88,6%), REC_007 (87,9%), and REC_014 (81,3%) (see *Table 39*, highlighted in bold). These recordings featured discussions in English about pertinent

topics, specifically politics and education, where speakers exhibited varying speech rates from very slow to fast. Notably, REC_006 and (especially) REC_007 had a very high number of speakers compared to the recordings. Furthermore, both REC_007 and REC_014 were classified in the “degraded audio quality" subset and had the lowest accuracy level among the fifteen recordings assessed. Lastly, the REC_007 audio recording included sections where speakers talked over each other.

The ASR system also demonstrated a consistent performance when transcribing audio recordings of native Italian speakers, regardless of the topic (*Politics*: REC_012: *N* = 3945, 95%; REC_013: *N* = 4069, 93,8%; *Education*: REC_015: *N* = 7012, 94,1%) and the number of speakers (*Politics*: REC_012: 1 F, 5 M; REC_013: 2 F, 3 M; *Education*: REC_015: 1 F).

Table 39. *Count and percentage (rounded values) of tokens correctly (No) and erroneously (Yes) transcribed for each audio recording.*

Audio recordings	**Transcription Error (n, percent)**	
	Yes	**No**
REC_001	152 (8,1%)	1731 (91,9%)
REC_002	19 (8,6%)	203 (91,4%)
REC_003	19 (9,1%)	190 (90,9%)
REC_004	10 (6,1%)	154 (93,9%)
REC_005	13 (6,5%)	187 (93,5%)
REC_006	**21 (11,4%)**	**163 (88,6%)**
REC_007	**603 (12,1%)**	**4380 (87,9%)**
REC_008	18 (6,1%)	278 (93,9%)
REC_009	14 (2,4 %)	573 (97,6%)
REC_010	23 (5,9%)	364 (94,1%)
REC_011	19 (6%)	296 (94%)
REC_012	207 (5%)	3945 (95%)
REC_013	271 (6,2%)	4069 (93,8%)
REC_014	**1487 (18,7%)**	**6479 (81,3%)**
REC_015	442 (5,9%)	7012 (94,1%)
Total	3318 (9,95%)	30024 (90,05%)

Note. The audio recordings highlighted in bold are those which scored the lowest rate of accuracy in the test set.

4.4.4.2 Error Type for each audio file

Consistent with the trends observed in the previous section, the highest number of deletions and substitutions occurred in the REC_014 audio file (N = 320, 4% and N = 1027, 13% respectively).This pattern of errors was expected due to the degradation of the speech signal (deletions) and the highly technical terminology presented during the lecture.

In contrast, a similar but inverted trend (compared to the one reported above) is evident for the error occurrences in the REC_006 and REC_007 audio recordings. *Table 40* reports a significant number of substitutions and deletions in these files, where speakers of English read their text during some political intervention and debate. On the one hand, the high number of substitutions errors in REC_006 may have stemmed from the fact that the spokesman was not a native speaker of English; on the other hand, substitution errors in REC_007 might be attributed to overlapping voices from multiple speakers or the noise in the Parliament.

Finally, the lowest number of errors is found in REC_009, which recorded only three deletions (0,5%), one insertion (0,2%), and ten substitutions (1,7%) out of 587 tokens. In this recording, a female native speaker of English read her intervention at a slightly slower pace than the average (183 wpm – Brysbaert, 2019), unconsciously (and involuntarily) contributing to the optimal performance of the ASR system.

Table 40. *Count and percentage (rounded values) of error types (deletions, insertions, substitutions, correct tokens) for each audio recording.*

Audio track	Error Type (n, percent)			
	Deletions	**Insertions**	**Substitutions**	**None**
REC_001	21 (1,1%)	42 (2,2%)	89 (4,7%)	1731 (91,9%)
REC_002	4 (1,8%)	3 (1,4%)	12 (5,4%)	203 (91,4%)
REC_003	1 (0,5%)	5 (2,4%)	13 (6,2%)	190 (90,9%)
REC_004	2 (1,2%)	2 (1,2%)	6 (3,7%)	154 (93,9%)
REC_005	2 (1%)	4 (2%)	7 (3,5%)	187 (93,5%)
REC_006	**6 (3,3%)**	**2 (1,1%)**	**13 (7,1%)**	**163 (88,6%)**
REC_007	**158 (3,2%)**	**80 (1,6%)**	**365 (7,3%)**	**4380 (87,9%)**
REC_008	3 (1%)	7 (2,4%)	8 (2,7%)	278 (93,9%)
REC_009	**3 (0,5%)**	**1 (0,2%)**	**10 (1,7%)**	**573 (97,6%)**
REC_010	2 (0,5%)	5 (1,3%)	16 (4,1%)	364 (94,1%)

REC_011	3 (1%)	2 (0,6%)	14 (4,4%)	296 (94%)
REC_012	61 (1,5%)	38 (0,9%)	108 (2,6%)	3945 (95%)
REC_013	50 (1,1%)	39 (0,9%)	182 (4,2%)	4069 (93,8%)
REC_014	**320 (4%)**	**140 (2%)**	**1027 (12,9%)**	**6479 (81,3%)**
REC_015	83 (1,1%)	125 (1,7%)	234 (3,1%)	7012 (94,1%)
Total	719 (2,2%)	495 (1,5%)	2104 (6,3%)	30024 (90%)

Note 1. Label 'None' refers to the tokens correctly transcribed.

Note 2. The audio recordings highlighted in bold are those which scored the lowest rate of accuracy in the test set.

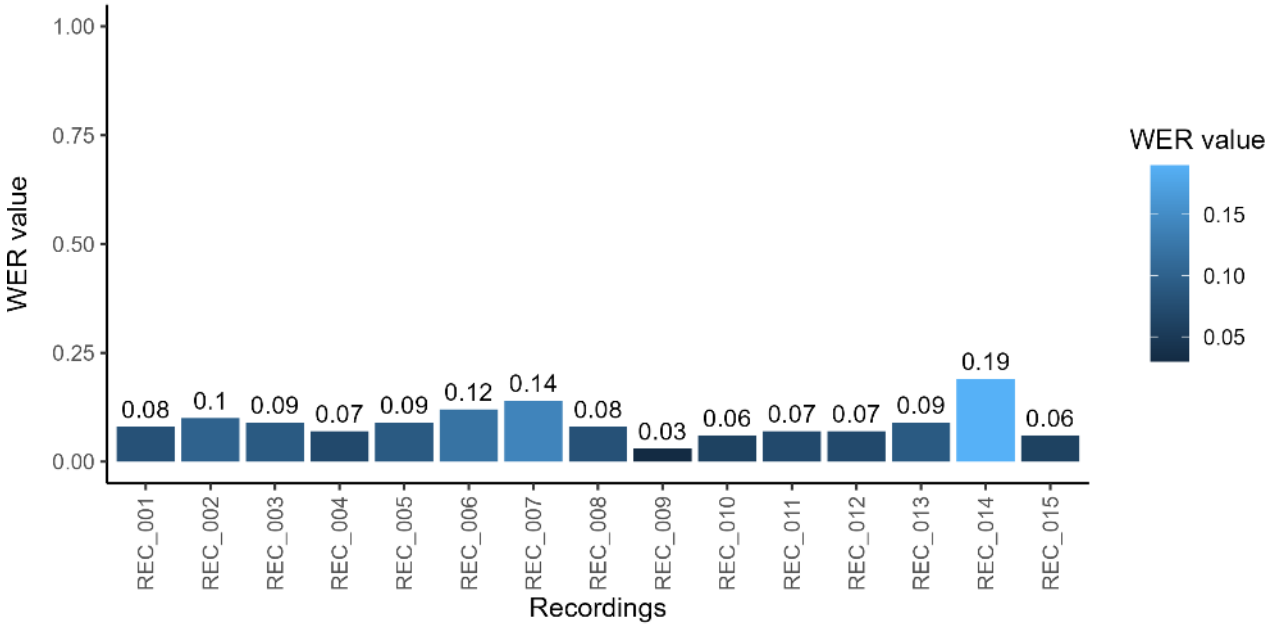
4.4.4.3 Word Error Rate (WER) for each audio file

The WER score was calculated for each audio recording and compared to evaluate the ASR system's performance. The values range is set between 0,03 (REC_009) and 0,19 (REC_014) (*Table 41* and *Figure 39*). The REC_014 recording was the one on which the system performed a worse transcription (0,19), along with the REC_006 and REC_007 (0,12 and 0,14 respectively). These results align with the data discussed in the previous sections.

Table 41. *Word Error Rate (WER) for each audio recording in the test set.*

Audio track	Topic	Language	Word Error Rate
REC_001	Politics	English	0,08
REC_002	Politics	English	0,1
REC_003	Politics	English	0,09
REC_004	Politics	English	0,07
REC_005	Politics	English	0,09
REC_006	**Politics**	**English**	**0,12**
REC_007	**Politics**	**English**	**0,14**
REC_008	Politics	English	0,08
REC_009	Politics	English	0,03
REC_010	Politics	English	0,06
REC_011	Politics	English	0,07
REC_012	Politics	Italian	0,07
REC_013	Politics	Italian	0,09
REC_014	**Education**	**English**	**0,19**
REC_015	Education	Italian	0,06

Figure 39. *Word Error Rate (WER) scores for each audio recording in the test set.*



4.4.4.4 Confidence score values distribution for each audio file

Lastly, the distribution of tokens across the color-coded ranges for each recording was analyzed (*Table 42*). For a considerable number of recordings, demonstrated a high level of confidence in its transcriptions (**black**, range: REC_012 = 75,2% - REC_014 = 47,2%), regardless of the characteristics of the recorded speech signal. Only two recordings in the English group of audio files displayed a rate of 50% or less of its tokens transcribed in **black** – namely, REC_006 and REC_014 (50,6% and 47,2%, respectively). Overall, these two recordings exhibit similar distributions for occurrences in the color-coded ranges. While we have already discussed the reasons for the ASR system's poor performance in transcribing the REC_014 recording, we may link the poor performance in transcribing speech in the REC_006 audio recording to the fact that the main spokesman was a non-native speaker of English. In comparison to the other recordings, the system showed less confidence in the transcription for these two cases, which is reflected in the distribution of tokens across the four color-coded ranges. Notably, the number of tokens in the grey, red, and orange ranges is higher than in the other recordings.

Table 42. *Number and percentage (rounded values) of correct tokens and errors in each color-coded confidence score range (black, grey, red, and ~~red~~).*

Audio track	Color-coded Label (n, percent)			
	Black	Grey	Red	~~Red~~
REC_001	1287 (69,1%)	408 (21,9%)	94 (5,1%)	73 (3,9%)
REC_002	144 (66,1%)	54 (24,8%)	10 (4,6%)	10 (4,6%)
REC_003	143 (68,8%)	42 (20,2%)	9 (4,3%)	14 (6,7%)
REC_004	111 (68,5%)	35 (21,6%)	8 (4,9%)	8 (4,9%)
REC_005	133 (67,2%)	36 (18,2%)	12 (6,1%)	**17 (8,6%)**
REC_006	**90 (50,6%)**	**55 (30,9%)**	**18 (10,1%)**	**15 (8,4%)**
REC_007	2954 (61,2%)	1112 (23,1%)	**348 (7,2%)**	**411 (8,5%)**
REC_008	198 (67,6%)	63 (21,5%)	**22 (7,5%)**	10 (3,4%)
REC_009	431 (73,8%)	113 (19,4%)	21 (3,6%)	19 (3,2%)
REC_010	281 (73%)	65 (16,9%)	19 (4,9%)	20 (5,2%)
REC_011	188 (60,3%)	**90 (28,8%)**	16 (5,1%)	18 (5,8%)
REC_012	3106 (75,2%)	649 (15,9%)	180 (4,4%)	156 (3,8%)
REC_013	3187 (74,3%)	752 (17,5%)	201 (4,7%)	150 (3,5%)
REC_014	**3605 (47,2%)**	**2266 (29,6%)**	**797 (10,4%)**	**978 (12,8%)**
REC_015	5085 (69%)	1507 (20,4%)	384 (5,2%)	395 (5,4%)
Total	**20943 (64%)**	**7247 (22,2%)**	**2139 (6,6%)**	**2294 (7%)**

Note. Percentages were calculated out of the total number of tokens in each recording.

4.4.4.5 Analysis by factor: Topic

We now move to assess the performance of the ASR system by analyzing the distributions of transcription type (correct, incorrect transcription) and confidence score by topic (in this section) and then, by language (in the next one).

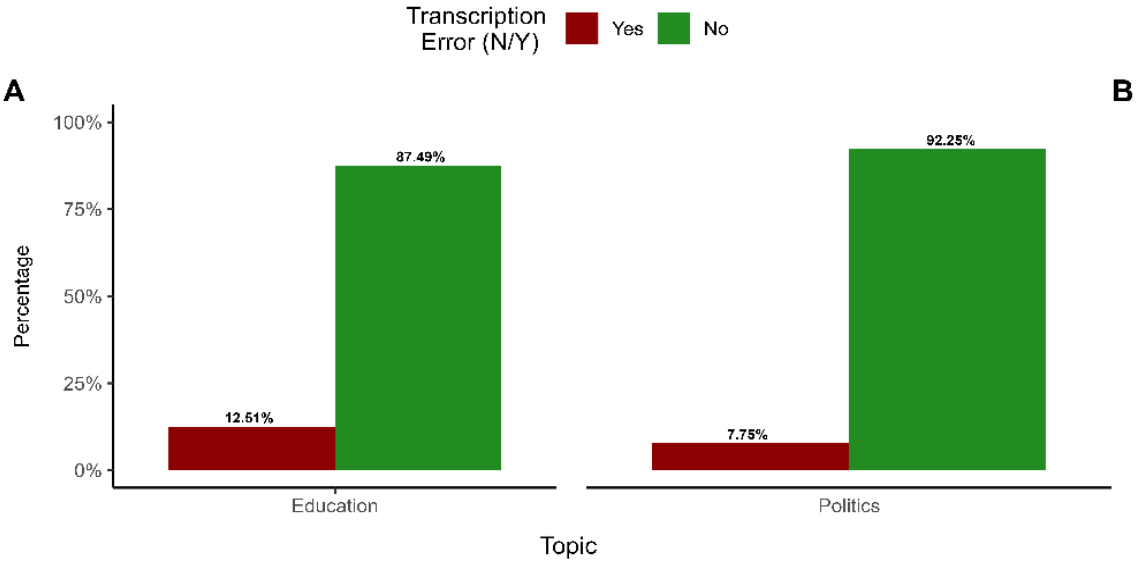
Table 43 summarizes the number of correct transcription and errors in the test set by *topic* (education, politics). Overall, the system transcribed more accurately the recordings about politics (92,3% of correctly transcribed tokens) compared to the recordings with the academic lectures (87,5% of correctly transcribed tokens) (*Figure 40*).

Table 43. *Count and percentage (rounded values) of tokens correctly (No: correct transcription) and erroneously (Yes: erroneous transcription) transcribed grouped by topic (education, politics).*

Topic	Error (n, percent)	
	Yes	No
Education	1929 (12,5%)	1389 (7,7%)
Politics	13491 (87,5%)	16533 (92,3%)

Note. Percentages of tokens correctly and erroneously transcribed were calculated out of the total number of tokens for each topic.

Figure 40. *Percentage (exact values) of tokens correctly (green bars = No – correct transcription) and erroneously (dark red bar = Yes – erroneous transcription) transcribed, grouped by topic (education,*



Note. Percentages of tokens correctly and erroneously transcribed were calculated out of the total number of tokens for each topic.

The majority of errors were found in the recordings from the '*education'* group: substitutions were the most frequent errors (*N* = 1240, 8% out of the total tokens in the group), followed by deletions (*N* = 403, 2,6%) and insertions (*N* = 286, 1,9%) (*Table 44*). *Figure 41* shows the distribution of error type (DEL: deletions, INS: insertions, SUB: substitutions, None: correct tokens) in both groups of recordings.

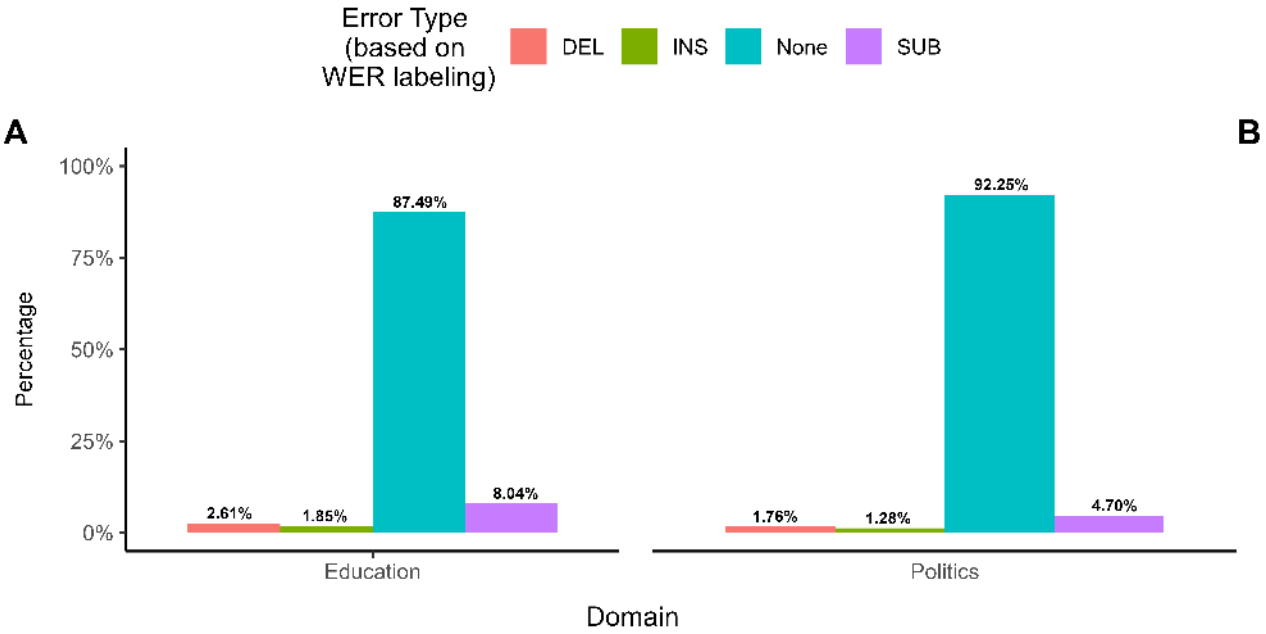
Table 44. *Count and percentage (rounded values) of error types grouped by topic (education, politics).*

Topic	Error Type (n, percent)			
	DEL	INS	SUB	None
Education	403 (2,6%)	286 (1,9%)	1240 (8%)	13491 (87,5%)
Politics	316 (1,8%)	230 (1,3%)	843 (4,7%)	16533 (92,3%)

Note 1. Percentages of tokens correctly and erroneously transcribed were calculated out of the total number of tokens for each topic.

Note 2. *DEL*: deletions, *INS*: insertions, *SUB*: substitutions, *None*: tokens correctly transcribed.

Figure 41. *Distribution of error types in percentage (exact values), grouped by topic (education, politics).*



Note 1. Percentages of tokens correctly and erroneously transcribed were calculated out of the total number of tokens for each topic.

Note 2. *DEL*: deletions, *INS*: insertions, *SUB*: substitutions, *None*: tokens correctly transcribed.

Table 45 and *Figure 42* show that the group of recordings about *education* have a higher WER score compared to the group of recordings concerning *politics* (0,12 versus 0,08 mean WER score). The WER score by *topic* was calculated by averaging the WER scores for the recordings in each group (*Education*: REC_014 = 0.19; REC_015 = 0.06; *Politics*: REC_001 = 0.08; REC_002 = 0.1; REC_003 = 0.09; REC_004 = 0.07; REC_005 = 0.09; REC_006 = 0.12; REC_007 = 0.14; REC_008

= 0.08; REC_009 = 0.03; REC_010 = 0.06; REC_011 = 0.07; REC_012 = 0.07; REC_013 = 0.09). The groups were unbalanced (number of recordings in the *Education* subset = 2; number of recordings in the *Politics* subset = 13), and for this reason, we ran the Mann-Whitney U test to assess if the two mean WER scores differed. The Wilcoxon rank sum test with continuity correction testing the difference in ranks between the average WER score of the audio recordings in the *Education* group and the average WER score of the audio recordings in the *Politics* group suggests that the effect is positive, statistically not significant, and small (W = 14.50, p = 0.864; r (rank biserial) = 0.12, 95% CI [-0.65, 0.76]). In sum, results of the statistical test show that the average WER scores of the two subsets did not differ significantly, and that the performance of the ASR system when transcribing the two topics was comparable, with a tendency of higher accuracy when transcribing audio recordings in the *Politics* subset.

Table 45. *Number of recordings and mean (SD) Word Error Rate (WER) for the audio recordings grouped by topic (education, politics).*

Topic	Sample size	Mean WER (SD)
Education	2	0,12 (0,09)
Politics	13	0,08 (0,03)

Figure 42. *Word Error Rate (WER) for the audio recordings grouped by topic (education, politics) (light blue bar: education; dark blue bar: politics).*

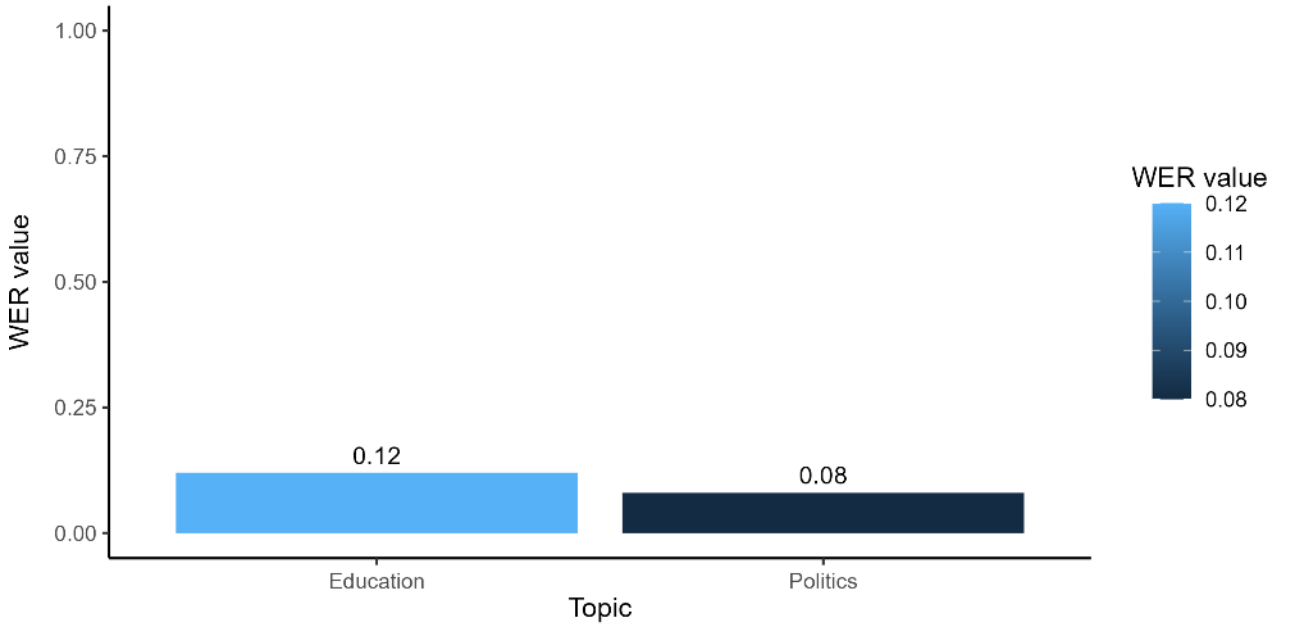
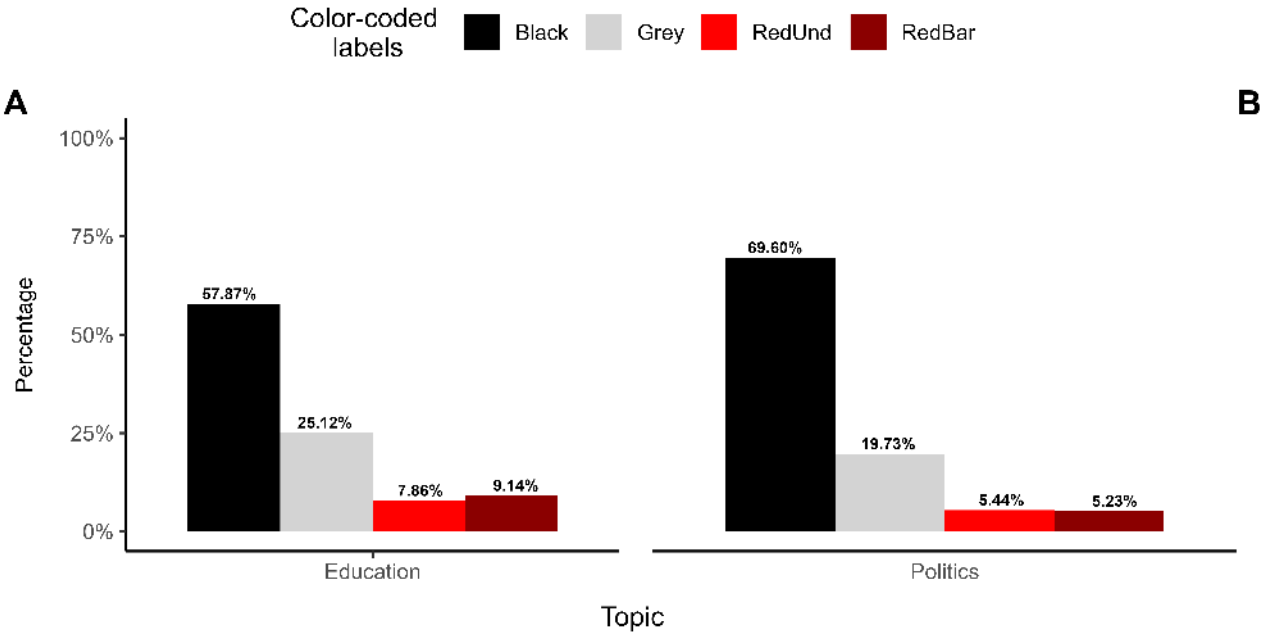


Table 46 summarizes the number of tokens in each color-coded confidence range by topic. The system showed high confidence in the transcriptions of a higher number of tokens in the *politics* group of recordings (**black**, N = 12253, 69,6%) compared to the number of tokens in the *education* group of recordings (**black**, *N* = 8690, 57,6%). Conversely, the system showed a higher number of words transcribed with low confidence scores in the *education* recordings (red: *N* = 1181, 7,9%; ~~red~~: *N* = 1373, 9,1%) compared to the number of tokens in the *politics* group of recordings (red: *N* = 958, 5,4%; ~~red~~: *N* = 921, 5,2%). *Figure 43* shows the distributions of the tokens in each color-coded range by topic.

Table 46. *Count and percentage (rounded values) of tokens in each color-coded confidence score range (**black**, grey, red, and ~~red~~) grouped by topic (education, politics).*

Topic	Color-coded Label (n, percent)			
	Black	Grey	Red	~~Red~~
Education	8690 (57,9%)	3773 (25,1%)	1181 (7,9%)	1373 (9,1%)
Politics	12253 (69,6%)	3474 (19,7%)	958 (5,4%)	921 (5,2%)

Figure 43. *Distribution of tokens (exact value in percentage) in each color-coded confidence score range (**black**, grey, red, and ~~red~~) grouped by topic (education, politics).*



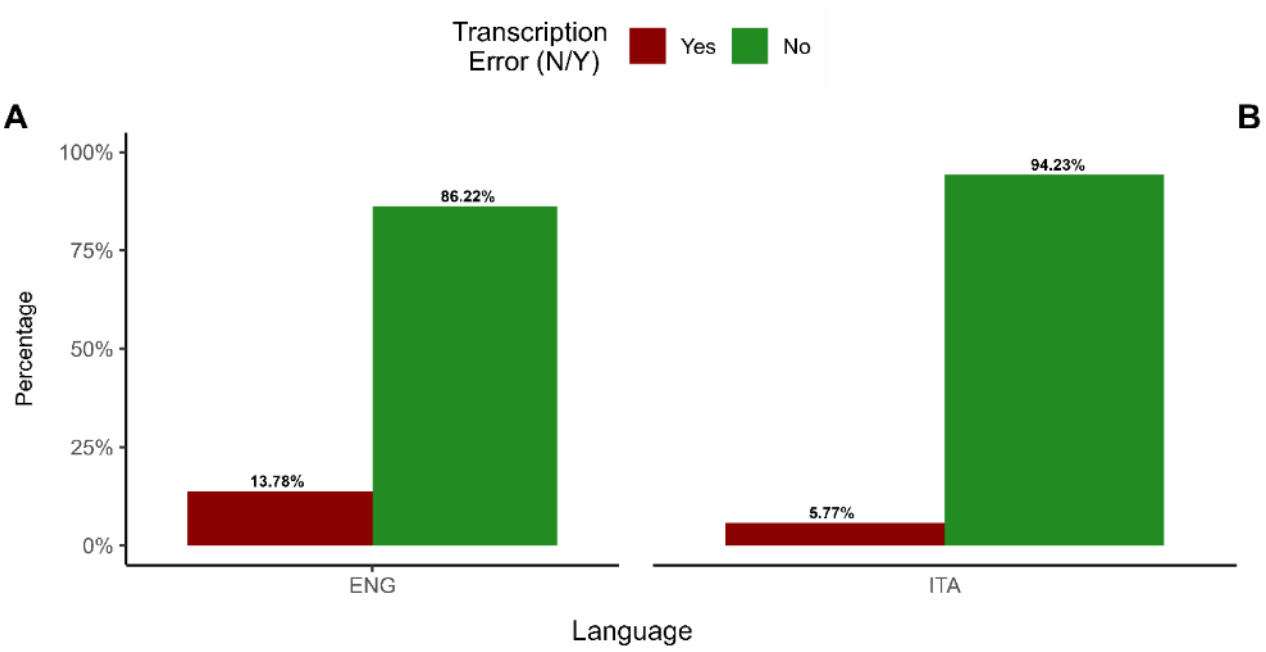
4.4.4.6 Analysis by factor: Language

When assessing the performance of the ASR system by language, we find that - in general - the accuracy rate for transcribing audio recordings in Italian is better than that for English (*Table 47* and *Figure 44*). The number of correctly transcribed tokens was higher in the Italian recordings (*N* = 15026, 94,2%) compared to the number of tokens in the English recordings (*N* = 14998, 86,2%).

Table 47. *Count and percentage (rounded values) of tokens correctly (No: correct transcription) and erroneously (Yes: erroneous transcription) transcribed by language (English, Italian).*

Language	Error (n, percent)	
	Yes	No
English	2398 (13,8%)	14998 (86,2%)
Italian	920 (5,8%)	15026 (94,2%)

Figure 44. *Percentage (exact values) of tokens correctly (green bars = No – correct transcription) and erroneously (dark red bar = Yes – erroneous transcription) transcribed grouped by language (ENG: English, ITA: Italian).*



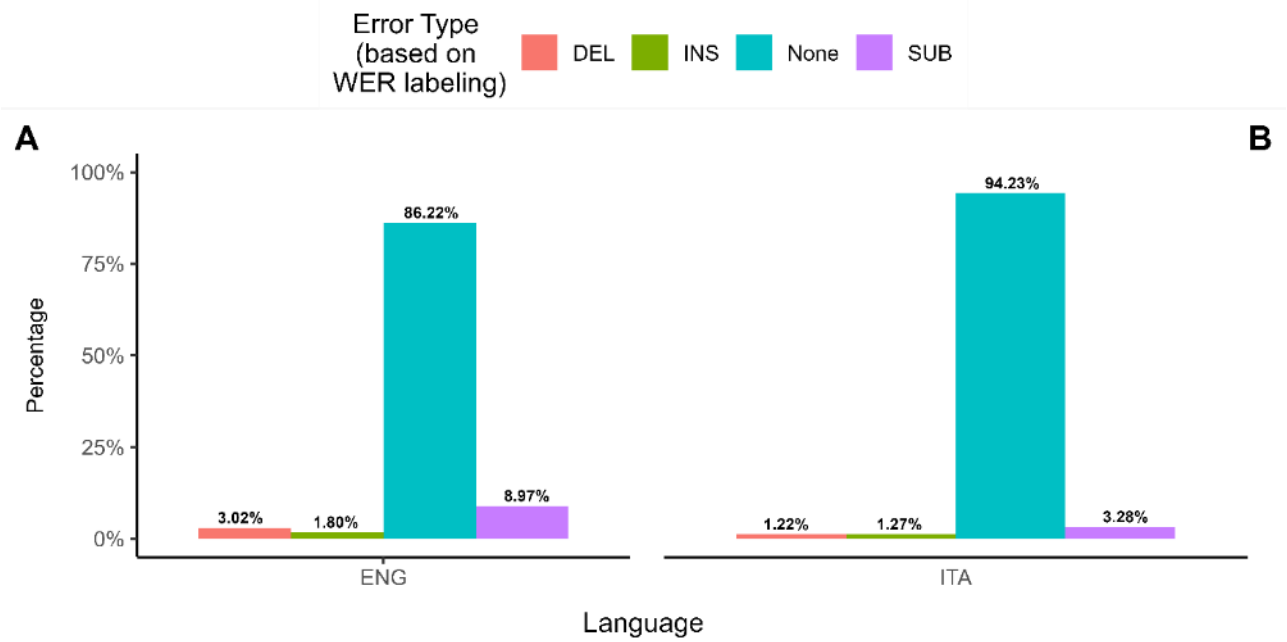
The analysis of the error types shows a high number of substitutions (*N* = 1560, 9%) and deletions (*N* = 525, 3%) in the English group of recordings (*Table 48*). *Figure 45* shows the distributions of error types by language of the recordings.

Table 48. *Count and percentage (rounded values) of error types grouped by language (English, Italian).*

Language	Error Type (n, percent)			
	DEL	INS	SUB	None
English	525 (3%)	313 (1,8%)	1560 (9%)	14998 (86,2%)
Italian	194 (1,2%)	203 (1,3%)	523 (3,3%)	15026 (94,2%)

Note. *DEL*: deletions; *INS*: insertions; *SUB*: substitutions; *None*: tokens transcribed correctly.

Figure 45. *Distribution of error types grouped by language (ENG: English, ITA: Italian) in percentage (exact values).*



Note. *DEL*: deletions; *INS*: insertions; *SUB*: substitutions; *None*: tokens transcribed correctly.

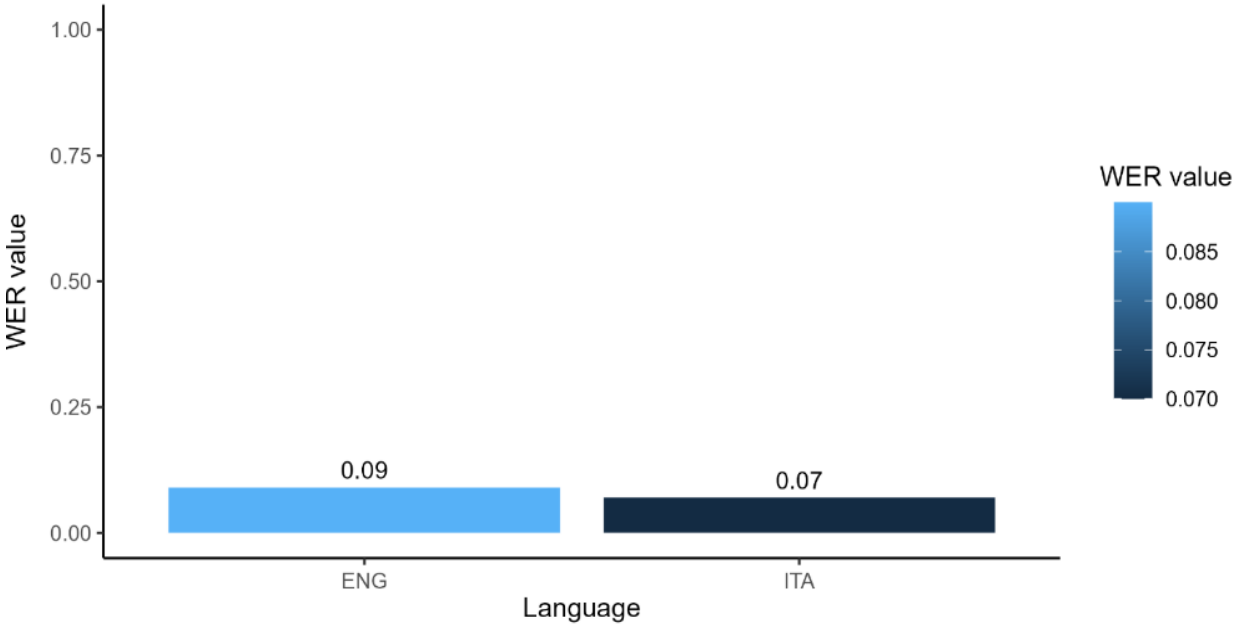
Table 49 and *Figure 46* show that the group of recordings where speakers talk in *English* have a slightly higher WER score compared to the group of recordings where speakers talk in *Italian* (0,09

versus 0,07 mean WER score). The WER score by *language* was calculated by averaging the WER scores for the recordings in each group (*English*: REC_001 = 0.08; REC_002 = 0.1; REC_003 = 0.09; REC_004 = 0.07; REC_005 = 0.09; REC_006 = 0.12; REC_007 = 0.14; REC_008 = 0.08; REC_009 = 0.03; REC_010 = 0.06; REC_011 = 0.07; REC_014 = 0.19; *Italian*: REC_012 = 0.07; REC_013 = 0.09; REC_015 = 0.06). Again, groups were unbalanced (number of recordings in the *English* subset = 12; number of recordings in the *Italian* subset = 3), and for this reason, we ran once again the Mann-Whitney U test to assess if the two mean WER scores differed. The Wilcoxon rank sum test with continuity correction testing the difference in ranks between the average WER score of the audio recordings in the *English* subset and the average WER score of the audio recordings in the *Italian* subset suggests that the effect is positive, statistically not significant, and large (W = 20.50, p = 0.395; r (rank biserial) = 0.37, 95% CI [-0.37, 0.82]). In sum, results of the statistical test show that the average WER scores of the two subsets did not differ significantly, and that the performance of the ASR system when transcribing the two languages was comparable, with a tendency of higher accuracy when transcribing audio recordings in *Italian*.

Table 49. *Number of recordings and mean (SD) Word Error Rate (WER) for the audio recordings grouped by language (English, Italian).*

Language	Sample size	Mean WER (SD)
English	12	0,09 (0,04)
Italian	3	0,07 (0,02)

Figure 46. *Word Error Rate (WER) for language (light blue bar: English; dark blue bar: Italian).*

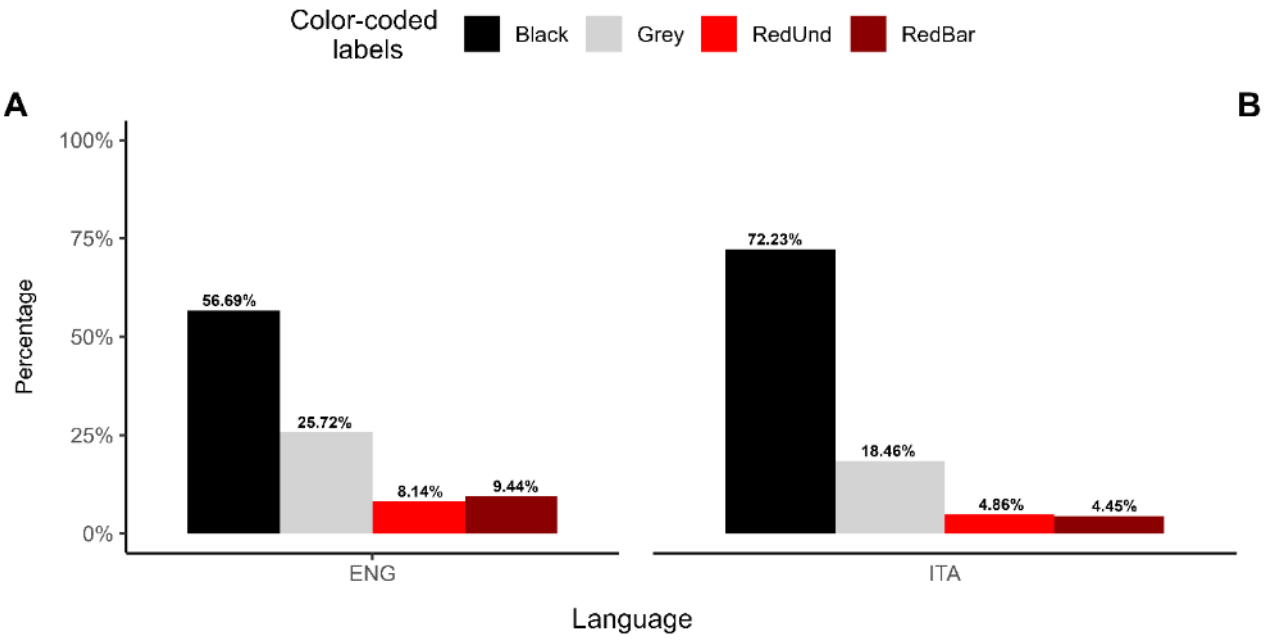


The system was more confident of its transcription for Italian in comparison to English (*Table 50*). *Figure 47* shows the differences in confidence between the transcriptions of the recordings in the two groups: it is evident that the system was more confident when transcribing speech in Italian, since there is a higher number of tokens transcribed in black compared to English. Furthermore, the number of tokens transcribed in grey, red and ~~red~~ is higher for English compared to Italian.

Table 50. *Count and percentage (rounded values) of tokens in each color-coded confidence score range (**black**, grey, red, and ~~red~~) grouped by language (English, Italian).*

Language	Color-coded Label (n, percent)			
	Black	Grey	Red	~~Red~~
English	9565 (56,7%)	4339 (25,7%)	1374 (8,1%)	1593 (9,4%)
Italian	11378 (72,2%)	2908 (18,5%)	765 (4,9%)	701 (4,4%)

Figure 47. *Distribution of tokens (exact value in percentage) in each color-coded confidence score range (**black**, grey, red, and ~~red~~) for language (English, Italian).*



4.4.4.7 Assessing the influence of factors on confidence scores and accuracy

The descriptive statistics highlighted a potential influence of some of the factors of interest on the performance of the ASR system. For this reason, we used the *Conditional Inference Tree* (CIT) analysis to assess which speaker- and/or environment-related factors have the strongest association with our dependent variables, namely *transcription type* (that is, accuracy - correct, incorrect transcriptions) and *confidence scores*.

We created a data file containing all the relevant information for the CIT analysis. We filtered out all tokens in the REF texts that were deleted and did not appear in the output of the ASR system (HYP texts). This ensured that we included only those tokens that were assigned a confidence score.

The factors we included in the model are those we listed in the methods:

- Topic
- Language
- Number of speakers
- Overlapping voices

- Gender of speaker(s)
- Speaker(s)' native language
- Speaking/reading rate

We performed our analysis on R using the package *party* (version 1.3 - 18) (Hothorn *et al*., 2006a; Hothorn *et al*., 2006b; Strobl *et al*., 2007; Strobl *et al*., 2008; Zeileis *et al*., 2008).

The model used to run the CIT analyses on confidence score was the following:

```
# Run the CIT analysis

ConfScore_OverallTestSet_CIT_Complete <- ctree(ConfScore ~ Language + Domain +
AudioQuality + NoOfSpeakers + OverlappingVoices + SpeakerGender + NativeSpeaker
+ SpeakingRateCoding, data = CIT_df_clean)


# Plot the graph


plot(ConfScore_OverallTestSet_CIT_Complete)
```

Figure 53 shows an individual CIT fitted to assess the influence of the relevant factors on confidence score. In this case, the most important predictor for these values is *Audio Quality* (*Node 1, p* < 0.001). Data included in both levels of that variable ("clean" and "degraded" audio quality) are then affected by the *Number of Speakers* in the audio recordings (*Node 2*, *p* < 0.001; and *Node 9*, *p* < 0.001). *Node 10* shows that the presence of more than three speakers affects the value of 4825 confidence scores ($mean_{ConfScore}$ = 0.939); in comparison, the presence of two speakers in the audio recordings affects 7646 confidence scores (*Node 11* - $mean_{ConfScore}$ = 0.908). This result is unexpected, since even if both nodes include tokens from recordings with a "degraded" quality of the audio, the system was less confident when transcribing speech from audio recordings with a low number of speakers.

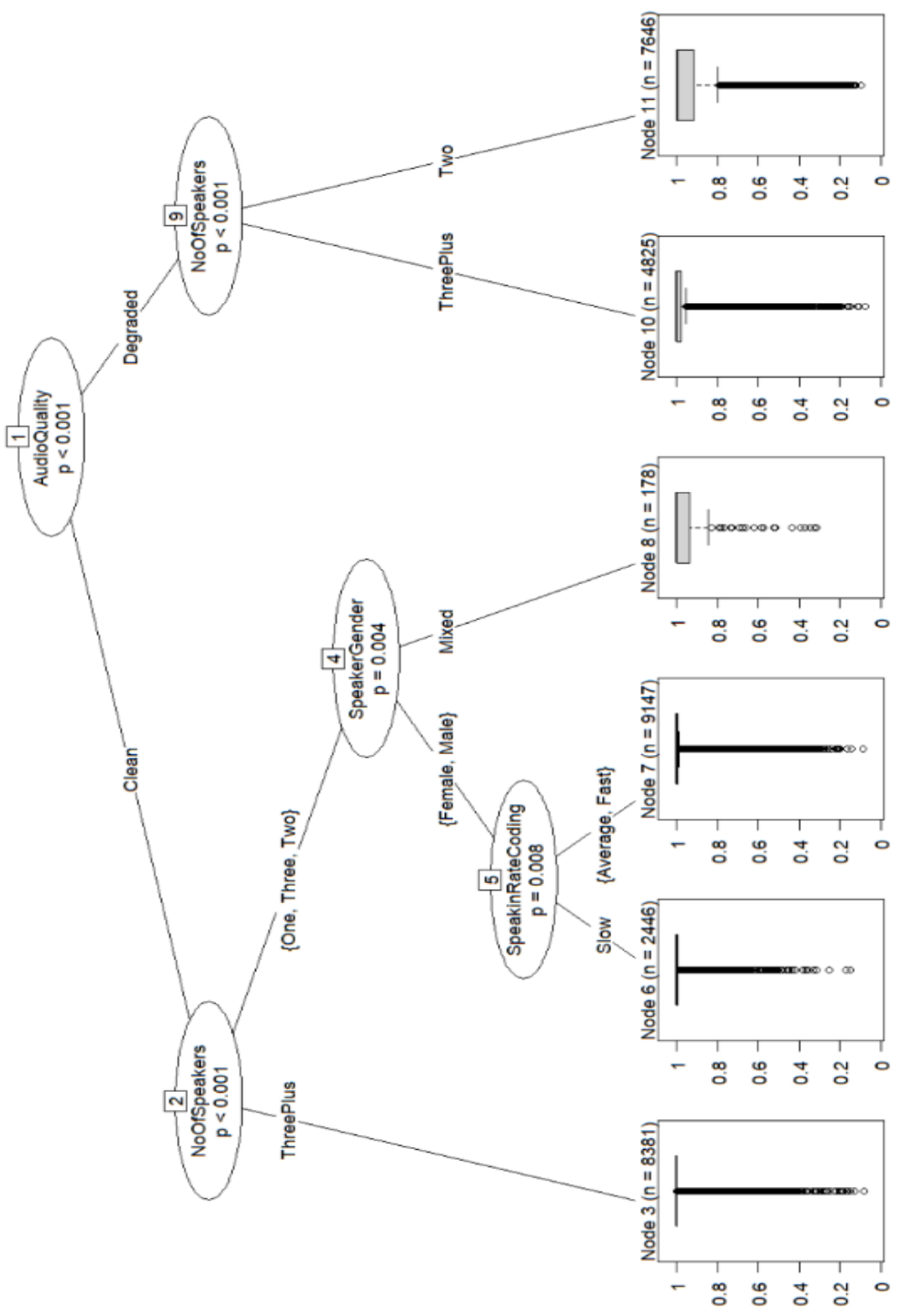


Figure 48. *CIT based on the assessment of the influence of speaker- and environment-related factors on confidence scores.*

In the recordings with a "clean" audio quality, confidence scores are affected by other variables depending on the number of speakers talking. On the one hand, *Node 3* shows that the confidence scores printed from the recordings where more than three speakers talk have an average confidence score of 0.971 (N = 8381). On the other hand, the confidence scores from the recordings with one, two, or three speakers are also affected by the speakers' gender (*Node 4*, $p = 0.004$). *Node 8* shows that 178 confidence scores were affected by the fact that there were speakers of different genders in the audio recording ($mean_{ConfScore} = 0.932$). Lastly, confidence scores of tokens transcribed from either female or male speakers were also affected by the speech rates of these persons (*Node 5*, p = 0.008). However, when compared, mean confidence scores in *Node 6* (N = 2446 – slow speaking/reading rate) and *Node 7* (N = 9147 – average and fast speaking/reading rate) are highly similar (0.969 vs 0.960, respectively). Lastly, we calculated the tree accuracy[52] to check its fit: results showed that this CIT has a classification accuracy of 0.91, which indicates a good fit for the tree.

We then performed the CIT analysis on transcription type using the following model:

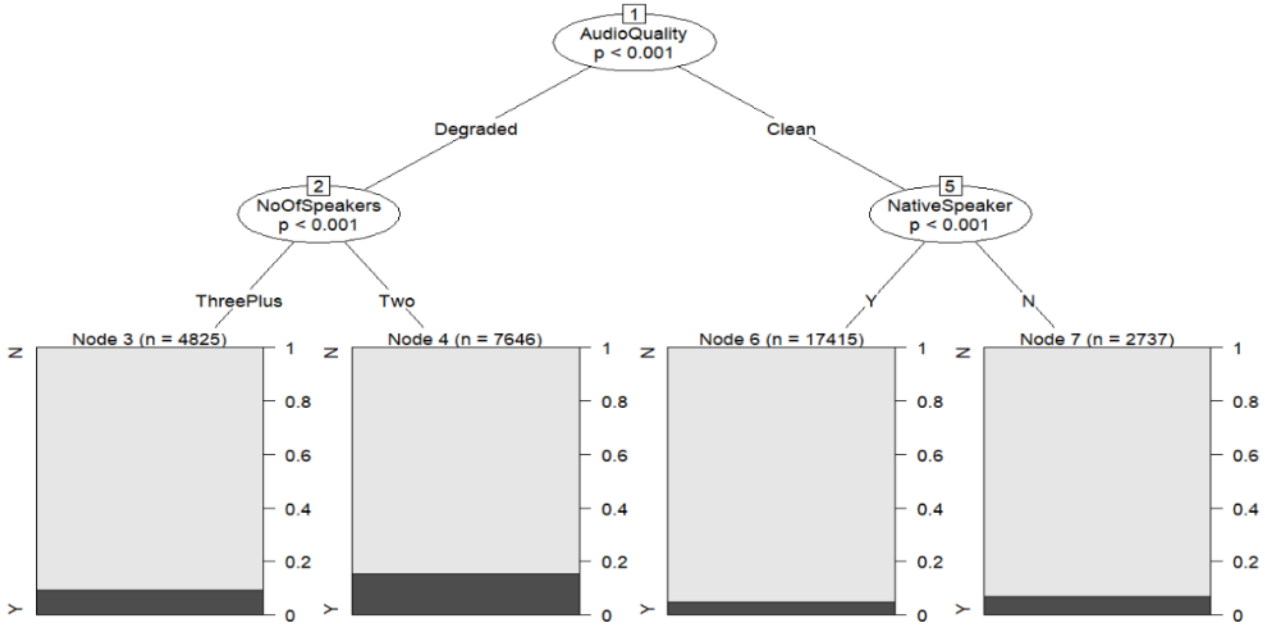
```
# Run the CIT analysis

ErrorWER_OverallTestSet_CIT_Complete <- ctree(ErrorWER ~ Language + Domain + Au
dioQuality + NoOfSpeakers + OverlappingVoices + SpeakerGender + NativeSpeaker +
SpeakingRateCoding, data = CIT_df_clean)


# Plot the graph


plot(ErrorWER_OverallTestSet_CIT_Complete)
```

The statistical analysis we conducted to assess the influence of *relevant factors* on accuracy marked *Audio quality* as the most important predictor (*Node 1, Figure 54*) The left-hand branch – which included the "degraded" tokens – are also affected by the *topic* of the recordings (*Node 2* – "Domain", $p < 0.001$): there is a higher proportion of errors for those tokens included in the recordings about *education* (academic lectures – *Node 4* - N = 7646, 85%) in comparison to proportion of errors for those tokens included in the recordings about *politics* (*Node 3* – N = 4825, 91%). Conversely, the tokens in the right-hand branch ("clean" audio quality) of the CIT are also affected by the *nativeness* of the speaker(s) talking in the audio recordings: in fact, the system seems to struggle more with the transcription of speech produced by non-native speakers (*Node 7*, N = 2737) than when transcribing speech produced by native speakers (*Node 6*, N = 17415) (higher proportions of errors in *Node 7* than

[52] "Accuracy (…) is defined as the number of correct predictions divided by the total number of observations" (Levshina, 2020: p. 632).

in *Node 6* – 0.07 vs 0.05). However, the predictive power is not satisfactory even for this model. While the CIT has a classification accuracy of 0.92, the concordance index C^{53} is 6.498791e-01, indicating that the model very much struggles to discriminate between outcomes.

Figure 49. *CIT based on the assessment of the influence of speaker- and environment-related factors on transcription type (accuracy).*



4.4.5 Creation of the second and third versions of the color-coded markup

After completing the analysis of the test set, we proceeded to create two new versions of the color-coded markup - namely, the V2 and V3 display formats. Before describing how we developed the new markups, we will quickly review the key results of the analysis of the performance of the ASR system.

- The ASR system exhibited a strong performance, achieving a transcription accuracy of 90%.

[53] The Concordance Index C (or C-Index) signals how well the model can discriminate between the response categories. It requires binary response variables to be computed. Following Levshina (2020), "a model has acceptable discrimination between the response categories if C is higher than 0.7, good if it is above 0.8, and excellent if it is above 0.9." (p. 633)

- The majority of the tokens were transcribed with values in the 'high' ranges of confidence - that is, black (confidence score value = 1; about 65%) and grey (range of confidence score: 99,9% - 90%; about 22%).
- WER scores indicate that the quality of transcriptions was generally good, regardless of the topic or language. However, the ASR system showed a tendency to transcribe recordings more accurately when they were about politics and when the speakers were Italian.
- Descriptive analysis of each audio recording of the test set revealed that the combination of some of the speaker- and environment-related factors we considered for our analysis affected both accuracy and confidence scores.
- The *Conditional Inference Tree* analysis highlighted that the performance of the ASR system (both accuracy and confidence scores) was mainly affected by the quality of the audio recordings. On the one hand, *accuracy* was also affected mainly by the number of speakers in the audio recordings, their gender, and their speaking/reading rate. On the other hand, *confidence scores* were also affected by the topic of the audio recordings (therefore, by the speech type – namely, semi-spontaneous speech and read speech) and the native language of the speakers. However, we must be cautious to interpret these results, since the C-index for one of the two models was extremely low, indicating its struggle to clearly determine the effects of the factors on the dependent variable and consequently group them.
- Correlation analysis revealed a statistically significant, positive and medium-sized correlation between transcription type and confidence score (greater chances of correct transcription at higher confidence scores).
- It was not possible to determine specific threshold values to define each range, since there are instances of tokens incorrectly transcribed even if they were assigned a confidence score = 1 (and vice versa - tokens correctly transcribed even if they were assigned a confidence score in the low range (≤ 69,9%).

These results, along with the opinions and insights of participants from the first study, were then utilized to develop the new experimental markups.

As previously described, the first markup (named *original* - OG) was created by one of the IT technicians of the partner company (*Table 54* also see section 4.3.2.1 for some additional details on the creation of the first markup). One of the IT technicians of the research team at the partner company (Cedat85) - based on their experience with the technology - arbitrarily defined all the elements of the

first markup, namely: 1) the threshold values to define the relevant confidence score (printed by the ASR system along with the textual output; range: between 0 and 1) ranges, 2) a color to signal the values in each range, and 3) a label to identify the ranges to be used during the correlation analysis.

For the first type of markup, the color-coding scheme was the following:

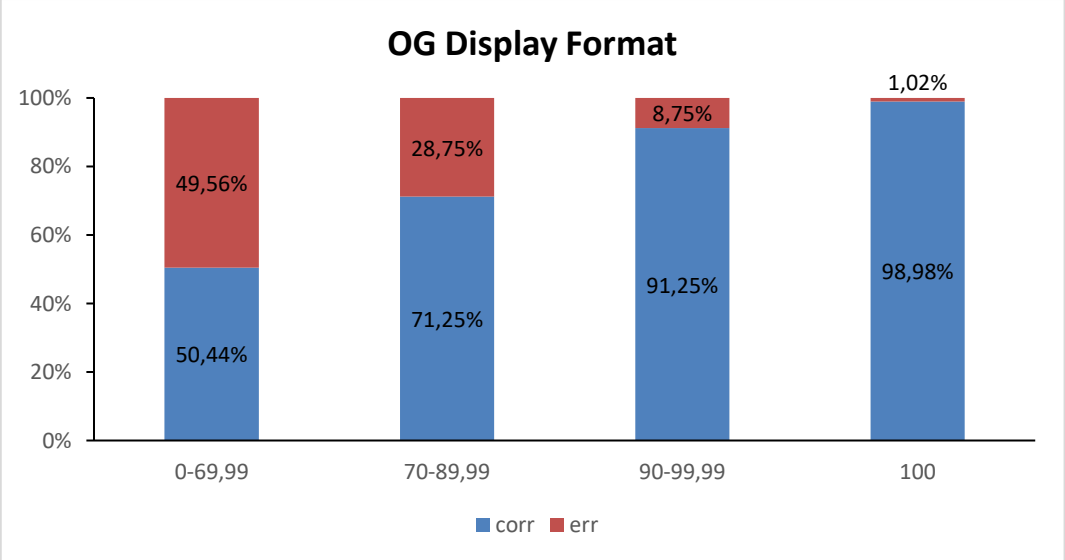
Table 51. *First version (OG) of the color-coded markup.*

Confidence Score Range	Color-coded Label	Range Labels
100%	**Black**	High
99,9% - 90%	Grey	
89,9% - 70%	Red	Mid
≤ 69,9%	~~Red~~	Low

Note. Each confidence score range was graphically represented with a color (**black**, grey, red, ~~red~~) and a label (high, mid, low). Both elements were defined arbitrarily by the research team from the partner company based on their experience in the field.

After the analysis of the test set, a second (called *version 2*, V2) and a third version (called *version 3*, V3) of the color-coded markup was created (*Table 55*). Below, the process of their creation is described.

Figure 50. *Distribution of tokens correctly and erroneously transcribed across the test set in each range (OG markup).*



First, the overall distribution of the tokens in the test set leaned toward the higher values of confidence score, with almost 90% of the tokens transcribed with a value included in the **black** and grey ranges (see *Figure 34*, p. 116). Analyzing the percentages of correctly and incorrectly transcribed tokens in each range, we observe that the majority of correct tokens fall within the **black** (100%) and grey (99,9% - 90%) ranges. In contrast, a larger percentage of incorrect tokens is found in the red (89,9% - 70%) and ~~red~~ (≤ 69,9%) ranges (*Figure 55* above). We therefore performed the statistical analysis (see section 4.4.3), which revealed weak and very weak correlations between the type of transcription and each range of the confidence scores (especially in the red range).

In light of these analyses, along with the research team of the partner company, we mutually agreed to change the threshold values of the ranges and their colors (*Table 55*).

First, we aimed to balance the distribution of correctly and incorrectly transcribed tokens from our test set: similar to what happened for the creation of the OG markup, and since our analysis did not reveal any specific confidence score that would serve as threshold values for each range, we arbitrarily decided to define six threshold values and three ranges. The **black** range (100% - 80%) was created by merging the two ranges under the 'high' label from the OG markup and by adjusting the threshold scores. We opted to merge these two ranges because the research team from our partner company determined that the percentage of errors in the grey range was sufficiently low. Similarly, the red and ~~red~~ threshold values were changed so as to balance the different range of confidence score values.

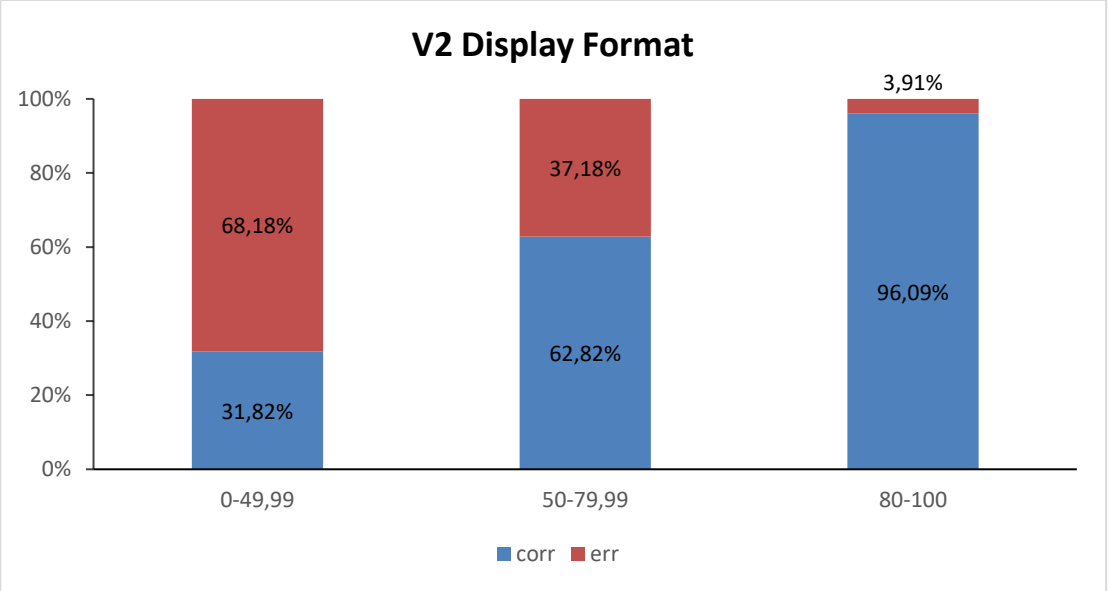
Table 52. *Second version (V2) of the color-coded markup.*

Confidence Score Range	Color-coded Label	Range Labels
100 - 80%	**Black**	High
79,9 – 50%	Red	Mid
49,9 – 0%	~~Red~~	Low

Figure 56 shows the distributions of correctly and incorrectly tokens in the test set used for our corpus analysis, but with the new color-coded ranges[54].

[54] This distribution is also valid for the V3 display format (see in the next page).

Figure 51. *Distribution of tokens correctly and erroneously transcribed across the test set, per range (V2 and V3).*



Finally, the third version of the markup (called *version 3*, V3) was defined (*Table 56*). In order to create this format, we kept the same ranges of confidence scores, and following the results of the questionnaire administered to university students in the first phase of this project (see Chapter 3, §3.5 for the discussion of results), we eliminated the remaining color from the markup (red and ~~red~~). Words transcribed with low confidence scores (below 80%) were then signaled with underlined or strikethrough words.

Table 53. *Third version (V3) of the color-coded markup.*

Confidence value range	Color-coded markup	Range Labels
100 - 80%	White	High
79,9 – 50%	White	Mid
49,9 – 0%	~~White~~	Low

4.5 Discussion

The aims of this study were to:

- Assess the performance of an ASR system when transcribing audio recordings from real-world contexts – specifically, academic lectures delivered in English;
- Test how reliable confidence scores are with the aim of developing color-coded markups to be implemented in the text of the automatic captions;
- Investigate the influence of speaker- and environmental-related factors on accuracy and confidence scores.

RQ1. *How well does a traditional, speaker-independent ASR system perform when dealing with the transcription of audio recordings collected in various real-world contexts?*

RQ2. *Which speaker- and environmental-related factors affect accuracy of transcriptions and confidence values the most?*

Overall, the ASR system performed well in transcribing the audio recordings from real-world settings (education, politics). Accuracy was assessed at 90% of correctly transcribed tokens, with a global WER score of 0.1. The most frequent errors were – in order - substitutions (6,25%), deletions (2,15%), and insertions (1,55%). In proportion to the total number of tokens in each color-coded range of confidence scores, the highest number of incorrectly transcribed tokens were included mainly in the red (*N* = 615, 28,75% out of the total number of tokens transcribed with a confidence score comprised between 70% and 89,9%) and ~~red~~ (*N* = 1137, 49,56% out of the total number of tokens transcribed with a confidence score below 69,9%) ranges of confidence scores. Concerning confidence in its transcription, the system transcribed most of the tokens in the test set with a confidence score in the range of 90%-100% (~86% out of the total number of the tokens). This type of performance highlights that this ASR system is robust enough to be used in various real-world contexts. The descriptive analysis also indicated a small difference in performance of the system between the categories included in the *topic* and *language* variables. The ASR system performed slightly better (in terms of accuracy and confidence in its transcriptions) when transcribing the majority of the audio recordings in the domain of *politics.* In this context, the speakers read from a script, and their speech had very few disfluencies. Speech signals, moreover, were recorded using unidirectional microphones in well-designed halls that featured good acoustics and sound systems. The combination of these factors resulted in a relatively low error rate and high confidence of the system in its transcription.

Conversely, the ASR system performed worse in the transcription of audio recordings in the domain of *education* (especially in English). In this case, the acoustic, speakers, and environment-related factors might have played a role along with the intrinsic nature of the speech itself, since the main speaker was engaged in semi-spontaneous speech production, which is notoriously more affected by phenomena such as connected speech, false starts, hesitations, and repairs (e.g., Benzeghiba *et al*., 2007). Regarding the performance of the ASR system when dealing with different languages, descriptive statistics highlighted a high accuracy and very good confidence of the system when transcribing in Italian compared to English - independently from the topic of the audio recordings. In fact, the WER error rate for the recording *REC_015* (the academic lecture) was similar to the WER score of recordings *REC_012* and *REC_013* (political interventions and debates) (WER scores: 0.06 versus 0.07 and 0.09, respectively). Overall, the ASR system performed better when compared to its English counterpart. This results also stem from the fact that the Italian recording contained a lower number of domain-specific lexical entries compared to the recording in English. In sum, the findings from our descriptive analysis are consistent with previous research, especially concerning the *topic* factor (e.g., Del Rosso & Brambilla, 2022; Kuhn *et al*., 2024). While its performance was overall good, the system struggled more when transcribing speech in English and from the education domain compared to its performance in transcribing Italian and audio recordings from the politics domain. However, further analyses that compared the mean WER scores of audio recordings by groups highlighted statistically non-significant differences, confirming the good performance of the ASR system in diverse real-life contexts.

After evaluating the ASR system's performance on the complete test set, we explored how the characteristics of individual audio files affect confidence scores and the accuracy of transcriptions. The system performed worse and was less confident with its transcription when transcribing the audio recordings *REC_006, REC_007,* and *REC_014*. The audio file *REC_007* was included in the 'politics' group of recordings: the speakers read a script, but the speech signal was affected not only by ambient noise and some reverberation but also by overlapping speakers, some of whom read at high speed. Conversely, the *REC_014* audio file was recorded in a lecture hall with poor acoustics, which affected the clarity of the speech signal. Additionally, one of the two speakers was positioned far from the microphone, causing the ASR system to recognize only a few words and degrading its performance. In this case, the location of the source was crucial in determining the quality of the speech signal and greatly affected the performance of the ASR system, which could not accurately detect the phones and match them to corresponding words, resulting in a high number of deletions (similar to the results of Del Rosso and Brambilla's 2022 study).

The effects we observed in the descriptive data were not confirmed by the *Conditional Inference Tree* (CIT) analysis, since for both accuracy and confidence scores, the models marked *audio quality* as the most important predictor (see 4.4.4.7). These results are clearly in line with previous research stating that acoustic and environment-related factors have a major impact both on the accuracy and the confidence of the system in its transcription (e.g., Benzeghiba *et al*., 2007). This is particularly relevant for the *REC_014* audio recording: this audio file, in fact, was characterized by particular acoustic and environment-related elements which significantly degraded the speech signals. The CIT analysis also highlighted the detrimental effects of accented speech (since some audio recordings contained speech from non-native speakers of English), speakers' gender and speech rates especially on the confidence scores. It is plausible that these factors had an impact on the performance of the ASR system during the extraction of the features and the word recognition processes (Benzeghiba *et al*., 2007). These factors could have affected the speech signal, creating some conflict between the data in the training set (especially if trained on the most common variants of English, such as Received Pronunciation or General American English - Kuhn *et al*., 2024) and the features in the speech signal, which in turn could have led to higher number of transcription errors and lower confidence scores. However, these results need to be interpreted with caution, since the C-index for the accuracy model was extremely low, indicating its struggle to clearly determine the effects of the factors on the dependent variable and consequently group them, potentially due to a very high correlation between multiple factors. These results therefore call for further investigation into the matter in order to better understand the cumulative effects of speaker- and environment-related factors on the performance of the ASR systems.

In conclusion, even if the ASR system performed well in transcribing audio recordings from diverse real-world settings; however, since we were not able to completely assess the effects of external, speaker- and environment-related factors on the performance of the ASR system, these results (especially the ones from the CIT analysis) call for the necessity to examine more in depth the use of ASR systems in academic settings in order to guarantee access to information to diverse populations (similar to those explored by researchers in the *Liberated Learning Project* - e.g., Wald & Bain, 2008).

RQ3. *Is there a correlation between confidence scores and the type of transcription (correct/erroneous)?*

RQ4. *Is there a threshold value determining that a certain correlation is always true (e.g., a specific low value corresponds 100% with an erroneous transcription)?*

Overall, the ASR system tends to be sure about most of its transcription. As we presented in section 4.5.3, errors and correct transcriptions appear on both ends of the graph - both at the highest and the lowest confidence scores. The lowest end of the scores includes the highest number of tokens incorrectly transcribed; on the contrary, the highest number of correctly transcribed tokens is on the highest end of the scores. This trend suggests that there may not be a straightforward correlation between confidence scores and transcription accuracy. However, it indicates that lower scores are likely associated with a *higher probability* of incorrect transcriptions, and higher scores are associated with a higher probability of correct transcriptions (also suggested by Jiang, 2005, and Vertanen & Kristensson, 2008). Statistical analysis of the data confirmed the trend, highlighting a positive, medium-strength, statistically significant correlation between confidence scores and transcription type for the entire test set. These results align with previous research (e.g., Dua *et al*., 2023) and confirm the assumption that it may not be possible to pinpoint a specific confidence score starting from which we will have only correct or incorrect transcriptions. This assumption may also be supported by our statistical analyses, which tested the potential correlation between specific values in each range of the first version of the color-coded markup and the type of transcription. Indeed, the analyses showed very weak and weak statistically significant strength of correlations between the type of transcription and the values in each range in the original markup, especially for the red range (where half of the tokens were correctly transcribed even if the system printed them out with a confidence score comprised between 70% and 89,99%).

When discussing these results, it is important to remember that 'correlation is not causation' (Winter, 2019: p. 70). Confidence scores are known to be affected by speaker variability (e.g., accents, speech rate) and acoustic/environmental factors (e.g., reverberation, ambient noise) (Jiang, 2005; Emara & Shaker, 2024). If speech signals collected from real-world contexts differ from the 'ideal' signal, the system will be less sure of its transcription, leading to a higher probability of errors (Li, 2018): in fact, the CIT analysis partially confirmed this statement.

4.6 Limitations and future directions

The primary limitation of this study lies in the structure of the test set and the criteria used to categorize the audio recordings for the analysis. As described in the Methods section, the test set was balanced according to the two main variables of interest: topic and language. The recordings were then categorized based on additional speaker- and environment-related factors that could have

influenced the performance of the ASR (Automatic Speech Recognition) system. However, in order to fully evaluate the impact of these additional factors on accuracy and confidence scores, it would have been beneficial to counterbalance the recordings according to these factors as well.

This limitation becomes evident when we revisit the results of the CIT analysis; as reported in the results and the discussion, one of the models struggled to clearly determine the effects of these factors on the performance of the ASR system. Future studies should take care also to predefine the secondary factors they wish to investigate and to be meticulous when choosing and categorizing the audio recordings.

4.7 Conclusion

Professionals, users, and researchers have highlighted the importance of providing accurate automatic captions to support speech processing and comprehension (e.g., Wald & Bain, 2008; Shimogori *et al.*, 2010; Romero-Fresco & Fresno, 2023). Despite the advancements in the robustness of ASR systems for processing and transcribing speech in real-world applications (e.g., O'Shaughnessy, 2024), these systems may operate poorly (in terms of both accuracy and confidence) when dealing with degraded speech signals and/or when used in environments with compromised acoustics (e.g., Del Rosso & Brambilla, 2022; Dua *et al*., 2023). Moreover, the confidence of the ASR system in its transcription decreases when dealing with speech recorded in adverse conditions. In these contexts, it could be helpful for users to visualize the degree of confidence the ASR system has via the integration of some graphical features in the text of automatic transcriptions/captions (e.g., Bain *et al*., 2002; Berke, 2017).

This study aimed to assess 1) the performance of an ASR system in transcribing audio recordings from real-world contexts, 2) the reliability of confidence scores to build a graphical feature (color-coded markups) that signals users how confident the ASR system is of its transcriptions, and 3) the degree of influence some factors have on both accuracy and confidence scores. Results highlighted an overall good performance and confidence of the ASR system; however, in line with previous research, the performance worsened when the system had to deal with degraded speech signals from a specific domain (the recording of an academic lecture delivered in English). Statistical analyses highlighted a weak correlation between transcription type (correct/incorrect) and confidence scores, also underlining the impossibility of establishing a threshold value associated with a specific type of transcription. Finally, the analysis performed on this test set allowed us to create two distinct versions of markups - V2 and V3 - to display the system's confidence. This development was based on the

findings from the corpus analysis and the results of the questionnaire administered to university students, as reported in Chapter 3.

The findings from this corpus analysis could initiate discussions on the implementation of ASR systems within educational settings (see Chapter 7).

In the next chapter, we will discuss the results of the third and final study of this doctoral project, which aimed to 1) explore the usefulness of displaying the confidence levels of the ASR system in automatic captions and 2) gather opinions and insights from L2 speakers of English regarding the three different display formats.

5 Color-coded markups in automatic captions: a pilot study

An investigation on the usefulness of displaying the ASR system's confidence to improve reliability on captions for L2 speakers of English

5.1 Introduction

In this chapter, we report on the third and last part of the research project.

After the administration of a questionnaire to university students to investigate their opinions and insights on the potential use of automatic captions in educational settings (Chapter 3) and the conduction of an analysis of a test set of ASR-generated transcriptions to evaluate the reliability of confidence scores (see Chapter 4), we developed two color-coded markups to be implemented in the text of captions to indicate the ASR system's confidence level. This chapter describes a pilot study we conducted to investigate the effects of automatic captions on speech processing and comprehension and to test the usefulness of presenting the text with color-coded markups to graphically display confidence in automatic captions within an educational setting.

5.2 Aims, design and research questions

We report on a study aimed at assessing:

- The effects of automatically generated captions on the speech processing and content comprehension of L2 speakers of English in an educational setting;

- The usefulness and potential benefits of showing users the degree of confidence the ASR system has in its transcriptions through a graphical feature (color-coded markup) in the text of automatic captions.

Specifically, we formulated the following research questions:

RQ1. Do errors in automatic captions affect speech processing and content comprehension in L2 speakers of English?

RQ2. Do color-coded markups affect content comprehension and attention in L2 speakers of English, hindering speech processing?

RQ3. What opinions do L2 speakers of English hold on the usefulness of displaying the confidence level of the ASR system through different (color-coded) markups?

Previous studies on the usefulness and effects of visualizing through graphical features the confidence level of the ASR system in its transcriptions have tested different types of characteristics of the display formats adopting a user-centered approach. Following this line of research, we conducted an experimental study to answer these empirical questions. We designed a between-subjects, one-way factorial study with four levels, each corresponding to a display format of captions, namely the classic format of captions (Classic), the display format with the original color-coded markup designed by the partner company (OG), and the display formats following the results of the analysis of the test set (V2) and the questionnaire administered to university students (V3) (see Chapter 4, for details and *Materials*, §4.3.2).

Color-coded markups can be informative of the confidence of the ASR system in its transcription and increase the reliability of users on the text of the automatic captions (e.g., Vertanen & Kristensson, 2008). However, the different colors in the text could increase the load imposed on the cognitive system (especially on attention), as users will be testing the formats in an already rich multimodal setting (Sweller, 2005). In this context, users may struggle to remember the color-coded schemes (Piquard-Kipffer *et al*., 2015), making them feel confused or distracted from listening (Berke *et al*., 2017) and potentially impacting their ability to understand the content of a lecture. Therefore, we predict that L2 speakers in the '*OG markup*' condition will perform worse on the comprehension task compared to those in the '*V2*' and '*V3 markup*' conditions. Additionally, we expect that the highest scores will come from participants who will watch the video in the baseline condition, which is the 'Classic captions' condition. This outcome is related to the number of colors and graphical features in

each display format: the OG markup, in fact, is made up of three colors (white, grey, and red) and two additional graphical features (underlined and strikethrough). In contrast, the V2 markup uses two colors (white and red) along with two graphical features (underlined and strikethrough), while V3 utilizes only one color (white) and the same two graphical features. Conversely, if users consider the markups to be informative and helpful for understanding the content of the lecture, we predict that participants assigned to the '*OG markup*' condition will perform better on the comprehension test than those in the baseline and other conditions. This prediction is based on the observation that the OG markup contains the highest number of graphical features, giving information about four different levels of confidence (it ranges from white: absolutely sure of its transcription, to red bar: absolutely not sure of its transcription). In contrast, the V3 markup provides the least information about the confidence level of the ASR system since it only has three levels of confidence, but only one color and two graphical features.

Errors in the transcriptions may also affect the performance of L2 speakers in the comprehension test (Cao *et al*., 2018). Following the corpus analysis of the automatically-generated transcriptions to assess the potential correlation of type of transcription (correct, incorrect) and confidence scores (see Chapter 4) and the results of Vertanen & Kristensson's study (2018), we hypothesize that the color-coded markups will help L2 speakers identify the errors in the transcriptions. In this context, participants assigned to the '*OG markup*' condition will perform better on the comprehension test and will find the markup to be the most reliable to identify the errors. Conversely, those assigned to the baseline condition will find this display format the least reliable for spotting transcription errors since it gives no graphical indication of the confidence of the ASR system.

5.3 Methods

5.3.1 Participants

Participants of this study were a small group of university students who participated in the first study (see Chapter 3). Sixteen native speakers of Italian and L2 speakers of English (13 F; M_{age} = 24 yo, SD = 2,63; age range: 21-31 years old) voluntarily agreed to participate in this study. The majority of participants were enrolled in degree programs related to *language sciences* or *linguistics*, while the rest of the participants were enrolled in various degree programs, ranging from *Archaeology* to *Electrical Energy Engineering*. *Table 57* provides a summary of participants' demographic data.

Table 54. *Participants' demographic data.*

N	Gender	Age	ENG AoA	Education	ENG Courses
16	12 F; 3 M; 1 ND	Mean: 24 yo SD: 2,6 Range: 21-31	Mean: 8 yo SD: 2,2 Range: 5-11	HS: 4 BA: 8 MA: 4	0: 1 1: 5 2: 1 3: 0 4: 1 5+: 8

Note. Labels in the *Education* column: HS: High School; BA: Bachelor's Degree; MA: Master's Degree. Label *ENG AoA*: age of acquisition of English; Label *ENG Courses*: number of courses of the degree program delivered in English.

We measured students' proficiency level of English and listening comprehension abilities with the same methods we used in the first study (see Chapter 3, §3.3.2). *Figure 57* shows the distribution of the self-rated proficiency level of participants. Seven participants (43,75%) self-reported their proficiency level at B2, seven participants C1 (43,75%), and two participants stated that their proficiency was attested at the C2 level (*N* = 12,5%). Overall, participants had an intermediate-to-advanced level of proficiency in English.

Figure 52. *Distribution of participants' level of proficiency of English.*

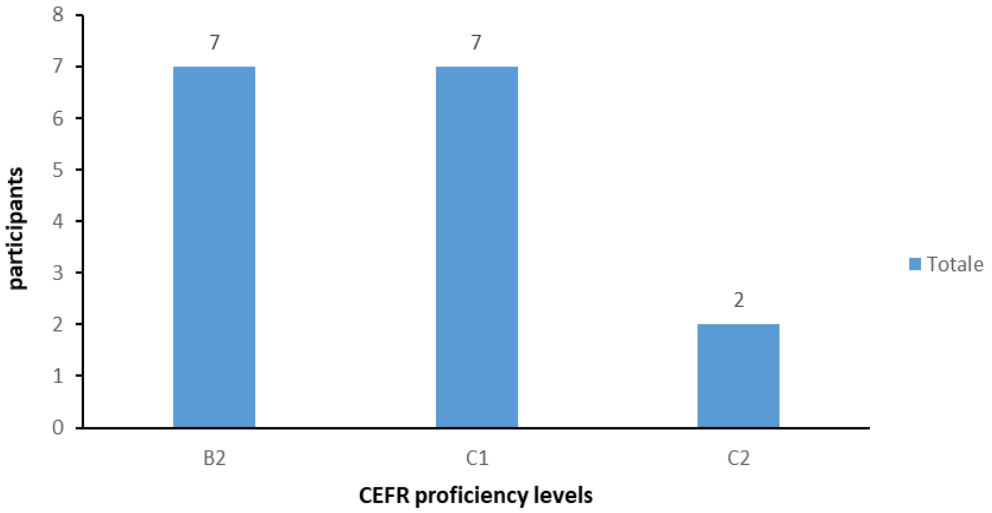


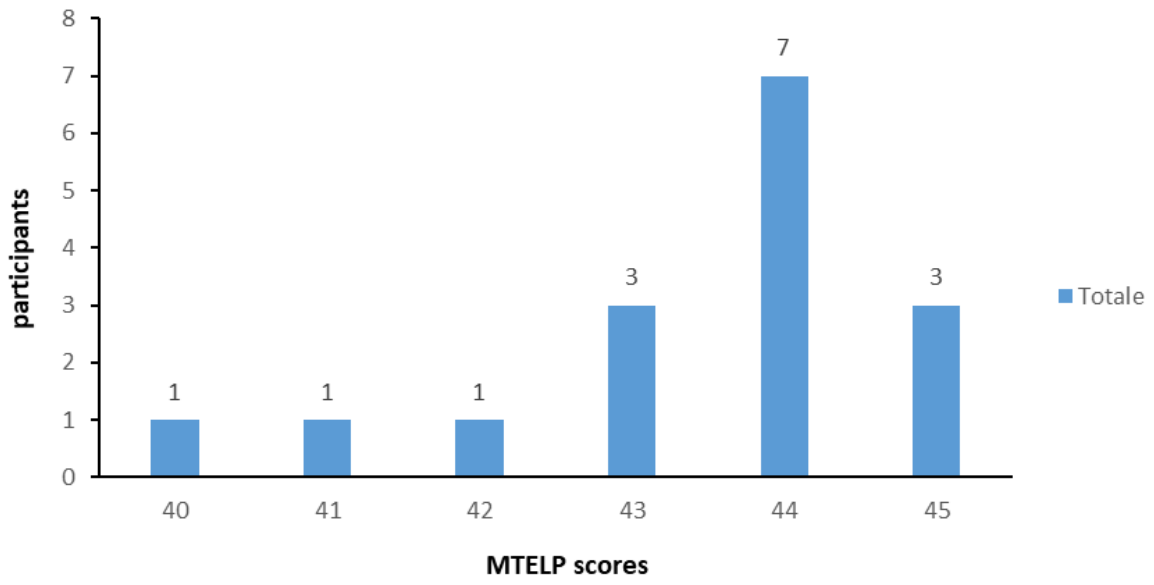
Table 58 summarizes data from the MTELP test that was administered to measure participants' listening abilities, while *Figure 58* shows the distribution of scores. The average MTELP score is 43,44 (SD = 1,41), while the range of scores is comprised between 40 and 45. Overall, these results indicate that participants have very good listening skills.

Table 55. *Descriptive statistics for participants' scores in the listening comprehension task.*

N	M_{MTELP}	SD_{MTELP}	$Median_{MTELP}$	$Range_{MTELP}$
16	43,44	1,41	44	40-45

Note. The participants' listening abilities were measured with a portion of the Michigan Test of Language Proficiency (MTELP).

Figure 53. *Distribution of the MTELP scores for all participants.*



5.3.2 Materials

Like the study reported in Chapter 3, participants completed the listening comprehension task from Michigan Test of Language Proficiency (MTELP) to assess their listening abilities and completed the questionnaire on their habits of watching audiovisual products in English and use of supporting written content (see Chapter 3, §3.3.2 for details).

5.3.2.1 Comprehension task

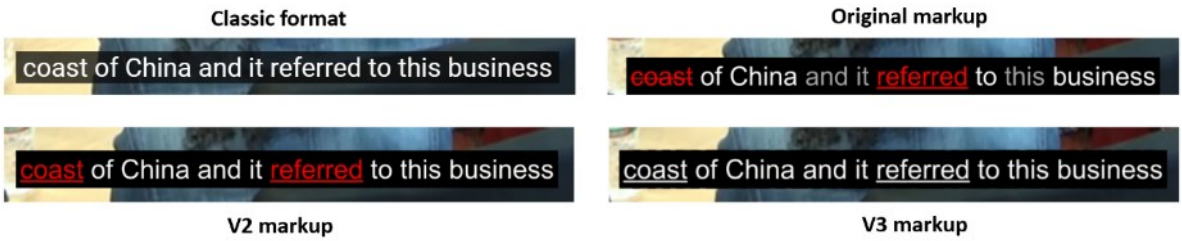
We designed a new questionnaire on Google Form aimed at investigating the impact of different markups in captions on comprehension of university students. The questionnaire contained a seminar-style video-lecture on Pidgins and Creole languages. The video was downloaded from the *MIT OpenCourseWare* website[55]. The video was chosen since it has some features that would make the listening process hard for students and the transcription less accurate from the ASR system. First, the main speaker in the video has a foreign accent while speaking in English. Second, a speaker stands too far from the microphone collecting speech, resulting in a degraded signal. These factors are the same factors we considered in our analysis in Chapter 4, which we know degraded the speech signal, lowering both accuracy and confidence of ASR systems. Moreover, these factors increase students' listening effort, impacting the listening process and have them rely on the textual input. Comprehension of the content of the video was assessed with a 10 multiple-choice questions task used by Venturini and colleagues in their 2022 study (see *Appendix F*). The questions aimed at testing not only the comprehension of the content, but also word recognition and retrieval. Automatic captions were generated using the partner company's ASR system (see §4.4.1 for details on the system). Two transcriptions were used: one for the Classic and OG markups, and one for V2 and V3 markups. The color-coded markups were created manually by the researcher using the open-source software Aegisub[56] (version 3.2.2; Montero & Hansen, 2008) and hardcoded in the video. The text was displayed in a small, rectangular black box in a *sans serif* font[57] at the bottom center of the screen (*Figure 59*).

The baseline condition was represented by the classic display format of captions, where text was presented in white in a black box to enhance text readability. The other three experimental markups

[55] MIT OpenCourseWare (2025). https://ocw.mit.edu. DeGraff, M. (2017) Lesson 1. Do "Pidgins" exist? Do creoles come from pidgin? Creole Languages and Caribbean Identities - MIT OpenCourseWare 24.908, Spring 2017.
[56] Montero & Hansen (2008). Aegisub [Computer software]. Retrieved from https://aegisub.org.
[57] We chose to use a *sans serif* font since *serif* fonts can be challenging to read on screen because they contain typographical embellishments that may impair text legibility (Rajendran *et al*., 2013: p. 7).

– namely, OG, V2, and V3 – were created during the second part of the doctoral project (see Chapter 4, §4.5.5 for details). An example of the four display formats is shown in *Figure 59* below.

Figure 54. *Screenshots from the video "Pidgin and Creole Languages" containing the four display formats.*



For both transcriptions, we calculated the WER score using the *JiWER* package[58] (and related packages) on *Google Colab*[59] (see *Appendix D* for the Python code). The WER score was attested respectively at 33% and 34% for the two transcriptions. The score was higher than the threshold signaled in their study by Shimogori and colleagues (2010 - who stated that a WER score below 20% would benefit L2 speakers' comprehension) and Cao and colleagues (2018 – the majority of the participants in this study found accurate enough the transcriptions with a WER score below 15%). We also used *sclite* to compute the number of substitutions, insertions, and deletions in the two transcripts. The number of correctly transcribed tokens for the file used to create the Classic and OG markups was attested at 68,8% ($N = 1127$) and at 68,9% ($N = 1128$) for the file used to create the V2 and V3 display formats. Substitutions were ~20% of errors for the two transcripts, insertions were ~2% of errors, and ~11% of errors. *Table 59* summarizes the results of the analysis conducted with *sclite*.

Lastly, an analysis of the five most frequent confusion pairs[60] revealed that the ASR system in both transcriptions misrecognized the same words with the same frequency. In general, technical, low-frequency words such as 'Pidgin(s)', 'Creole', 'Spanglish', and proper nouns were misrecognized by the ASR system and substituted with other nouns, namely 'pigeon(s)', 'career', and 'stingless/things'. These results are in line with the analysis of the test set we presented in Chapter 4.

[58] JiWER. (2025). Retrieved March 4, 2025, from https://github.com/jitsi/jiwer?tab=readme-ov-file.
[59] Google. (2024). Google Colaboratory. Retrieved December 5, 2024, from https://colab.research.google.com/.
[60] The confusion pairs list is one of the analyses that sclite returns to users while detecting the differences in the REF and HYP texts. The script identifies the most frequently misrecognized words in the REF by the ASR system and prints out the errors in a list.

Table 56. *Results of the analysis conducted with the JiWER package on Google Colab on the two files containing the automatic captions shown in the comprehension task.*

	Captions Classic & OG Markups	**Captions V2 & V3 Markups**
Accuracy	1127 (68,8%)	1128 (68,9%)
Substitutions	330 (20,1%)	335 (20,5%)
Insertions	32 (2,0%)	35 (2,1%)
Deletions	181 (11,1%)	175 (10,7%)
WER score	0.33	0.34

Table 60 presents the main characteristics of the automatic captions used in the study. The report was generated using the software Subtitle Edit[61] (version 4.0.7). The two transcripts, even though generated by the same ASR system, had some differences in the output (especially in the number of total words and characters): this could be due to the higher number of insertion errors (see *Table 59* above). The average presentation time (17 cps) of automatic captions was above the recommended average subtitling speed of 12 cps signaled by Díaz-Cintas and Remael (2007) and was the same for both transcripts. This speed therefore could be too high for users to read comfortably and without losing chunks of text.

[61] Lynge Olsson (2024). Subtitle Edit [Computer software]. Retrieved from https://www.nikse.dk/subtitleedit.

Table 57. *Analysis of the automatic captions for both transcripts.*

	Captions Classic & OG Markups	**Captions V2 & V3 Markups**
Total captions lines	224	225
Total words	1489	1498
Total characters	16815	16894
Duration (sec)	07:51,481	07:52,201
Average presentation speed (cps)	16,9	16,9
Average presentation speed (wpm)	201,081	201,858
Average caption time on screen (sec)	2,105	2,099
WER score	0.33	0.34

5.3.2.2 Questionnaires

We asked participants to fill out two questionnaires.

The first questionnaire we asked participants to fill out was the second version of the one on viewing habits and supporting written content use we used for the first study of this project (see Chapter 3). The questionnaire was revised to explore in greater detail why L2 speakers have utilized in the past (when their proficiency level in English was lower) and currently use various types of supporting written content to facilitate speech processing, language comprehension, and learning (see *Appendix A.II*).

The second questionnaire aimed at investigating the opinions and insights of university students on automatic captions and the display formats shown in the video. The questionnaire was divided into three sections. The first section asked university students to state their difficulties while watching the video and to assess the impact of errors in the automatic captions on their attention and listening

process. The second part aimed at collecting students' opinions and insights about the display format they saw in the video: questions focused on investigating the impact the markups on their attention and overall listening process, as well as their preferences for other display formats. The third and last section of the questionnaire aimed at investigating the potential use of this type of automatic captions and display formats in class during lectures delivered in English. The questionnaire had a total of thirty questions (thirty-two multiple-choice questions and one optional, open question). Like the first questionnaire, this one was created and disseminated through Google Forms to be administered remotely.

5.3.3 Procedure

Prior to participating in the experimental activities, all participants signed an informed consent form. Then, they were randomly assigned to one of the conditions (*ClassicCaps*, *OG*, *V2*, *V3*) and asked to complete the three activities.

Like in the first study, first we asked participants to complete the listening comprehension portion of the MTELP and the questionnaire on viewing habits and captions use, which included questions on the potential use of live captioning in academic settings (see Chapter 3, §3.3.2). Then, students had to complete the questionnaire containing the comprehension task. First, participants had to watch the video and answer the ten multiple-choice questions test. Then, they completed the questionnaire aimed at investigating their preferences, opinions and insights on the display format they were presented with and the use of automatic captions in class during lectures delivered in English.

The activities included in this study were approved by the Ethics Committee of Ca' Foscari University, Venice.

5.4 Results

In this section, we discuss the results of the two questionnaires and the comprehension task.

5.4.1 Questionnaire on participants' viewing habits and use of supporting written content to aid speech processing and comprehension

Participants were asked to fill out the questionnaire on their viewing habits and use of supporting written content to support speech processing and learning of English (see Chapter 3, §3.3.2). Results highlighted that most participants habitually use supporting written content while watching audiovisual products in English (*Table 61*), preferring captions and intralingual subtitles to interlingual subtitles (*Table 62*).

Table 58. *Answers (count and percentage) for the question "Do you use captions/subtitles when watching audiovisual content in English?".*

Answer	Total (n°)	Total (%)
Yes	13	81,25%
No	3	18,75%

Table 59. *Answers (count and percentage) for the question "Which type of supporting written content (captions, subtitles, etc.) do you prefer using when watching audiovisual content in English?".*

Answer	Total (n°)	Total (%)
Captions	9	56,25%
Intralingual subtitles	6	37,5%
Interlingual subtitles	4	25%
It depends (on some factors)	1	6,25%
I don't use captions/subtitles	3	18,75%
Other	1	6,25%

They mainly use the supporting written content of their choice to make sure to understand the content of what they are watching (*option H, N* = 10, 62,5%), but also to learn new words (*option F*, *N* = 6, 37,5%) and improve their pronunciation (*option G, N* = 4, 25%) (*Table 63*). They also use captions

or subtitles to help them listening to varieties of English they are unfamiliar with ("I only use captions when listening to a particular dialect" and "It depends on the content and especially the English variety. I use subtitles for varieties I'm not used to hear often, such non-standard British English") or to help them lower the degree of listening effort the second language imposes on cognition ("It requires less effort to read than to listen").

Table 60. *Answers (count and percentage) for the question "Why do you prefer that supporting written content (captions, subtitles, etc.)? Could you please motivate your previous answer?".*

Option	Total (n°)	Total (%)
A	1	6,25%
B	1	6,25%
C	2	12,5%
D	3	18,75%
E	-	-
F	6	37,5%
G	4	25%
H	10	62,5%
I	-	-
L	4	25%
M	3	18,75%
N	2	12,5%
O	2	12,5%
P	3	18,75%
Other	5	31,25%

Note. Labels in the *Answer* column: *Option A*: Yes: I find it hard to concentrate if I have pay attention both to the audio and the text; *Option B*: Yes: I do not turn captions/subtitles on because I try to improve my listening skills, and I don't think that written support helps; *Option C*: I don't use captions/subtitles because I prefer listening to what is being said

rather than reading the text; *Option D*: No: I don't use captions/subtitles because I don't need them; *Option E*: I need to identify where words begin and end by reading the written transcript; *Option F*: I want to learn new words; *Option G*: I want to improve/learn the pronunciation of words; *Option H*: I want to make sure I understand the content of what I am watching/listening; *Option I*: I want to know what's the translation in my native language of the words in the speech; *Option L*: I want to know what's the meaning of a word in my native language; *Option M*: I prefer reading each word as soon as they are pronounced by speakers; *Option N*: I prefer reading the exact words pronounced by speakers; *Option O*: I prefer reading the text in my native language; *Option P*: It requires less effort to read the text in my native language.

All participants used to watch audiovisual products in English, with the majority of them preferring to use both intralingual and interlingual subtitles to improve their knowledge of English when their proficiency was lower (*N* = 16) (*Table 64* and *Table 65*, *Past* column). The majority of them still watch audiovisual products in English to improve their proficiency (*N* = 12, 75%), but way less students use supporting written content to improve their knowledge, with the majority of them rarely using captions and intralingual subtitles (*Table 64* and *Table 65*, *Present* column).

Table 61. *Answers (count and percentage) for the questions "Did you use to watch audiovisual content in English to improve your knowledge of the language when your proficiency in English was lower?" (Past column) and "Do you watch audiovisual content in English to improve your knowledge of English in the present day?" (Present column).*

Answer	Past (total n°, %)	Present (total n°, %)
Yes	16 (100%)	12 (75%)
No	-	4 (25%)

Table 62. *Answers (count and percentage) for the questions " Did you use captions/subtitles when watching audiovisual content in English to improve your knowledge of the language when your proficiency in English was lower?" (Past column) and "Do you use captions/subtitles to improve your knowledge of English in the present day?" (Present column).*

Option	Past (total n°, %)	Present (total n°, %)
Captions	Rarely (*N* = 6, 37,5%)	Rarely (*N* = 7, 43,75%)
Intralingual subtitles	Rarely (*N* = 6, 37,5%) Frequently (*N* = 6, 37,5%)	Rarely (*N* = 6, 37,5%)
Interlingual subtitles	Frequently (*N* = 8, 50%)	Never (*N* = 9, 56,25%)

Note. The table reports the option selected by the highest number of participants.

Table 66 summarizes the reasons why participants use supporting written content to improve their knowledge of English in the past when their proficiency was lower and today. In the past, most participants used captions/subtitles especially to become familiar with the least known accents of English (*Option E*: *N* = 13, 81,25%). A high percentage of participants used these supporting written content to spot unknown words (*Option B*: *N* = 10, 62,5%) and learn both their pronunciation (*Option D*: *N* = 9, 56,25%) and graphic form in English (*Option F*: *N* = 9, 56,25%) or its meaning in their L1 (*Option A*: *N* = 7, 43,75%). In the present, many participants still use supporting written content to familiarize themselves with less known accents[62] (*Option E*: *N* = 8, 50%) or learn the pronunciation of words (*Option D*: *N* = 6, 37,5%). However, a small group of participants stated that they do not use captions/subtitles to improve their knowledge of English (*Option H*: *N* = 6, 37,5%).

Table 63. *Answers (count and percentage) for the questions "Why did you use captions/subtitles to learn English when your proficiency was lower?" (Past column) and "Why do you use captions/subtitles to learn English in the present day?" (Present column).*

Option	Past (total n°, %)	Present (total n°, %)
A	7 (43,75%)	2 (12,5%)
B	10 (62,5%)	5 (31,25%)
C	7 (43,75%)	3 (18,75%)
D	9 (56,25%)	6 (37,5%)
E	13 (81,25%)	8 (50%)
F	9 (56,25%)	4 (25%)
G	1 (6,25%)	-
H	1 (6,25%)	6 (37,5%)

[62] A participant added the following statement regarding the use of supporting written content to familiarize with different accents: “I also enjoy watching films or TV series in peculiar accents to have further exposure to those, and in that case, I definitely need captions/subtitles to catch all words.” On the same note, another participant stated that “I only use captions if I hear a word I’m not familiar with, either because it's a technical term or a common word pronounced with a strong accent.”

Other	-	1 (6,25%)

Note. Labels in the *Answer* column: *Option A*: I could/can easily spot words I don't know and learn them when the translation is provided in the subtitles; *Option B*: I could/can spot words I don't know and learn them in English; *Option C*: I could/can learn the meaning of a word by reading the supporting written text; *Option D*: I could/can learn the pronunciation of a word; *Option E*: It helped/helps me becoming acquainted with English accents I am not familiar with; *Option F*: I could/can learn how a word is written in English; *Option G*: The text in English helped/helps me identifying where words pronounced by speakers begin and end; *Option H*: I don't use captions/subtitles to improve my proficiency in English.

Two participants reflected on their habits of using the different types of supporting written content. Both participants mentioned that they use intralingual subtitles when watching audiovisual content in their L1 (P1: "I use subtitles and/or captions almost 100% of the time, not only when consuming content in English but also in my native language (Italian), as it helps me process what I am hearing"; P2: "Sometimes I like using captions and subtitles even for videos in my native language"). Specifically, P1 stated that

"I use subtitles and/or captions almost 100% of the time, not only when consuming content in English but also in my native language (Italian), as it helps me process what I am hearing. Regarding translated subtitles, I do not usually use them for English, as I find it a bit distracting to read in one language and listen in another, but I might turn them on temporarily if I hear a sentence that I think might have been particularly tricky to localize, just to see how the translators handled that part."

Similarly, P2 stated that

"I think I've become used to use captions and subtitles because it requires less effort to read than to listen and I can watch videos with a lower volume (because I can easily read what I cannot hear e.g. if there's noise or I'm in a crowded space). Sometimes I like using captions and subtitles even for videos in my native language."

5.4.2 Comprehension Task

Scoring was similar to the one conducted by Venturini and colleagues in their study (2022). We assigned one point to each correct answer, and zero points to each wrong answer. The final score was the sum of the points divided by the total number of questions (percent correct).

The statistical analyses were conducted using R software (R Core Team, 2020).

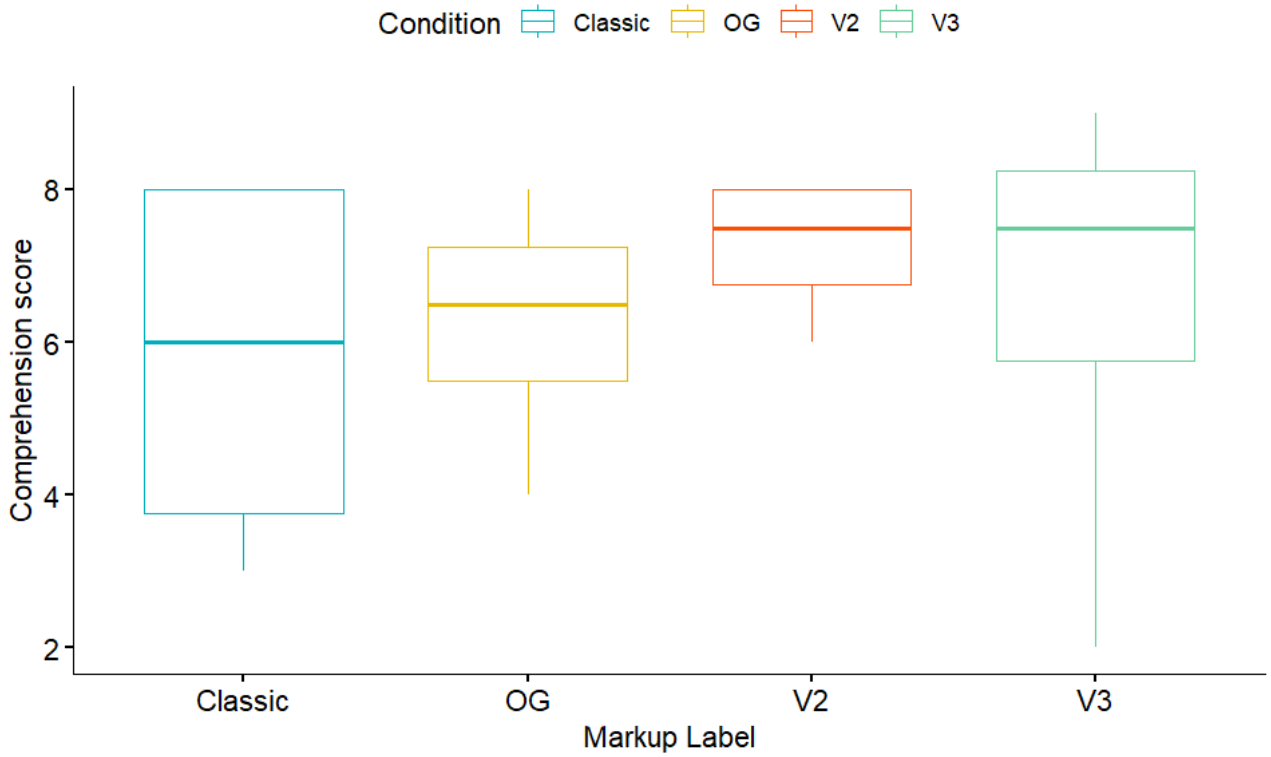
Overall, participants' average comprehension score was attested at 6,4 (SD = 0,625). *Table 67* shows the descriptive statistics of the results of the comprehension task by condition (classic, OG, V2, AND V3). Participants performed better when presented with the V2 markup (M_{Score} = 7.25, SD = 1), while participants performed worse when presented with the OG markup (M_{Score} = 5.75, SD = 2.6).

Table 64. *Descriptive statistics of the results of the comprehension task by condition (classic, OG, V2, and V3).*

Condition	N	M_{Score}	M_{Score} (% correct)	SD	SE	CI
Classic	4	5.75	0.6	2.6	1.3	4.2
OG	4	6.25	0.6	1.71	0.8	2.7
V2	4	7.25	0.7	1	0.5	1.5
V3	4	6.50	0.6	3.1	1.5	4.9

Figure 60 shows the boxplots with the mean comprehension score by condition (classic display format, OG, V2, and V3). We compared these scores using the one-way ANOVA to check any statistically significant difference.

Figure 55. *Average comprehension scores by condition (classic, OG, V2, V3).*



We fitted a linear model to predict if different experimental markups significantly predicted comprehension score (percent correct) (formula: *ScorePercent ~ caption.e*). The three experimental markups (*OG*, *V2*, and *V3*) were contrasted with the *Classic* condition, and follow-up pairwise comparisons contrasted each condition to one another. *Table 68* summarizes the outcome of the linear model.

Table 65. *Results of the linear model.*

	ScorePercent		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.58	0.33 – 0.82	**<0.001**

caption.e [2]	0.05	-0.30 – 0.40	0.760
caption.e [3]	0.15	-0.20 – 0.50	0.366
caption.e [4]	0.07	-0.27 – 0.42	0.647
Observations	16		
R^2 / R^2 adjusted	0.071 / -0.161		

Note. Labels in the column *Predictors*: caption.e [2] = OG markup; caption.e [3] = V2 markup; caption.e [4] = V3 markup.

The overall fit of the model explains a statistically not significant and weak proportion of variance (R^2 = 0.07, F(3, 12) = 0.31, p = 0.821, adj. R^2 = -0.16). The model's intercept, corresponding to Markup = Classic display format, is at 0.58 (95% CI [0.33, 0.82], t(12) = 5.09, p < .001). Within this model:

- The effect of markup *OG* is statistically non-significant and positive (beta = 0.05, 95% CI [-0.30, 0.40], t(12) = 0.31, p = 0.760; Std. beta = 0.05, 95% CI [-0.30, 0.40]);
- The effect of markup *V2* is statistically non-significant and positive (beta = 0.15, 95% CI [-0.20, 0.50], t(12) = 0.94, p = 0.366; Std. beta = 0.15, 95% CI [-0.20, 0.50]);
- The effect of markup *V3* is statistically non-significant and positive (beta = 0.07, 95% CI [-0.27, 0.42], t(12) = 0.47, p = 0.647; Std. beta = 0.07, 95% CI [-0.27, 0.42]).

In sum, none of the experimental markups tested in this study predict comprehension scores.

Finally, we performed post-hoc analysis using the pairwise comparison of estimated marginal means of linear trends method. *Table 69* summarizes the results of this analysis.

Table 66. *Results of the pairwise estimated marginal means of linear trends method.*

Comparison	estimate	SE	df	t ratio	p value
Classic - OG	-0.050	0.16	12	-0.313	0.988
Classic – V2	-0.150	0.16	12	-0.939	0.785
Classic – V3	-0.075	0.16	12	-0.469	0.964
OG – V2	-0.100	0.16	12	-0.626	0.921
OG – V3	-0.025	0.16	12	-0.156	0.998
V2 - V3	0.075	0.16	12	0.469	0.964

Results highlight no statistically significant differences between conditions; however, a trend is evident. Negative values in the “estimates” column indicate respectively that participants who were assigned the experimental markups performed better (measured in higher comprehension scores) than those who watched the video with classic captions. Similarly, participants who were assigned with the markups V2 and V3 performed better than those who were assigned with the OG markup. Lastly, the positive estimates indicate that participants assigned to the V2 condition outperformed those assigned to the V3 condition. Overall, the V2 condition tentatively predicts better comprehension scores in comparison to the other conditions.

An analysis of the questions that participants frequently answered incorrectly revealed a potential influence of errors in the automatic captions on content comprehension. Specifically, two questions were critical. Question n° 1 asked participants the definition of "creole languages", while question n° 8 asked to specify which structure pidgin languages have. In both cases, most of the answers selected by participants were possibly influenced by the text of captions. For Question n° 1, three participants (18,75% - among those, two did not have any previous knowledge of the topic) selected the options where the term 'pidgin(s)' was substituted by the term 'pigeon(s)'. Similarly, in Question n° 8, thirteen participants (81,25%) selected one of the two answers where the negative morphemes were inserted ('possible' was transcribed as 'impossible') or deleted ('un-language-like' was transcribed as 'language') compared to the original linguistic element in the speech.

5.4.3 Questionnaire on opinions and insights of university students on automatic captions and the experimental markups

In this paragraph we will discuss the results from the questionnaire on participants' opinions on the display format they saw in the video. Each question will be discussed on its own.

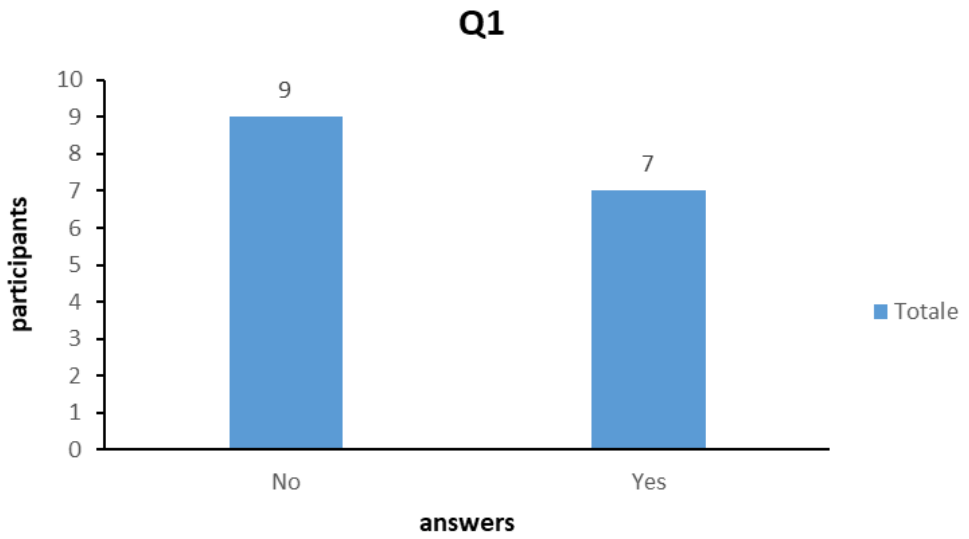
Q1. *Did you already know something about the topic (Pidgin and Creole languages) before watching the video?*

The majority of participants did not have any previous knowledge of the topic discussed in the video (*N* = 9, 56,25%) (*Table 70* for details and *Figure 61* to see the distribution of answers).

Table 67. *Number (count, percentage) of answers for Question n° 1 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
No	9	56,25%
Yes	7	43,75%

Figure 56. *Distribution of answers for Question n° 1 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



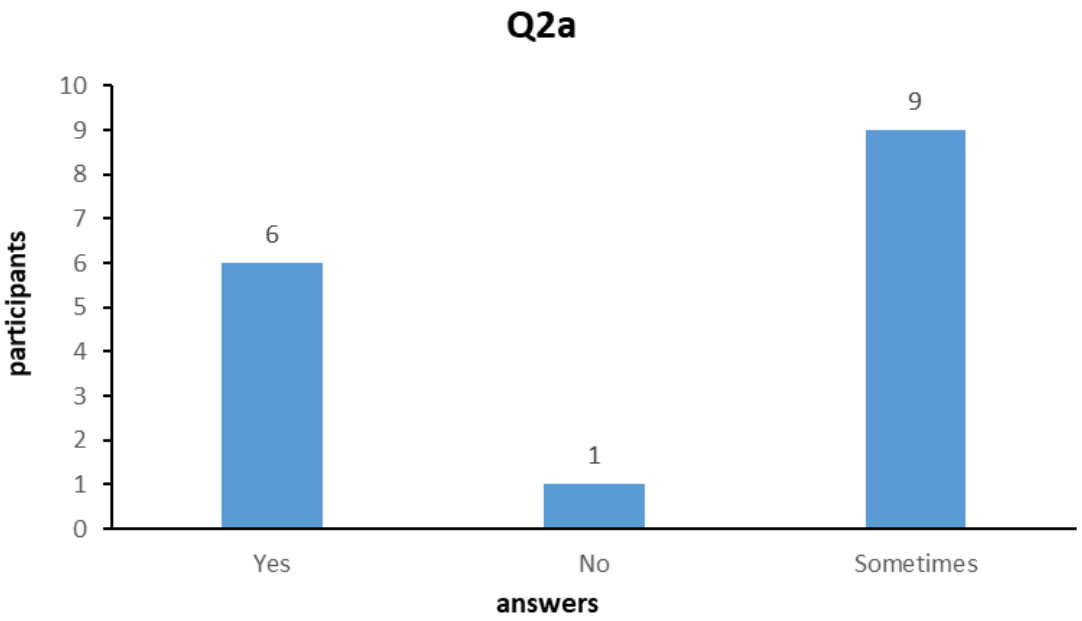
Q2.*Did you have a hard time understanding the content of the clip? Why? Could you motivate your previous answer? Choose all that apply to you.*

Almost all participants had trouble understanding the content of the clip ('Yes': *N* = 6, 37,5%; 'Sometimes': *N* = 9, 56,25%) (*Table 71* and *Figure 62*).

Table 68. *Number (count, percentage) of answers for Question n° 2a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
No	1	6,25%
Yes	6	37,5%
Sometimes	9	56,25%

Figure 57. *Distribution of answers for Question n° 2a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



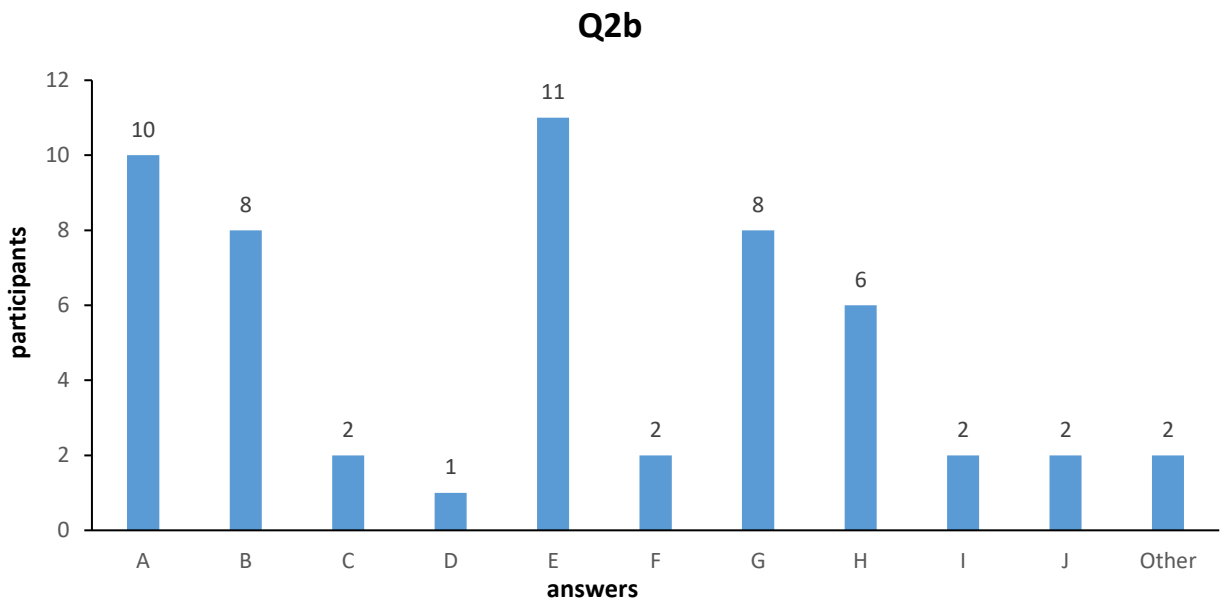
The reason behind these difficulties mainly lie in the fact that many participants were not familiar with the topic discussed in the video clip (*Option A*: *N* = 10, 62,5%). They had to pay attention to the speakers (*Option E*: *N* = 11, 68,75%) because they had trouble recognizing some words - for example, proper nouns - (*Option B*: *N* = 8, 50%) and were not familiar with the speakers' accents (*Option G*: *N* = 8, 50%). Six participants (37,5%) stated that their troubles with comprehension were related to the fact that the quality of the audio was bad (*Option H*). *Table 72* summarizes the answers for Question n° 2b, *Figure 63* shows the distribution of answers.

Table 69. *Number (count, percentage) of answers for Question n° 2b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
A	10	62,5%
B	8	50%
C	2	12,5%
D	1	6,25%
E	11	68,75%
F	2	12,5%
G	8	50%
H	6	37,5%
I	2	12,5%
J	2	12,5%
Other	2	12,5%

Note. Labels in the *Answer* column: *Option A*: I was not familiar with the topic; *Option B*: I had troubles recognizing some words; *Option C*: I didn't know enough vocabulary; *Option D*: It was not difficult to me; *Option E*: I had to pay close attention to the speakers' speech; *Option F*: It was hard to follow the speakers' interactions; *Option G*: I wasn't familiar with speakers' accents; *Option H*: The audio quality of the clip was bad; *Option I*: Speakers were talking too fast, I couldn't keep up with the conversation; *Option J*: I have some knowledge of the topic discussed in the clip; *Other*: Other answer.

Figure 58. *Distribution of answers for Question n° 2b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in each column: *Option A*: I was not familiar with the topic; *Option B*: I had troubles recognizing some words; *Option C*: I didn't know enough vocabulary; *Option D*: It was not difficult to me; *Option E*: I had to pay close attention to the speakers' speech; *Option F*: It was hard to follow the speakers' interactions; *Option G*: I wasn't familiar with speakers' accents; *Option H*: The audio quality of the clip was bad; *Option I*: Speakers were talking too fast, I couldn't keep up with the conversation; *Option J*: I have some knowledge of the topic discussed in the clip; *Other*: Other answer.

Q3. *Did the captions help you understand the content of the clip? Why? Please, motivate your previous answer. Choose all that apply to you.*

Overall, 87,5% ($N = 14$) of participants did not find captions helpful to understand the content of the video clip (*Table 73* and *Figure 64*).

Table 70. *Number (count, percentage) of answers for Question n° 3a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
No	14	87,5%
Yes	2	12,5%

Figure 59. *Distribution of answers for Question n° 3a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

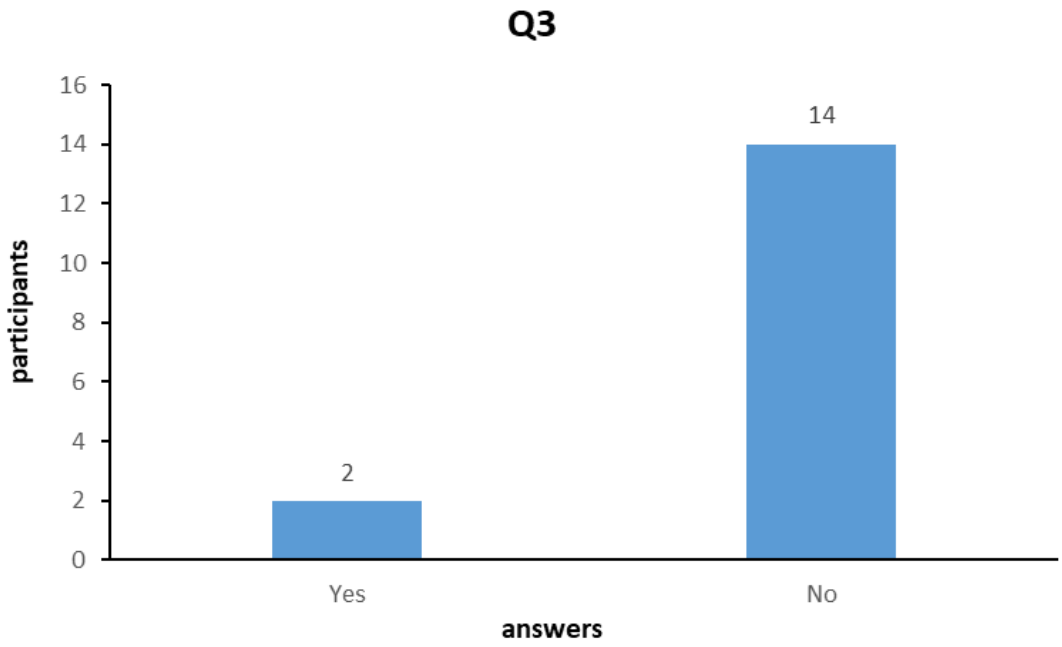


Table 74 summarizes the answers for Question n° 3b. Almost all participants stated that captions had too many errors (*Option C*: *N* = 12, 75%) and they distracted *("I found myself constantly having to focus on looking at the captions, as my instinct was to avoid looking at them since I felt they were hindering my understanding - e.g. I knew they were talking about pidgins, but looking at the captions made me hear pidgeons"*) and/or annoyed them (*Option G*: *N* = 14, 87,5%). A group of participants also found the font of the text of the captions distracting (*Option M*: *N* = 6, 37,5%) or annoying (*Option O*: *N* = 3, 18,75%). A very low number of participants found captions helpful to recover words they missed (*Option D*: *N* = 2, 12,5%) and understand speakers since they were not familiar with the accents (*Option B*: *N* = 2, 12,5%). In this regard, one of the participants also added that

"captions helped me since I was not familiar with the accents, but sometimes the errors distracted me." Finally, a group of participants found captions redundant (*Option J*: *N* = 2, 12,5%), and they would prefer reading the slides instead of captions (*Option Q*: *N* = 5, 31,25%) (see also *Figure 65* for the distribution of answers).

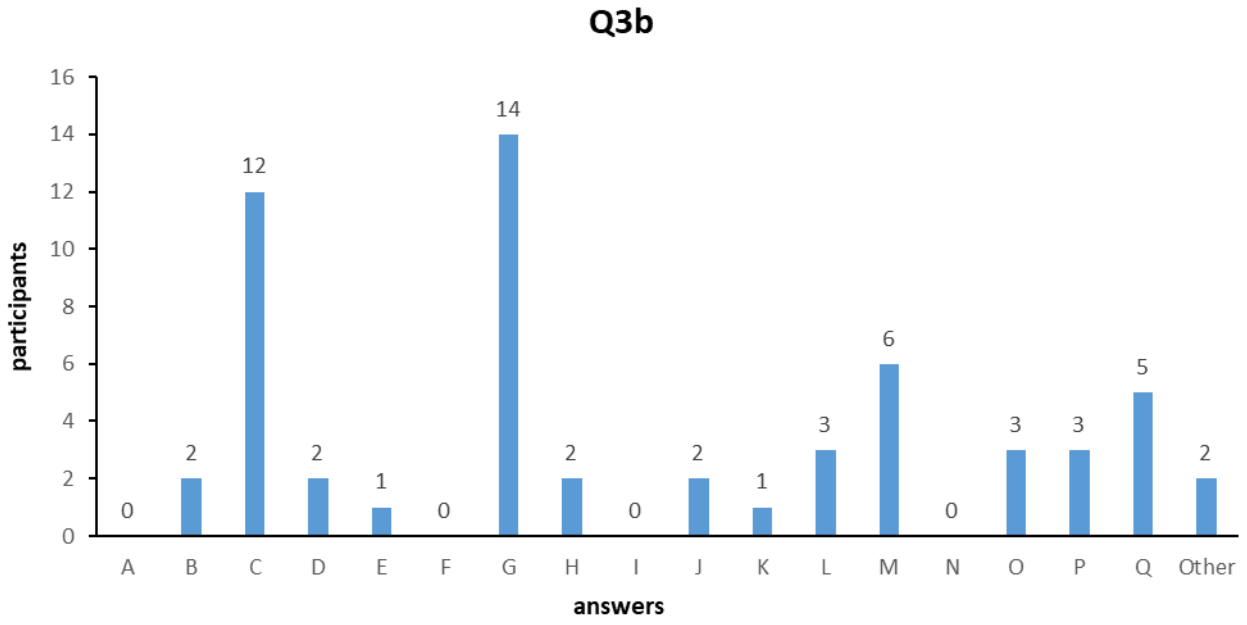
Table 71. *Number (count, percentage) of answers for Question n° 3b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
A	-	-
B	2	12,5%
C	12	75%
D	2	12,5%
E	1	6,25%
F	-	-
G	14	87,5%
H	2	12,5%
I	-	-
J	2	12,5%
K	1	6,25%
L	3	18,75%
M	6	37,5%
N	-	-
O	3	18,75%
P	3	18,75%
Q	5	31,25%

Other	2	12,5%

Note. Labels in the *Answer* column: *Option A*: I could rely on captions to make sure I understood the content of the clip; *Option B*: I could rely on captions to understand the speakers since I was not familiar with the different accents; *Option C*: The captions had too many errors - I could not rely on the text to improve comprehension; *Option D*: I could rely on captions to recover words I missed for some reason; *Option E*: I could rely on captions since the quality of the audio was bad; *Option F*: Captions helped me pay more attention to the discussion; *Option G*: The errors in the captions distracted/annoyed me; *Option H*: I think my knowledge of English was too low to fully understand the content of the clip; *Option I*: I think my knowledge of English was high enough to fully understand the content of the clip without captions; *Option J*: Captions were redundant; *Option K:* I look at the speakers' faces while they speak to follow the discussion, I don't need captions; *Option L:* I prefer listening to speakers talking; *Option M:* The font of the text (e.g., colors, underlined, etc.) of the captions distracted me; *Option N:* I could rely on captions to recognize words while the speakers were talking; *Option O:* The format with which captions were displayed annoyed me; *Option P:* I could not concentrate on the speakers' voices, captions kept distracting me; *Option Q:* I prefer reading the text in the slides (if provided in this setting) rather than reading the captions; *Other*: Other answer.

Figure 60. *Distribution of answers for Question n° 3b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in each column: *Option A*: I could rely on captions to make sure I understood the content of the clip; *Option B*: I could rely on captions to understand the speakers since I was not familiar with the different accents; *Option C*: The captions had too many errors - I could not rely on the text to improve comprehension; *Option D*: I could rely on captions to recover words I missed for some reason; *Option E*: I could rely on captions since the quality of the audio was bad; *Option F*: Captions helped me pay more attention to the discussion; *Option G*: The errors in the captions distracted/annoyed me; *Option H*: I think my knowledge of English was too low to fully understand the content of the clip without captions; *Option I*: I think my knowledge of English was high enough to fully understand the content of the clip without captions; *Option J*: Captions were redundant; *Option K:* I look at the speakers' faces while they speak to follow the discussion, I don't need captions; *Option L:* I prefer listening to speakers talking; *Option M:* The font of the text (e.g., colors, underlined, etc.) of the captions distracted me; *Option N:* I could rely on captions to recognize words while the speakers were talking; *Option O:* The format with which captions were displayed annoyed me; *Option P:* I could not concentrate on the speakers' voices, captions kept distracting me; *Option Q:* I prefer reading the text in the slides (if provided in this setting) rather than reading the captions; *Other*: Other answer.

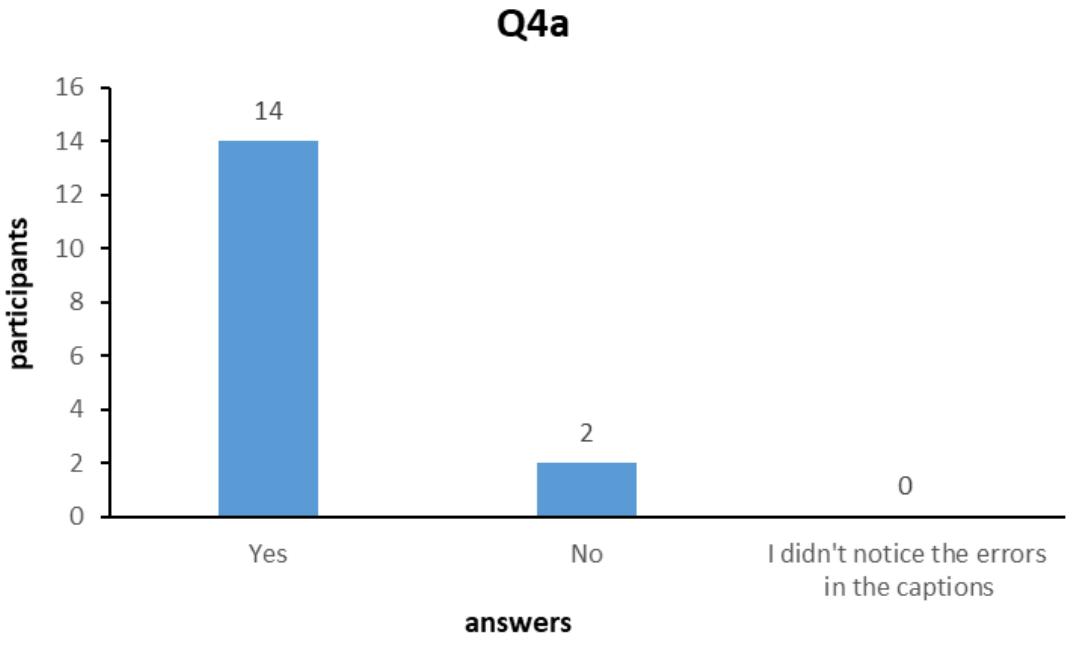
Q4. *It is common that captions generated by ASR systems - like the ones in the video you previously watched - contain errors in the transcription of speech. Did finding errors in the captions affect your listening experience? How? Please, motivate your previous answer. Choose all that apply to you.*

The majority of participants stated that errors in the text of captions affected their listening process (*N* = 14, 87,5%) (*Table 75* and *Figure 66*). It is remarkable that all participants spotted errors in the automatic captions, suggesting that at some point they relied on the text to aid speech processing.

Table 72. *Number (count, percentage) of answers for Question n° 4a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
No	2	12,5%
Yes	14	87,5%
I didn't notice the errors in the captions	-	-

Figure 61. *Distribution of answers for Question n° 4a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



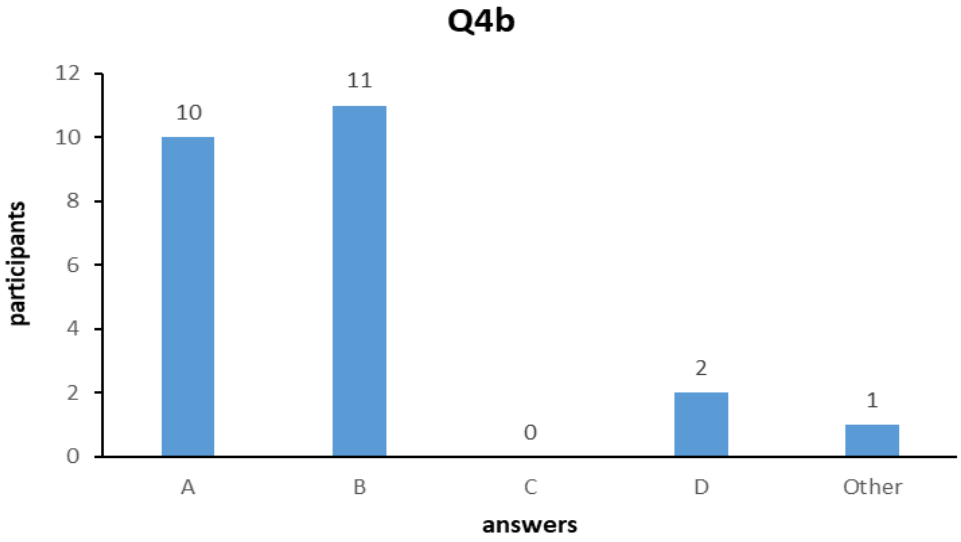
Errors in the text of automatic captions mainly confused (*Option A: N* = 10, 62,5%) and distracted participants from listening to speakers talking in the video (*Option B: N* = 11, 68,75%). Only two participants ignored the captions and only focused on listening (*Option D:* 12,5%) (*Table 76* and *Figure 67*). One participant also added that *"Once I realized the captions were not reliable, I ignored them altogether and focused on what the speaker was saying."*

Table 73. *Number (count, percentage) of answers for Question n° 4b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
A	10	62,5%
B	11	68,75%
C	-	-
D	2	12,5%
Other	1	6,25%

Note. Labels in the *Answer* column: *Option A*: Errors confused me; *Option B*: Errors distracted me from listening to the speakers; *Option C*: I didn't pay attention to the errors; *Option D*: I focused on listening rather than reading the captions; *Other*: Other answer.

Figure 62. *Distribution of answers for Question n° 4b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in each column: *Option A*: Errors confused me; *Option B*: Errors distracted me from listening to the speakers; *Option C*: I didn't pay attention to the errors; *Option D*: I focused on listening rather than reading the captions; *Other*: Other answer.

Q5.*Did the errors in the captions have an impact on your attention? Choose all that apply to you. If you want to add an answer, please select the box Altro and write your answer down.*

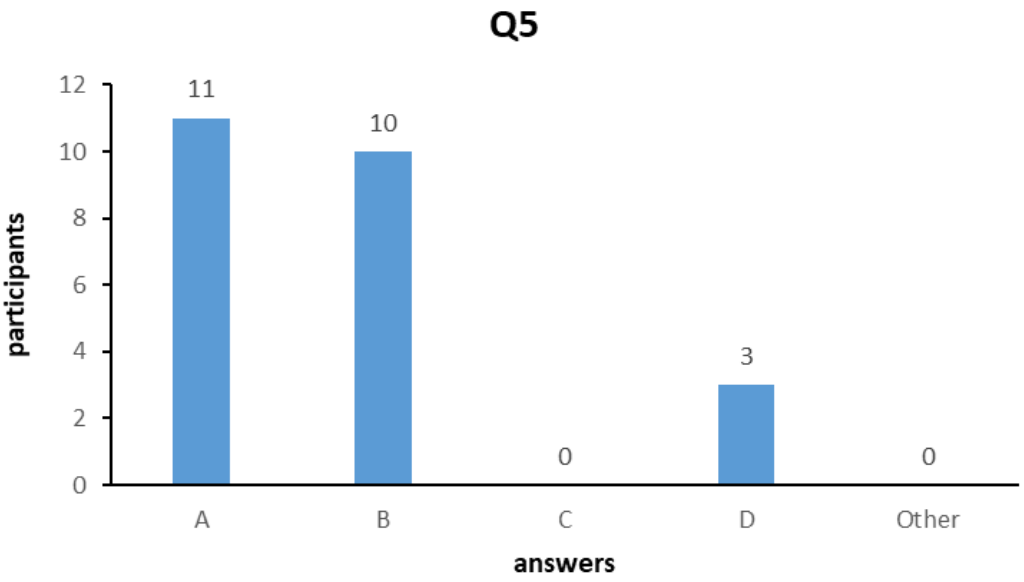
Most participants confirmed that errors had an impact on their attention, making them feel confused (*Option A*: *N* = 11, 68,75%) and distracting them from listening to speakers (*Option B*: *N* = 5, 31,25%) (*Table 77* and *Figure 68* to see the distribution of answers).

Table 74. *Number (count, percentage) of answers for Question n° 5 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
A	11	68,75%
B	10	62,5%
C	-	-
D	3	18,75%
Other	-	-

Note. Labels in the *Answer* column: *Option A*: Yes, errors confused me; *Option B*: Yes, errors distracted me from listening to the speakers; *Option C*: No, I didn't pay attention to the errors; *Option D*: No, I focused on listening rather than reading the captions; *Other*: Other answer.

Figure 63. *Distribution of answers for Question n° 5 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in each column: *Option A*: Yes, errors confused me; *Option B*: Yes, errors distracted me from listening to the speakers; *Option C*: No, I didn't pay attention to the errors; *Option D*: No, I focused on listening rather than reading the captions; *Other*: Other answer.

Q6. *Did the errors in the captions have an impact on the comprehension of the content of the video? Choose all that apply to you. If you want to add an answer, please select the box Altro and write your answer down.*

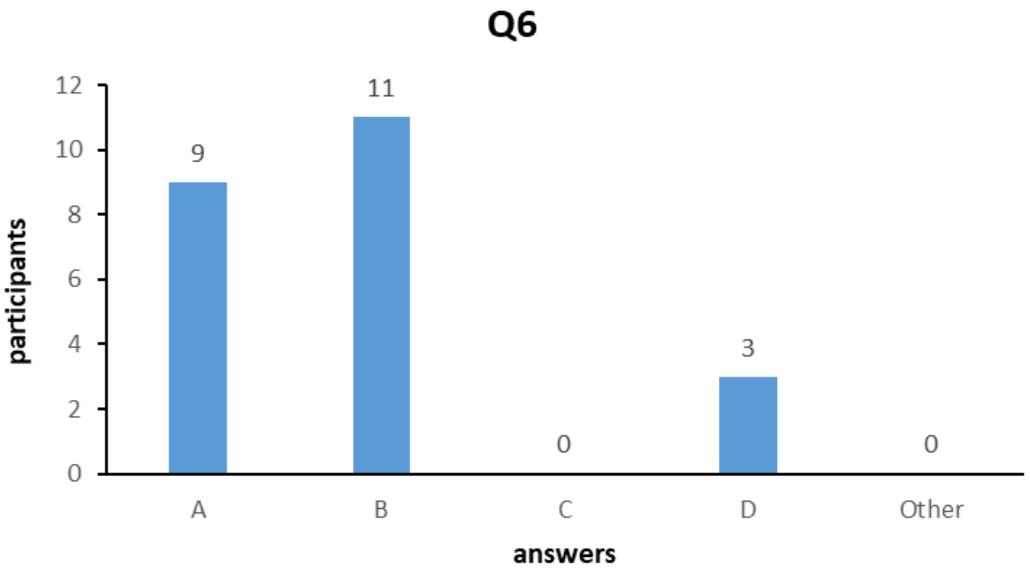
Similarly to the previous question, most participants confirmed that errors hindered the comprehension of the content of the video by distracting them from listening to speakers (*Option B*: *N* = 11, 68,75%) and making them feel confused (*Option A*: *N* = 9, 56,25%) (*Table 78* and *Figure 69* to see the distribution of answers).

Table 75. *Number (count, percentage) of answers for Question n° 6 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
A	9	56,25%
B	11	68,75%
C	-	-
D	3	18,75%
Other	-	-

Note. Labels in the *Answer* column: *Option A*: Yes, errors confused me; *Option B*: Yes, errors distracted me from listening to the speakers; *Option C*: No, I didn't pay attention to the errors; *Option D*: No, I focused on listening rather than reading the captions; *Other*: Other answer.

Figure 64. *Distribution of answers for Question n° 6 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in each column: *Option A*: Yes, errors confused me; *Option B*: Yes, errors distracted me from listening to the speakers; *Option C*: No, I didn't pay attention to the errors; *Option D*: No, I focused on listening rather than reading the captions; *Other*: Other answer.

Q7. *Think about the display format of the captions (...) and say if you agree or not with the following statements[63]: 1) Captions were easy to read; 2) Captions helped me understand the content of the video; 3) It was easy for me to tell how accurate the captions were by looking at the display format; 4) I like the display format of the captions shown in the video[64].*

Table 79 summarizes the answers for statement n° 1 for each markup. On the one hand, interviewed participants agreed (*N* = 1, 25%) or strongly agreed (*N* = 3, 75%) that classic captions were easy to read; a similar trend was found for the markup V2 (*Agree*: *N* = 2, 50%). On the other hand, original and V3 markups were deemed not easy to read (*Disagree*: *N* = 2, 50%). *Figure 70* shows the distributions of answers for statement n° 1 by condition (different color-coded markups).

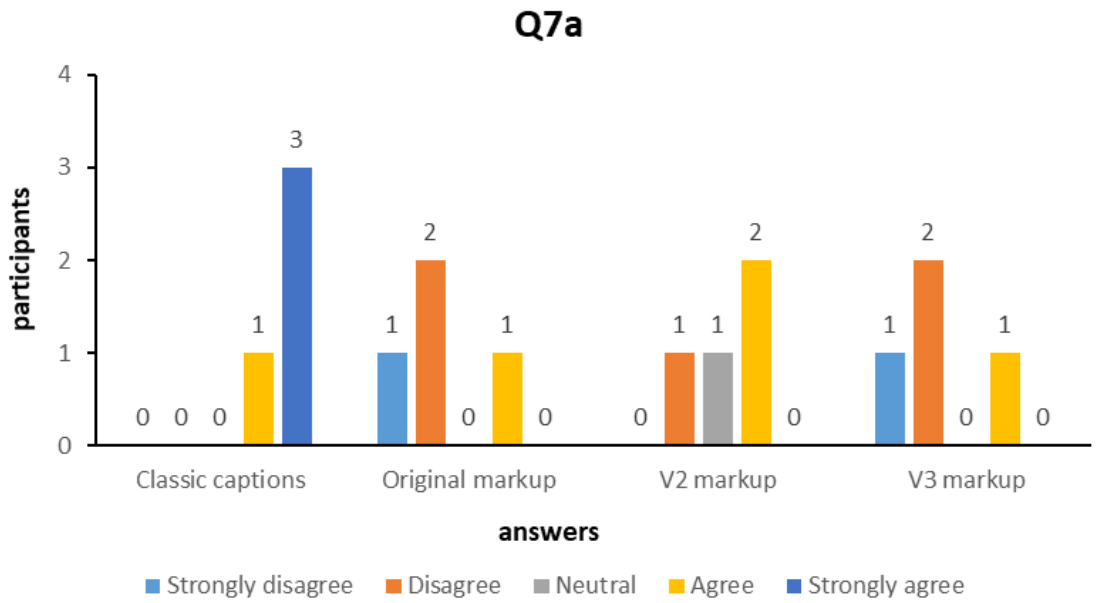
Table 76. *Number (count, percentage) of answers for Question n° 7, statement n° 1 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Condition	Captions were easy to read.				
	SD	**D**	**N**	**A**	**SA**
Classic captions	0	0	0	1	3
Original (OG) markup	1	2	0	1	0
V2 markup	0	1	1	2	0
V3 markup	1	2	0	1	0

Note. Labels in the columns: *SD*: Strongly disagree; *D*: Disagree; *N*: Neutral; *A*: Agree; *SA*: Strongly Agree.

[63] Participants needed to select one option among the five available: strongly disagree, disagree, neutral, agree, strongly agree. Statements were formulated following Shiver & Wolfe (2015).
[64] A screenshot of the video and the captions in the markup assigned to each condition was included in this question.

Figure 65. *Distribution of answers for Question n° 7, statement n° 1 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



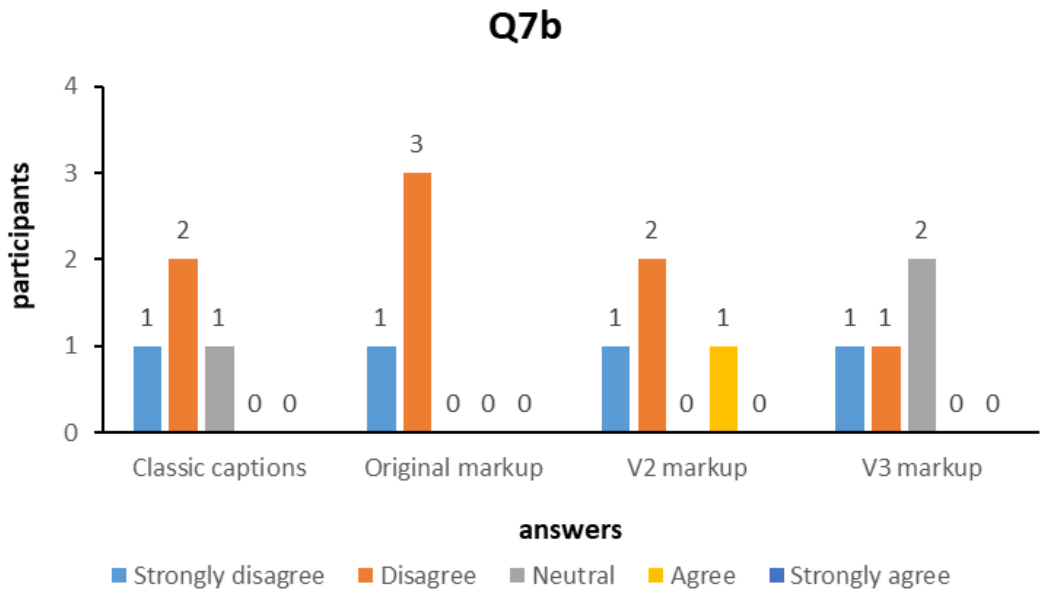
Irrespective of the groups and display formats they were assigned to, the majority of participants overall disagreed that automatic captions helped them understand the content of the video (statement n° 2, see *Table 80* and *Figure 71* for the distributions of answers). Only one participant in the V2 markup group stated that they found automatic captions helpful to understand the content of the video.

Table 77. *Number (count, percentage) of answers for Question n° 7, statement n° 2 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Condition	Captions helped me understand the content of the video.				
	SD	**D**	**N**	**A**	**SA**
Classic captions	1	2	1	0	0
Original (OG) markup	1	3	0	0	0
V2 markup	1	2	0	1	0
V3 markup	1	1	2	0	0

Note. Labels in the columns: *SD*: Strongly disagree; *D*: Disagree; *N*: Neutral; *A*: Agree; *SA*: Strongly Agree.

Figure 66. *Distribution of answers for Question n° 7, statement n° 2 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



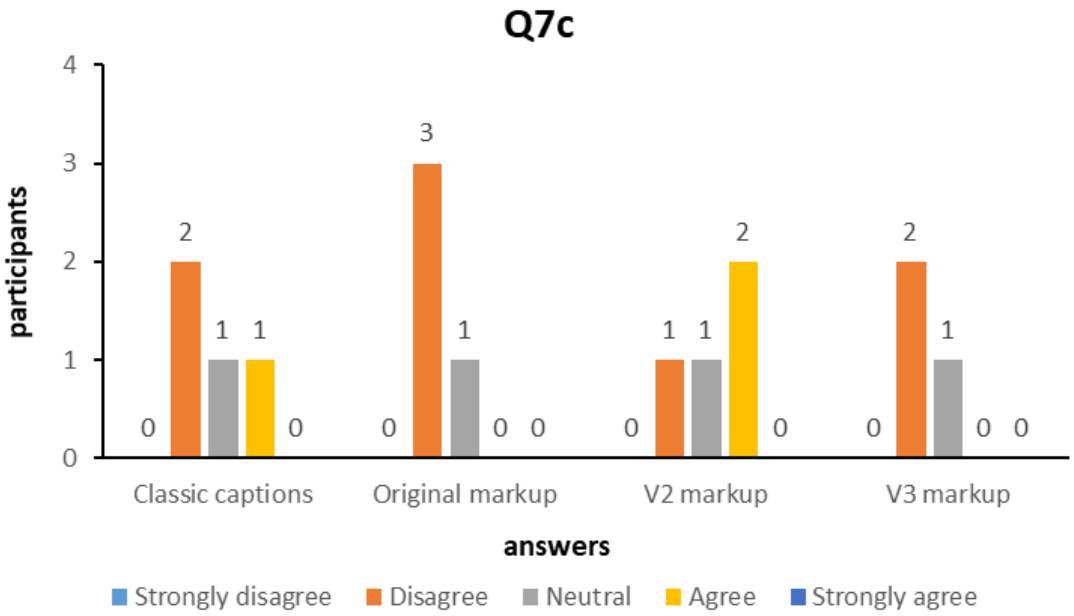
A similar distribution of answers was found for statement n° 3 (*Table 81* and *Figure 72*). Overall, participants did not find it easy to tell how accurate the captions were by looking at the markup. This is relevant to all display formats, except for the V2 markup, where two participants agreed with the statement.

Table 78. *Number (count, percentage) of answers for Question n° 7, statement n° 3 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Condition	It was easy for me to tell how accurate the captions were by looking at the display format.				
	SD	**D**	**N**	**A**	**SA**
Classic captions	0	2	1	1	0
Original (OG) markup	0	3	1	0	0
V2 markup	0	1	1	2	0
V3 markup	0	2	1	0	0

Note. Labels in the columns: *SD*: Strongly disagree; *D*: Disagree; *N*: Neutral; *A*: Agree; *SA*: Strongly Agree.

Figure 67. *Distribution of answers for Question n° 7, statement n° 3 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



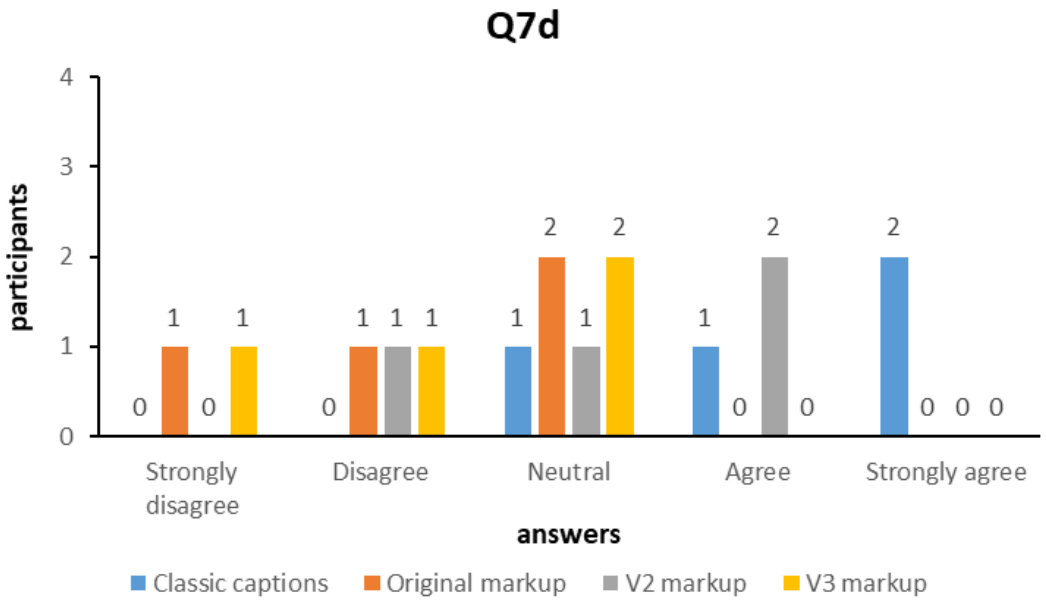
Finally, participants preferred the display format displayed in the classic and V2 markups, where the majority of students agreed or strongly agreed with the statement (*Table 82* and *Figure 73*).

Table 79. *Number (count, percentage) of answers for Question n° 7, statement n° 4 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Condition	I like the display format of the captions shown in the video.				
	SD	**D**	**N**	**A**	**SA**
Classic captions	0	0	1	1	2
Original (OG) markup	1	1	2	0	0
V2 markup	0	1	1	2	0
V3 markup	1	1	2	0	0

Note. Labels in the columns: *SD*: Strongly disagree; *D*: Disagree; *N*: Neutral; *A*: Agree; *SA*: Strongly Agree.

Figure 68. *Distribution of answers for Question n° 7, statement n° 4 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Q8. *Overall, did you find this type of display format useful to support comprehension of the content of the video? Why? Please, motivate your previous answer. You can select more than one option.*

Overall, participants did not find the display formats helpful to support comprehension of content (*Table 83*). This was especially true for the OG and V3 markups; however, a small percentage of participants did find the V2 and classic display formats helpful. *Figure 74* reports on the distribution of answers by condition.

Table 80. *Number (count, percentage) of answers for Question n° 8a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	**Condition**			
	Classic captions	Original (OG) markup	V2 markup	V3 markup
Yes	2	0	1	0
No	2	4	2	4
Other	0	0	1	0

Note. *Other* answer: "More yes than no."

Figure 69. *Distribution of answers for Question n° 8 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

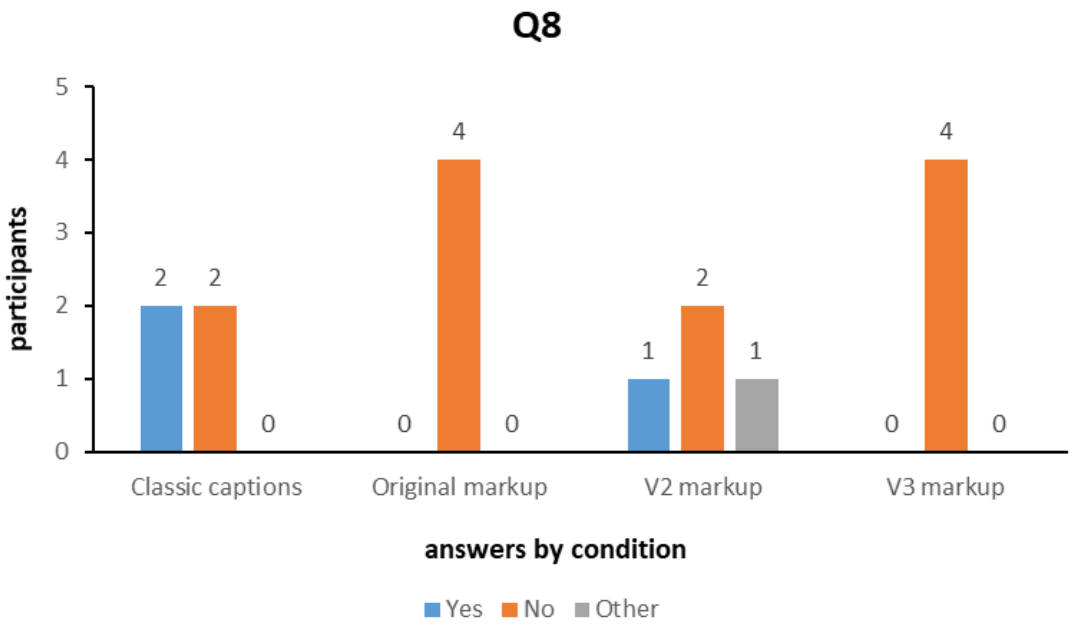


Table 84 sums up the answers given by participants to *Question n° 8b*. Participants in the *classic captions* group found their markup confusing (*Option C*: $N = 2$, 50%), and the presence of errors[65] and the additional text unnecessarily increased the load on their attention (*Option E*: $N = 1$, 25%), distracting them (*Option A*: $N = 1$, 25%). Similarly, participants who were shown the V3 markup found the display format distracting (*Option A*: $N = 4$, 100%), along with the presence of errors in the text[66]. Lastly, some participants in the OG and V2 conditions found the use of different colors to display the system's confidence helpful (*Option B*: $N = 2$, 50%; $N = 1$, 25% respectively), even if the display formats and the errors[67] sometimes distracted (*Option A*: $N = 1$, 25% for both groups) and confused them (*Option C*: $N = 2$, 50%; $N = 1$, 25% respectively).

[65] *Other* answer: "I liked the format of the text, but the content was distracting and full of mistakes."
[66] *Other* answer: "I think the captions were completely useless and distracting, they contained too many errors, the format was too confusing. It is easier to listen without reading them than to try to understand what is written. At a certain point I stopped reading them."
[67] *Other* answer: "It helped me but sometimes confused me. Some errors were too serious (Pidgin-pigeon / Creole / …))."

Table 81. *Number (count, percentage) of answers for Question n° 8b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Condition			
	Classic captions	Original (OG) markup	V2 markup	V3 markup
A	1	1	1	4
B	0	2	1	0
C	2	2	1	1
D	0	0	1	0
E	1	1	1	1
Other	1	0	1	1

Note. Labels in the *Answer* column: *Option A*: Seeing the different display format (i.e., words underlined, etc.) within the captions distracted me; *Option B*: It helped me to know how confident the system is with the transcription; *Option C*: I didn't find it useful, it made me feel confused; *Option D*: It made me feel more confident with my listening skills; *Option E*: I couldn't focus on the text nor speech since there was too much information in the video (text, colors, audio, etc.) - I felt overwhelmed.

Q9. *[Classic captions only] If captions included in the video you watched had had a markup that conveyed such information[68], would you have found them useful? Why? Could you motivate your previous answer? Choose all that apply to you.*

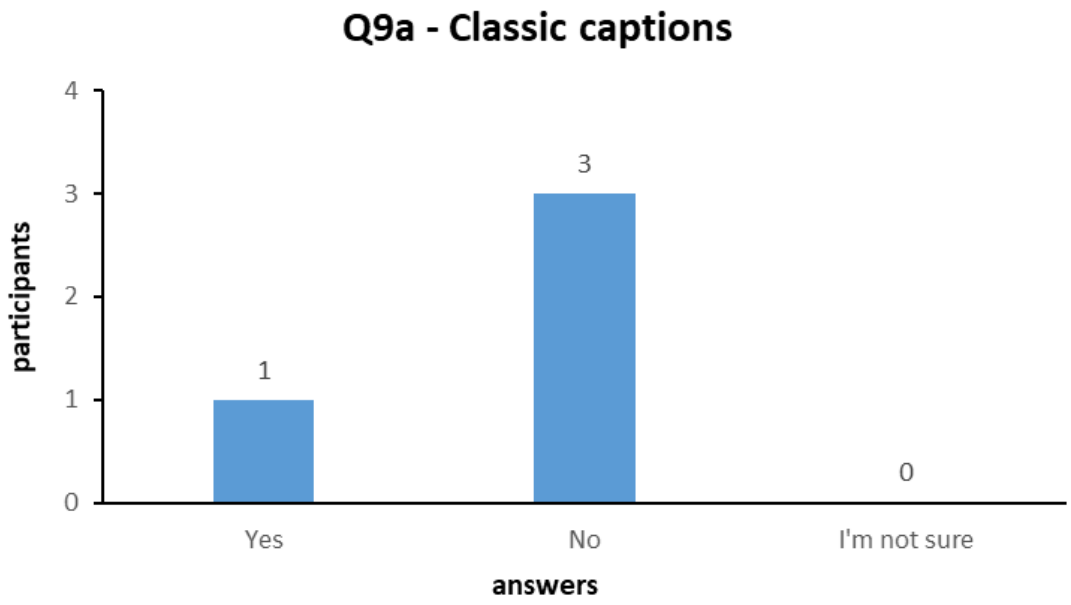
The majority of participants in the *classic captions* group stated that they would have not find it useful to see the confidence of the system through the experimental markups (No: $N = 3$, 75%) (*Table 85* and *Figure 75*).

[68] I.e., the confidence the ASR system has in its transcription ("…information regarding how confident the system is that the transcribed words match the speakers' words").

Table 82. *Number (count, percentage) of answers for Question n° 9a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
Yes	1	25%
No	3	75%
I'm not sure	-	-

Figure 70. *Distribution of answers for Question n° 8 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



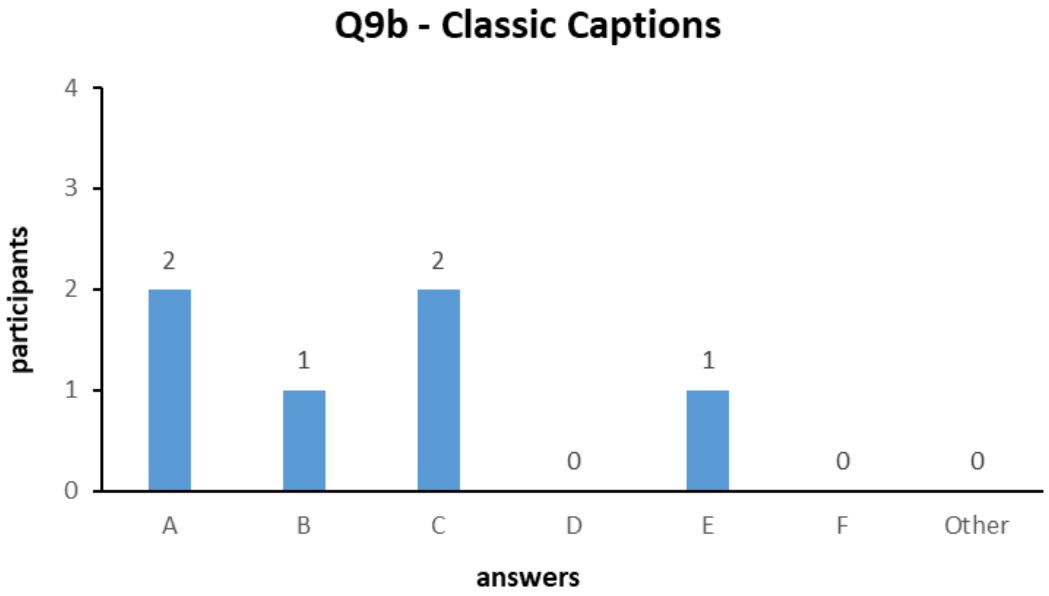
The majority of participants would have found the experimental markups distracting (*Option A*: *N* = 2, 50%) and confusing (*Option C*: *N* = 2, 50%) (*Table 86* and *Figure 76*).

Table 83. *Number (count, percentage) of answers for Question n° 9b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
A	2	50%
B	1	25%
C	2	50%
D	-	-
E	1	25%
F	-	-
Other	-	-

Note. Labels in the *Answer* column: *Option A*: I would have found the different display format (i.e., change of colors) within the captions distracting; *Option B*: It would have helped me to know how confident the system was with the transcription; *Option C*: I wouldn't have found it useful, it would have made me feel confused; *Option D*: I would like to try it in class before making my decision; *Option E*: I wouldn't be able to concentrate on the text nor speech since there would be too much information in the video (text, colors, audio, etc.); *Option F*: It would have made me feel more confident with my listening skills.

Figure 71. *Number (count, percentage) of answers for Question n° 8b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in each column: *Option A*: I would have found the different display format (i.e., change of colors) within the captions distracting; *Option B*: It would have helped me to know how

confident the system was with the transcription; *Option C*: I wouldn't have found it useful, it would have made me feel confused; *Option D*: I would like to try it in class before making my decision; *Option E*: I wouldn't be able to concentrate on the text nor speech since there would be too much information in the video (text, colors, audio, etc.); *Option F*: It would have made me feel more confident with my listening skills.

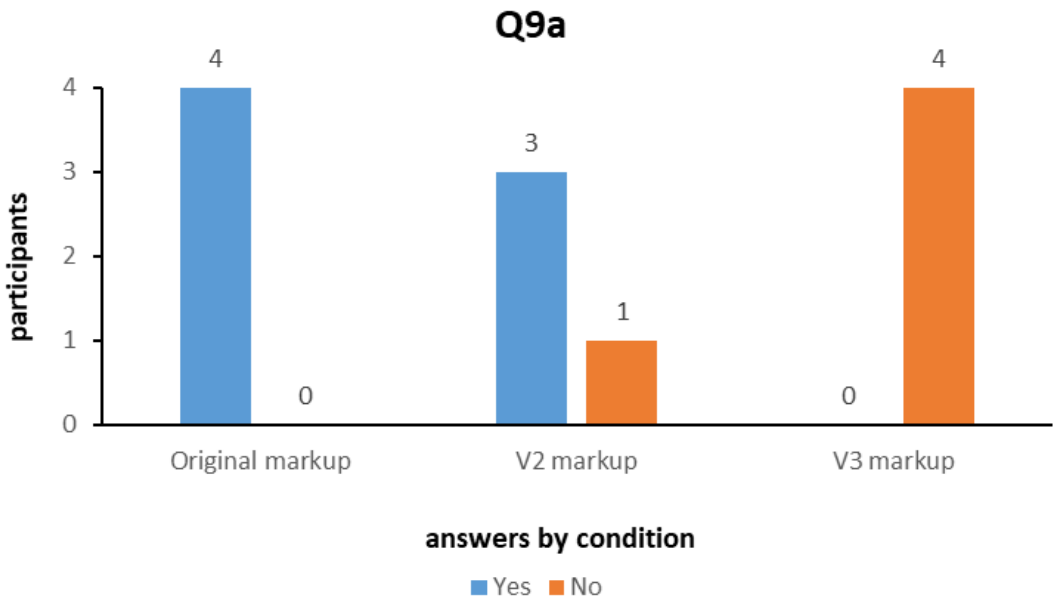
Q9. *Think about the **markup** implemented in the captions. Did you find it useful that it signaled how confident the system was with its transcription? Why? Could you please briefly motivate your previous answer? Choose all that apply to you.*

A large number of participants who watched the video with the OG and V2 markups found the display format useful (Yes: $N = 4$, 100%; $N = 3$, 75% respectively), as opposed to participants in the V3 markup group (No: $N = 4$, 100%) (*Table 87* and *Figure 77*).

Table 84. *Number (count, percentage) of answers for Question n° 9a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Original (OG) markup	V2 markup	V3 markup	Total
Yes	4 (100%)	3 (75%)	-	58,4%
No	-	1 (25%)	4 (100%)	41,6%

Figure 72. *Distribution of answers for Question n° 9a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



In respect to participants in the OG and V2 markup groups, they appreciated that the display formats helped them determine if they could trust the text in the captions or not (*Option B*: $N = 4$, 100%; $N = 3$, 75% respectively). Concerning participants in the V3 markup group, overall, they found the display format unhelpful (*Option A*: $N = 1$, 25%), confusing and distracting. A participant stated that

"I intuitively interpret underlined words as relevant, rather than potential errors."

Two other participants stated that overall, captions were useless due to the combination of too many errors in the transcription and the display format[69,70] (*Table 88* and *Figure 78*).

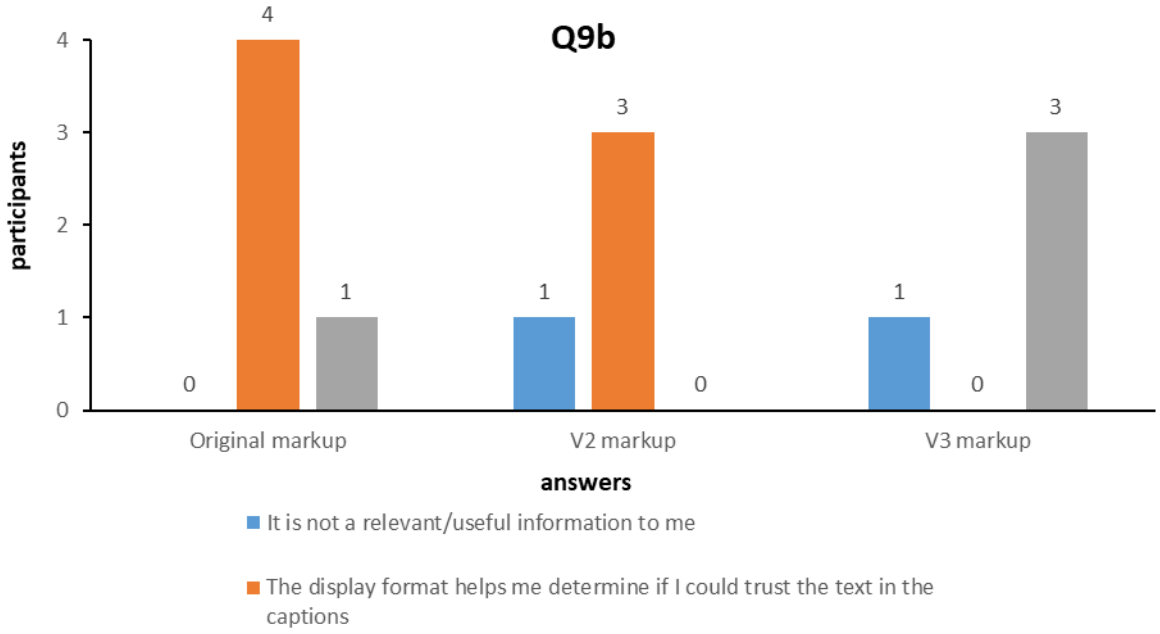
[69] *Other* answer, *V3 markup* group: "I think the captions were completely useless and distracting, they contained too many errors, the format was too confusing. It is easier to listen without reading them than to try to understand what is written. At a certain point I stopped reading them."
[70] *Other* answer, *V3 markup* group: "It doesn't work well so it's just a bother."

Table 85. *Number (count, percentage) of answers for Question n° 9b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Original (OG) markup	V2 markup	V3 markup
A	-	1 (25%)	1 (25%)
B	4 (100%)	3 (75%)	-
Other	1 (25%)[71]	-	3 (75%)

Note. Labels in the *Answer* column: *Option A*: It is not relevant/useful information to me; *Option B*: The display format helps me determine if I could trust the text in the captions.

Figure 73. *Distribution of answers for Question n° 9b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



[71] *Other* answer, *OG markup* group: "I think the concept is useful, but there were too many colors, underlining etc. and it also felt distracting."

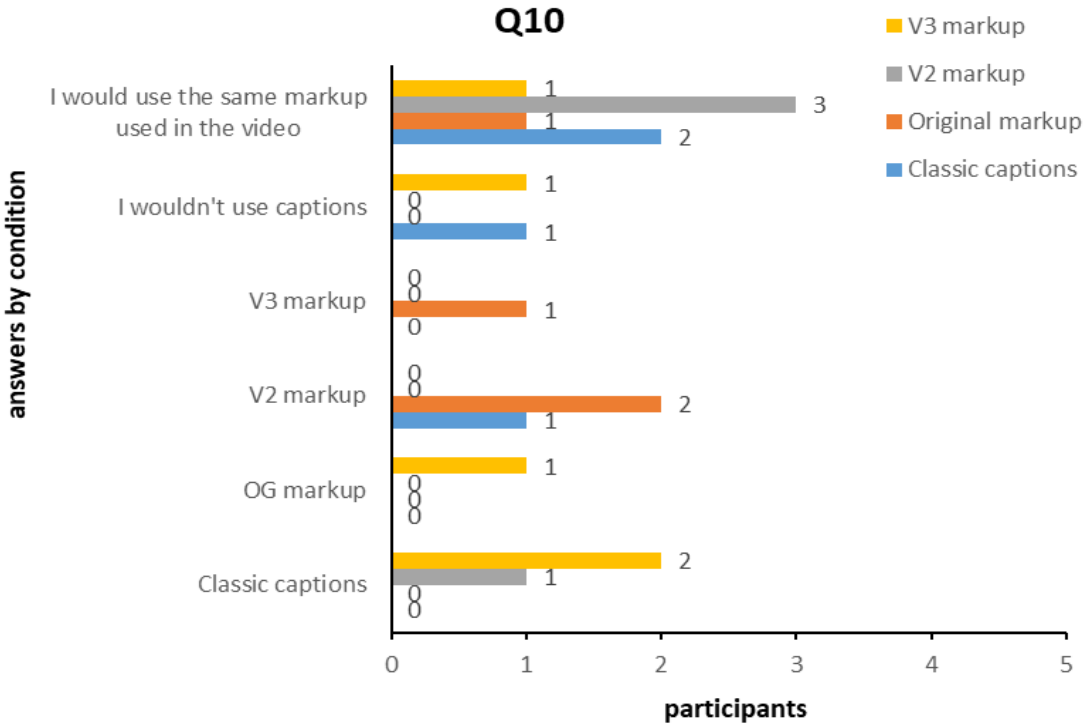
Q10. *If you were given the possibility to watch the video again and choose how captions would be displayed on screen, which format would you choose?*

On the one hand, overall, participants would prefer using the V2 markup, especially those who watched the video in the same condition, or in the OG display format (*OG markup*: *N* = 2, 50%). On the other hand, the classic markup was preferred by participants in the same condition or in the V3 markup group (*Classic captions*: *N* = 2, 50%; *V3 markup*: *N* = 2, 50% respectively) (*Table 89* and *Figure 79*).

Table 86. *Number (count, percentage) of answers for Question n° 10 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Classic captions	Original (OG) markup	V2 markup	V3 markup
Classic captions	-	-	1	2
Original (OG) markup	-	-	-	-
V2 markup	1	2	-	-
V3 markup	-	1	-	-
I wouldn't use captions	1	-	-	1
Same markup of the video	2	1	3	1

Figure 74. *Distribution of answers for Question n° 10 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Q11. *Do you think the markup implemented in the captions had an impact on your attention? Choose all that apply to you. If you want to add something, please select the option Altro and write your answer there.*

The majority of participants found the markups confusing (*Option A: N* = 7, 43,75%) and distracting (*Option B: N* = 5, 31,25%; *Other answer*[72]) (*Table 90* and *Figure 80*). Only three participants respectively did not pay attention to the markup (*Option C: N* = 1, 6,25%) and focused on listening, ignoring the captions (*Option D: N* = 1, 6,25%; *Other answer*[73]).

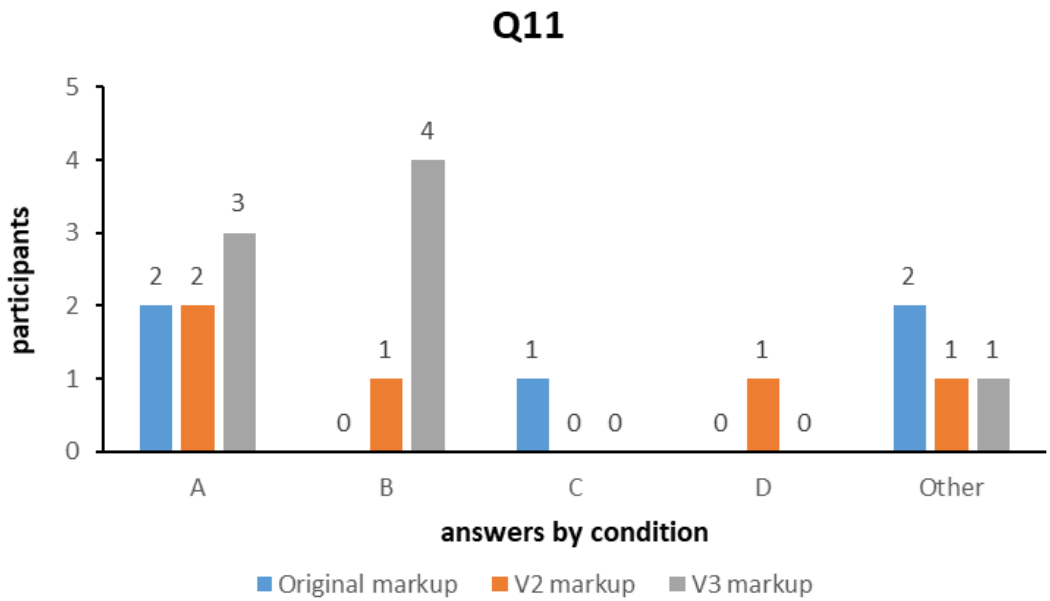
[72] *Other* answer, *V2 markup*: "Yes, the format distracted me from listening to the speakers (sometimes)."
[73] *Other* answer, *OG markup*: "I found it hard to focus on the captions, since the audio was easier to follow."

Table 87. *Number (count, percentage) of answers for Question n° 11 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Original (OG) markup	V2 markup	V3 markup	Total
A	2	2	3	7 (43,75%)
B	-	1	4	5 (31,25%)
C	1	-	-	1 (6,25%)
D	-	1	-	1 (6,25%)
Other	2[74]	1[75]	1[76]	4 (25%)

Note. Labels in the *Answer* column: *Option A*: Yes, the markup confused me; *Option B*: Yes, the markup distracted me from listening to the speakers; *Option C*: No, I didn't pay attention to the markup; *Option D*: No, I focused on listening rather than reading the captions.

Figure 75. *Distribution of answers for Question n° 11 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in each column: *Option A*: Yes, the markup confused me; *Option B*: Yes, the markup distracted me from listening to the speakers; *Option C*: No, I didn't pay attention to the markup; *Option D*: No, I focused on listening rather than reading the captions.

[74] *Other* answer, *OG markup*: "I liked it, but it wasn't a very good caption."
[75] *Other* answer, *V2 markup*: "Yes, the format distracted me from listening to the speakers (sometimes)."
[76] *Other* answer, *V3 markup*: "I think the captions were completely useless and distracting, they contained too many errors, the format was too confusing. It is easier to listen without reading them than to try to understand what is written. At a certain point I stopped reading them."

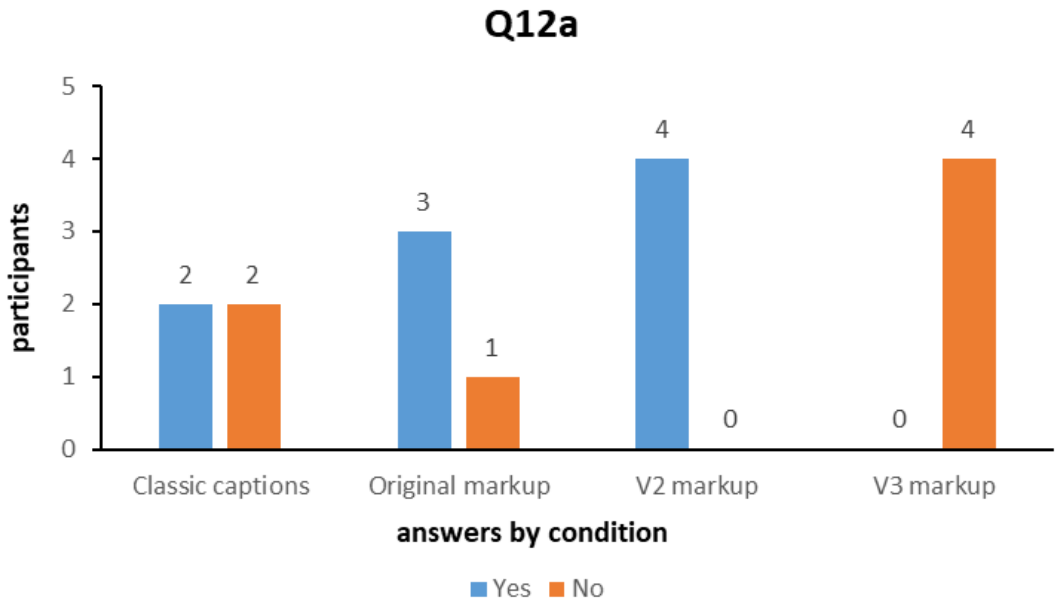
Q12. *In general, did you like the markup implemented in the captions? Why? Could you motivate your previous answer? Choose all that apply to you.*

Overall, participants liked the OG ($N = 3$, 75%) and the V2 ($N = 4$, 100%) markups the most (*Table 91* and *Figure 81*). All participants assigned the V3 markup did not like the display format, while the preferences were split for the *classic captions* (Yes: $N = 2$, 50%; No: $N = 2$, 50%).

Table 88. *Number (count, percentage) of answers for Question n° 12a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Classic captions	Original (OG) markup	V2 markup	V3 markup
Yes	2	3	4	-
No	2	1	-	4

Figure 76. *Distribution of answers for Question n° 12a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Participants from the *OG* and *V2 markup* groups found the display formats helpful to understand the content of the video (*OG markup*: *N* = 2, 50%; *V2 markup*: *N* = 2, 50% respectively) and informative (*V2 markup*: *N* = 2, 50%), also because "it shows possible mistakes" (*OG markup, other* answer) (*Table 92* and *Figure 82*). One participant from the *OG markup* group also stated that

"I liked the concept, but the experience was a bit overwhelming. Also, the captions had so many errors that it was easier to just listen. If the captions had been more accurate, this system might have been useful to know when to listen more closely."

All participants in the *V3 markup* group found the display format distracting, and one participant added that they would have preferred a slightly different format for words with low confidence to be displayed:

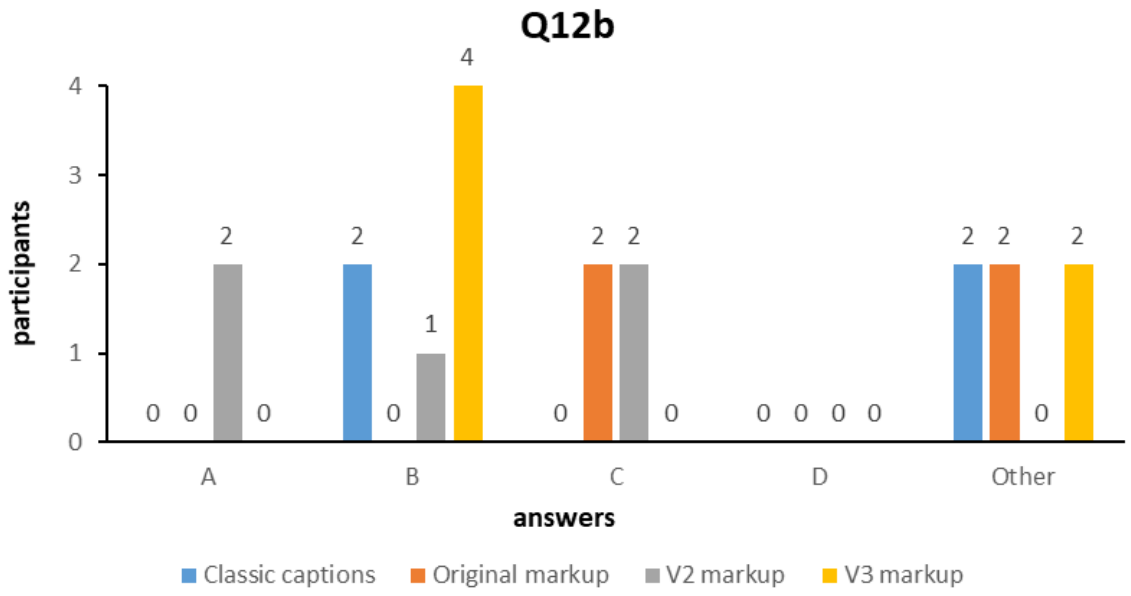
"If the markup is simple (underlined but that's it) then fine, otherwise no."

Table 89. *Number (count, percentage) of answers for Question n° 12b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Classic captions	Original (OG) markup	V2 markup	V3 markup
A	-	-	2	-
B	2	-	1	4
C	-	2	2	-
D	-	-	-	-
Other	2	2	-	2

Note. Labels in the *Answer* column: *Option A*: I found it really informative; *Option B*: It distracted me; *Option C*: It helped me understand the content of the video; *Option D*: It highlighted some words I did not know.

Figure 77. *Distribution of answers for Question n° 12b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in each column: *Option A*: I found it really informative; *Option B*: It distracted me; *Option C*: It helped me understand the content of the video; *Option D*: It highlighted some words I did not know.

Q13. *Would you like to add some thoughts or opinions regarding the markup used in the captions? If not, click on Avanti [optional, open-ended question].*

Four participants, respectively from the *classic captions*, *V2*, and *V3 markup* groups wrote additional thoughts and opinions on the display formats of the automatic captions.

One participant from the *classic captions* group stated that

"The format was perfect, but the content was too distracting, so I ended up ignoring it and going back to it just to spot mistakes. It took me a full minute to understand we were talking about languages and not animals at first."

Another participant from the *V3 markup* group stated that

"The idea is good, but in fact it only distracted me, it wasn't helpful."

Similarly, two participants from the *V2* and *V3 markup* groups remarked that automatic captions were not helpful due to the errors in the transcription. On the one hand, the participant from the *V2 markup* group stated that

"I found ASR very useful but maybe it still needs improvement, e.g. [to] be trained to recognize technical terms that are crucial to understand academic lectures."

On the other hand, the participant from the *V2 markup* group stated the following:

"I think the captions were completely useless and distracting, they contained too many errors, the format was too confusing. It is easier to listen without reading them than to try to understand what is written. At a certain point I stopped reading them."

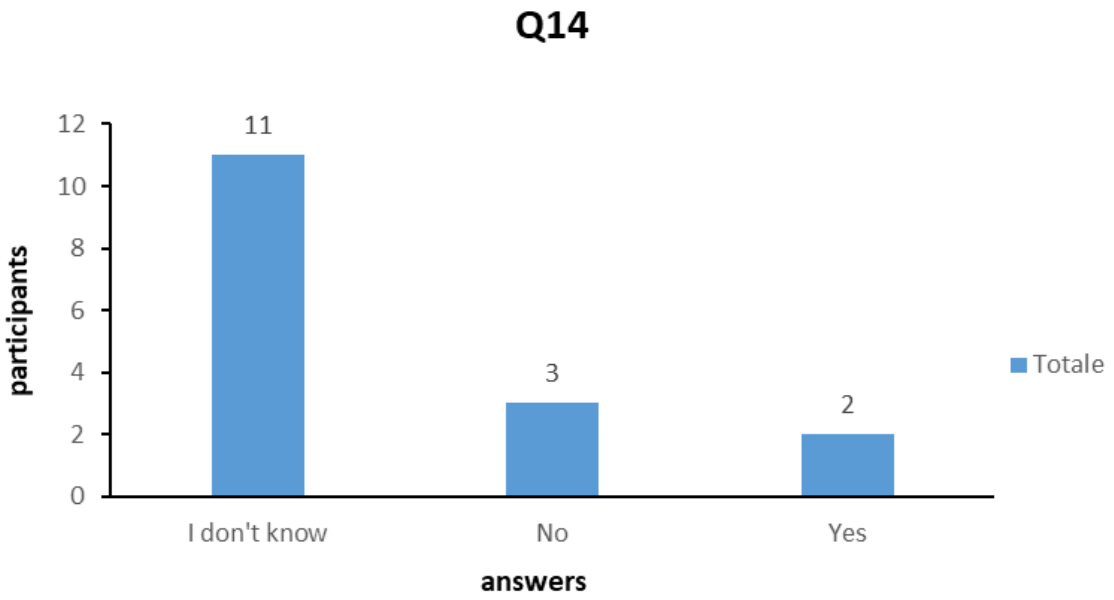
Q14. *Would you consider using live captions (generated by an ASR system) during courses taught in English if the University provided this service? Why? Could you motivate your previous answer? Choose all that apply to you.*

The majority of participants were not sure about using automatic live captions during lectures delivered in English at university (Yes: *N* = 11, 68,75%) (*Table 93* and *Figure 83*).

Table 90. *Number (count, percentage) of answers for Question n° 14a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
Yes	2	12,5%
No	3	18,75%
I don't know	11	68,75%

Figure 78. *Distribution of answers for Question n° 14a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



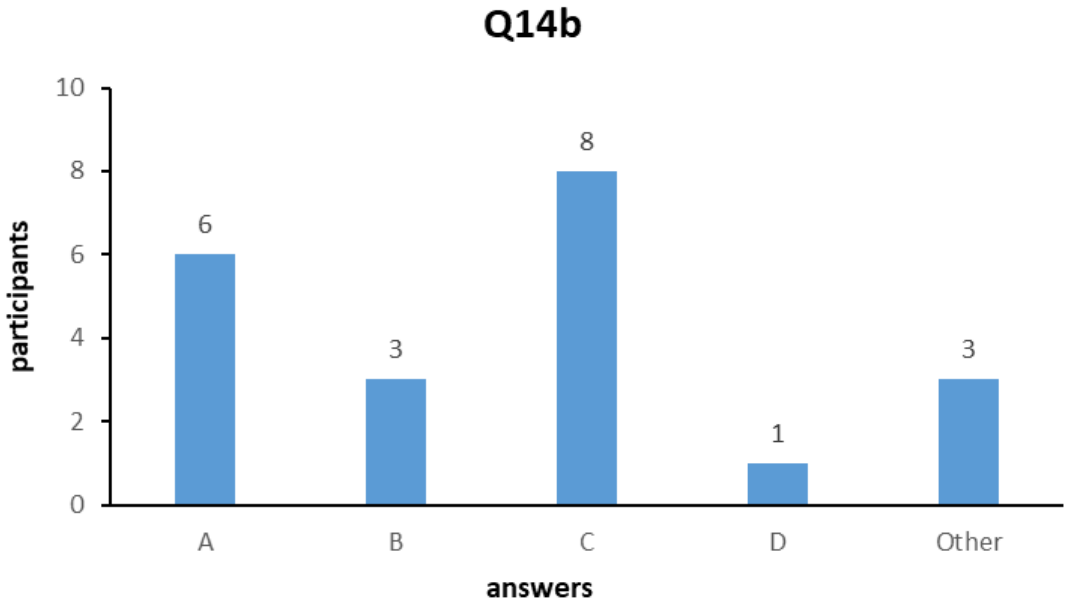
Half of the students stated that they would like to try the system in class before deciding if they would use automatic captions in class (*Option C*: *N* = 8, 50%); six of them also stated that they would find it distracting to read the captions in addition to the text on the slides (*Option A*: *N* = 6, 37,5%) (*Table 94* and *Figure 84*).

Table 91. *Number (count, percentage) of answers for Question n° 14b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
A	6	37,5%
B	3	18,75%
C	8	50%
D	1	6,25%
Other	3	18,75%

Note. Labels in the *Answer* column: *Option A*: I would find it distracting to see the additional text on screen; *Option B*: It would help me to know how confident the system is with the transcription; *Option C*: I would like to try it in class before making my decision; *Option D*: My listening comprehension would benefit by the presence of captions.

Figure 79. *Distribution of answers for Question n° 14b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in each column: *Option A*: I would find it distracting to see the additional text on screen; *Option B*: It would help me to know how confident the system is with the transcription; *Option C*: I would like to try it in class before making my decision; *Option D*: My listening comprehension would benefit by the presence of captions.

Some participants stated their concerns about using ASR-generated captions during lectures. Two participants respectively said that "*If too much trust is put in a ASR system to understand the topic of a discussion, there is a serious risk of being confused by the numerous errors and leave the lectures without understanding anything or being convinced of untrue facts*" and that "*I think it [ASR systems] still need improvement to be used for academic lectures*".

Another participant stated that "*I am usually looking at my notebook when listening to lectures, so I do not think I would benefit from captions*."

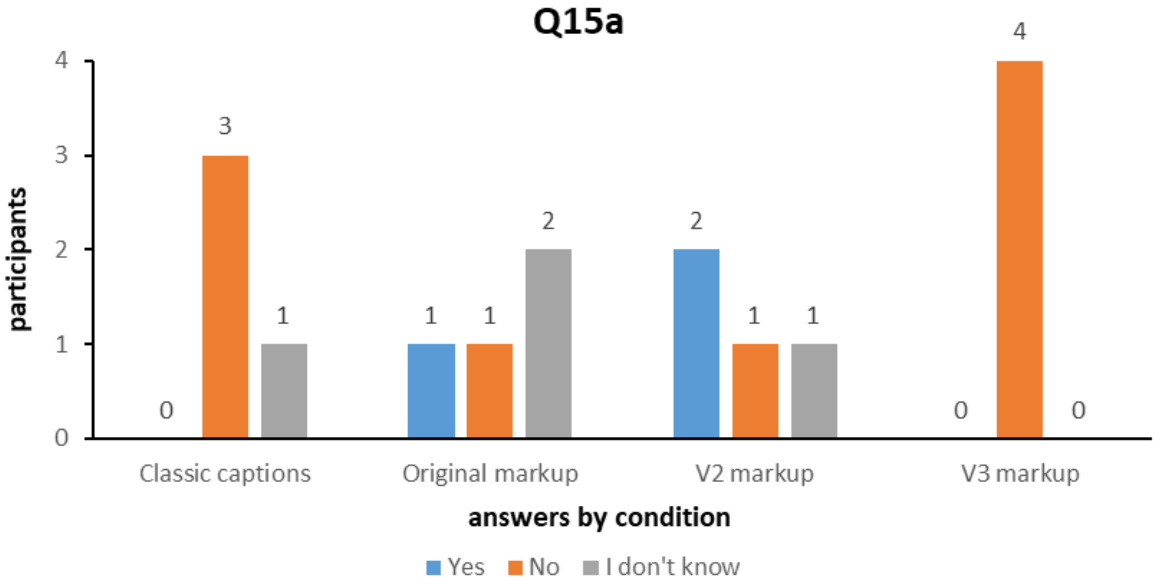
Q15. *Would you consider using live captions (generated by an ASR system) with the display format you saw in the video during courses taught in English if the University provided this service? Why? Could you motivate your previous answer? Choose all that apply to you.*

Overall, participants could be again divided into two groups based on the display formats they were exposed to (*Table 95* and *Figure 85*). On the one hand, participants in the *classic format* and *V3 markup* groups would not consider using live captions in class with those display formats. On the other hand, participants in the *OG* and *V2 markup* groups were more inclined to try these formats in class.

Table 92. *Number (count, percentage) of answers for Question n° 15a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Classic captions	Original (OG) markup	V2 markup	V3 markup
Yes	-	1	2	-
No	3	1	1	4
I don't know	1	2	1	-

Figure 80. *Distribution of answers for Question n° 15a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Participants in the *classic format* and *V3 markup* groups justified their previous answer saying that they would find it distracting to see the different display formats within the captions (*Option A,* C*lassic captions*: N = 2, 50%; *V3 markup*: N = 4, 100%) (*Table 96* and *Figure 86*). Instead, participants in the *OG* and *V2 markup* groups stated that they would like to try these formats in class (*Option C, OG markup*: N = 2, 50%; *V2 markup*: N = 2, 50%), but that it would them knowing how confident the system was with its transcription (*Option B, OG markup*: N = 2, 50%; *V2 markup*: N = 2, 50%).

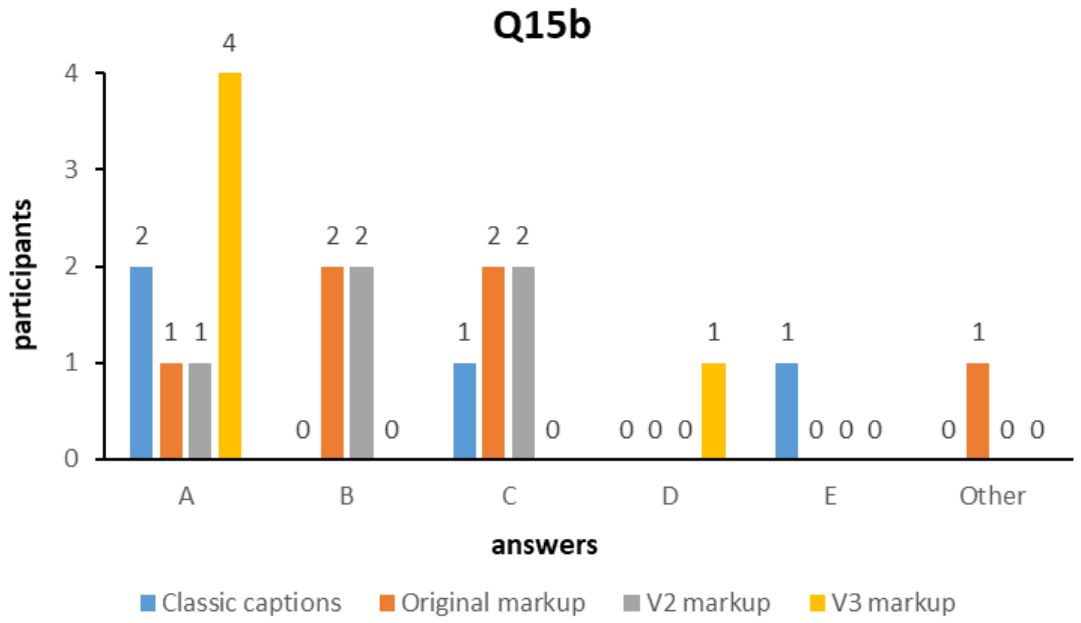
Table 93. *Number (count, percentage) of answers for Question n° 15b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Classic captions	Original (OG) markup	V2 markup	V3 markup
A	2	1	1	4
B	-	2	2	-
C	1	2	2	-
D	-	-	-	1
E	1	-	-	-
Other	-	1[77]	-	-

Note. Labels in the *Answer* column: *Option A*: I would find it distracting to see the different display format (e.g., words underlined, etc.) within the captions; *Option B*: It would help me to know how confident the system is with the transcription; *Option C*: I would like to try it in class before making my decision; *Option D*: I would prefer to see the standard format for captions; *Option E*: [Classic captions] It wouldn't help me to know how confident the system is with the transcription.

[77] *Other* answer, *OG markup* group: "I would prefer to see another format."

Figure 81. *Distribution of answers for Question n° 15b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in the *Answer* column: *Option A*: I would find it distracting to see the different display format (e.g., words underlined, etc.) within the captions; *Option B*: It would help me to know how confident the system is with the transcription; *Option C*: I would like to try it in class before making my decision; *Option D*: I would prefer to see the standard format for captions; *Option E*: [Classic captions] It wouldn't help me to know how confident the system is with the transcription.

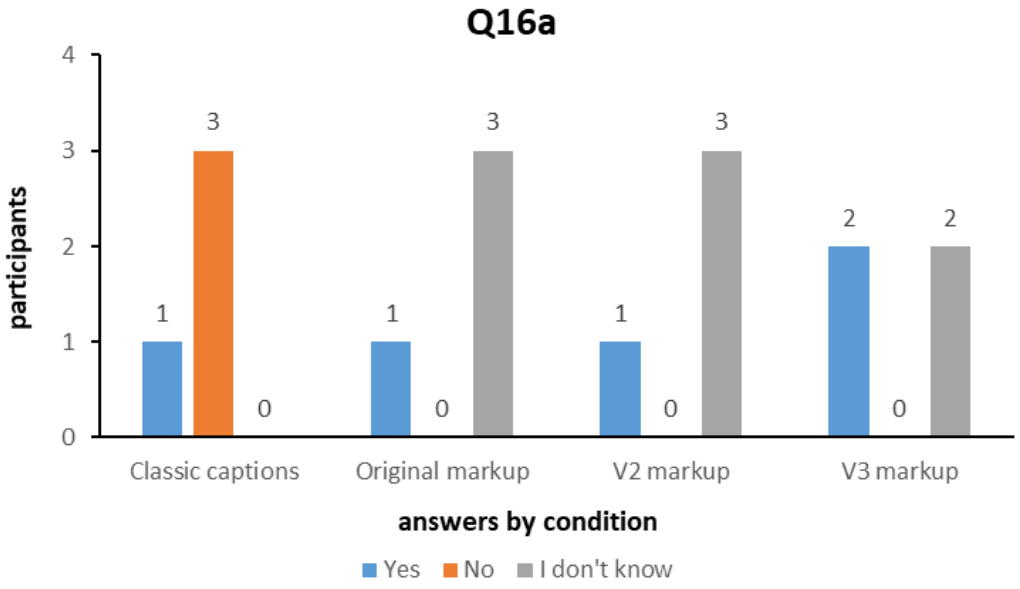
Q16. *Would you consider using live captions (generated by an ASR system) with other types of display format during courses taught in English if the University provided this service? Why? Could you motivate your previous answer? Choose all that apply to you.*

The majority of participants were unsure of whether they would consider other types of display formats to be displayed in the automatic captions (*Table 97* and *Figure 87*).

Table 94. *Number (count, percentage) of answers for Question n° 16a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Classic captions	Original (OG) markup	V2 markup	V3 markup
Yes	1	1	1	2
No	3	-	-	-
I don't know	-	3	3	2

Figure 82. *Distribution of answers for Question n° 16a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



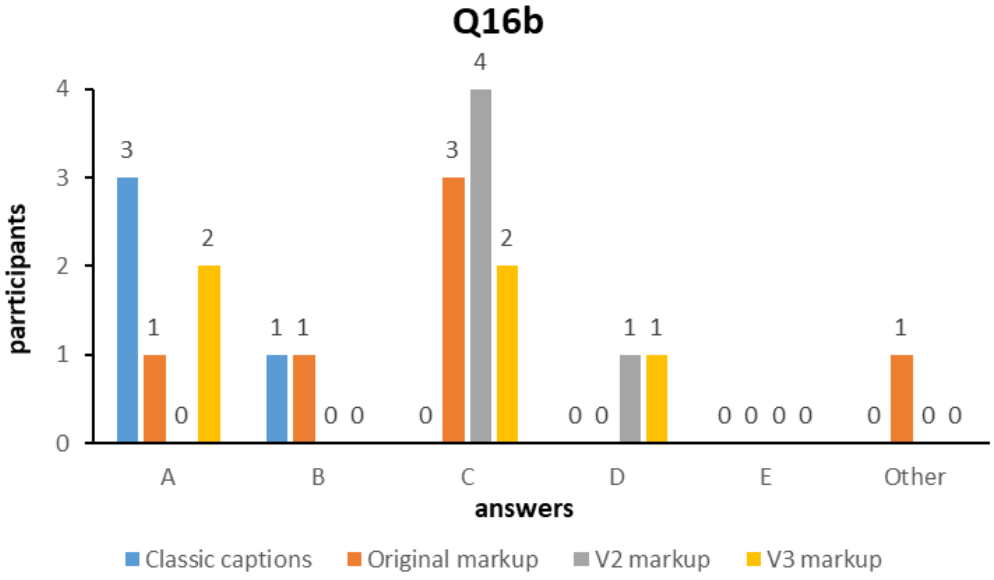
The reason is that they would like to try these different display formats in class before making their decision (*Option C*). The majority of participants in the classic captions justified their answer by stating that seeing the changing graphical features within the text would distract them (*Option A*) (*Table 98* and *Figure 88*).

Table 95. *Number (count, percentage) of answers for Question n° 16b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Classic captions	Original (OG) markup	V2 markup	V3 markup
A	3	1	-	2
B	1	1	-	-
C	-	3	4	2
D	-	-	1	1
E	-	-	-	-
Other	-	1[78]	-	-

Note. Labels in the *Answer* column: *Option A*: I would find it distracting to see the different display format (e.g., words underlined, etc.) within the captions; *Option B*: It would help me to know how confident the system is with the transcription; *Option C*: I would like to try it in class before making my decision; *Option D*: I would prefer to see the standard format for captions; *Option E*: [Classic captions] It wouldn't help me to know how confident the system is with the transcription.

Figure 83. *Distribution of answers for Question n° 16b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in the *Answer* column: *Option A*: I would find it distracting to see the different display format (e.g., words underlined, etc.) within the captions; *Option B*: It would help me to know how

[78] *Other* answer, *OG markup*: "I would prefer to see another format."

confident the system is with the transcription; *Option C*: I would like to try it in class before making my decision; *Option D*: I would prefer to see the standard format for captions; *Option E*: [Classic captions] It wouldn't help me to know how confident the system is with the transcription.

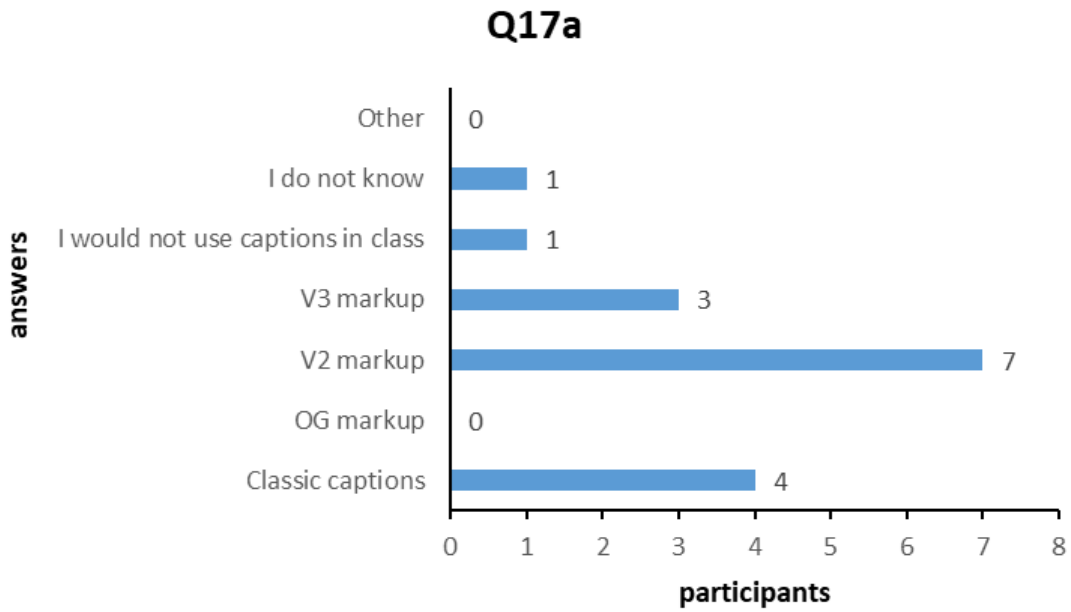
Q17. *If your university offered live captioning during courses taught in English, which captioning style would you consider using? Why? Could you motivate your previous answer? Choose all that apply to you.*

Almost half of the participants would consider using the *V2 markup* to display the confidence level of the ASR system in the automatic captions (*N* = 7, 43,75%) (*Table 99* and *Figure 89*), followed by the classic format (*N* = 4, 25%).

Table 96. *Number (count, percentage) of answers for Question n° 17a in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
Classic captions	4	25%
Original (OG) markup	-	-
V2 markup	7	43,75%
V3 markup	3	18,75%
I wouldn’t use captions in class	1	6,25%
I don’t know	1	6,25%
Other	-	-

Figure 84. *Distribution of answers for Question n° 16b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



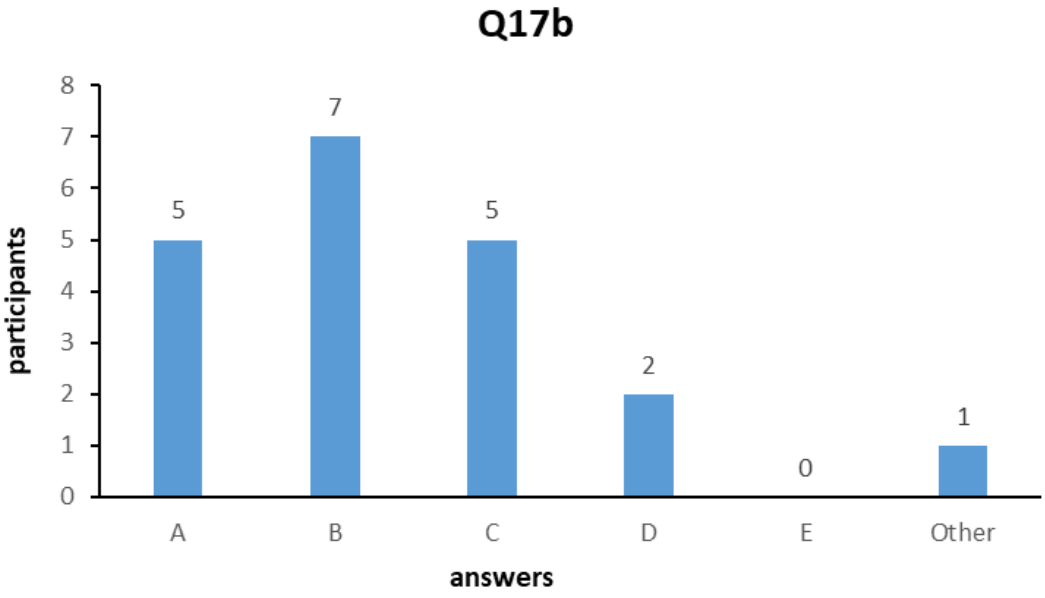
Participants stated that their preferred format would be useful to know the confidence of the system in its transcription (*Option B*: *N* = 7, 43,75%). Some of them expressed their concern about the changes in graphical features in the text (*Option A*: *N* = 5, 31,25%), but they would be open to trying it in class (*Option C*: *N* = 5, 31,25%) (*Table 100* and *Figure 90*).

Table 97. *Number (count, percentage) of answers for Question n° 17b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
A	5	31,25%
B	7	43,75%
C	5	31,25%
D	2	12,5%
E	-	-
Other	1[79]	6,25%

Note. Labels in the *Answer* column: *Option A*: I would find it distracting to see the different display format (e.g., words underlined, etc.) within the captions; *Option B*: It would help me to know how confident the system is with the transcription; *Option C*: I would like to try it in class before making my decision; *Option D*: I would prefer to see the standard format for captions; *Option E*: [Classic captions] It wouldn't help me to know how confident the system is with the transcription.

Figure 85. *Distribution of answers for Question n° 17b in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in each column: *Option A*: I would find it distracting to see the different display format (e.g., words underlined, etc.) within the captions; *Option B*: It would help me to know how confident

[79] *Other* answer, *V3 markup*: "I would know that underlined means it might be wrong."

the system is with the transcription; *Option C*: I would like to try it in class before making my decision; *Option D*: I would prefer to see the standard format for captions; *Option E*: [Classic captions] It wouldn't help me to know how confident the system is with the transcription.

Q18. *Do you think your knowledge of English would get better if live captions were provided in class for courses taught in English? Why? Choose all that apply to you.*

Half participants are mainly concerned about the high error rate in the transcriptions, and for this reason, they do not think that automatic captions would help improving their knowledge of English (*Option E*: N = 8, 50%). Other participants, however, stated that automatic captions would help them recover words they missed (*Option E*: N = 6, 37,5%) or improve their vocabulary knowledge (*Option E*: N = 5, 31,25%) (*Table 101* and *Figure 91*). One participant specified that

"*on one hand captions could help me recovering words, but on the other hand sometimes the ASR didn't recognize wrong words as errors and that could negatively affect my knowledge.*"

Similarly, another participant stated that

"*[using captions] it may be helpful for the vocabulary, but I fear it would be a source of distraction*."

Lastly, two participants stated their concern about using captions to learn English: while one participant declared that

"*I think using subtitles is a bit of an old school way of teaching and learning a language*",

the other stated that

"*in uni lectures I would be focused on content more than my English learning.*"

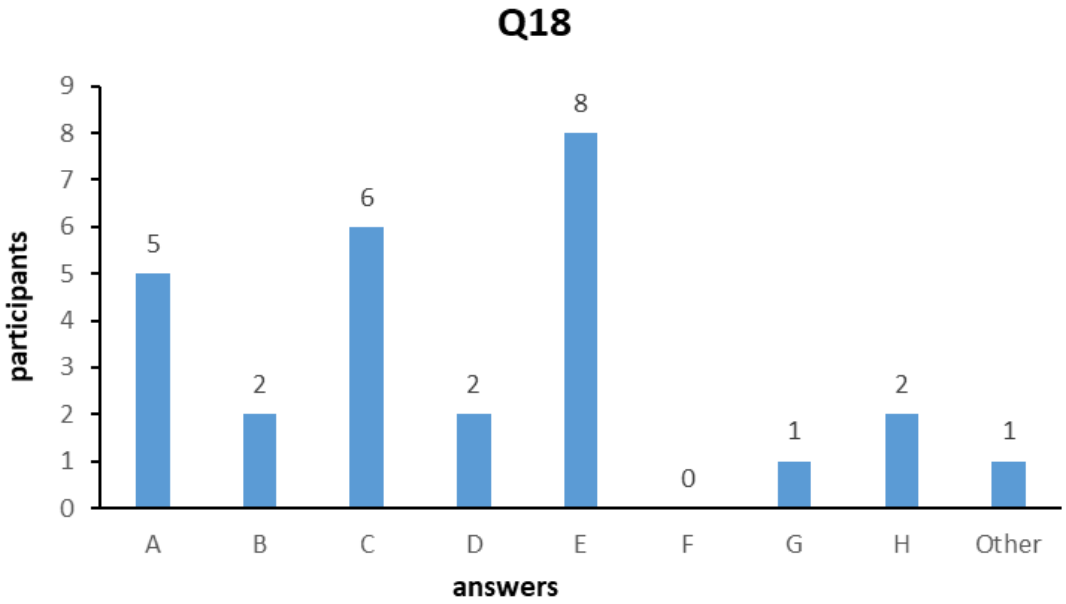
Table 98. *Number (count, percentage) of answers for Question n° 18 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*

Answer	Total (n°)	Total (%)
A	5	31,25%
B	2	12,5%
C	6	37,5%

D	2	12,5%
E	8	50%
Yes: Other reason	-	-
No: Other reason	1	6,25%
Maybe: Other reason	2	12,5%
Other	1	6,25%

Note. Labels in the *Answer* column: *Option A*: Yes: captions could help me improving my vocabulary knowledge; *Option B*: Yes: captions could help me improving my pronunciation; *Option C*: Yes: captions could help me recovering words I missed/I did not understand for some reason; *Option D*: No: captions would only be redundant, I prefer listening to the professor; *Option E*: No: too many errors make live captions useless.

Figure 86. *Distribution of answers for Question n° 18 in the questionnaire investigating the preferences of students for the color-coded markups in automatic captions.*



Note. Labels in the *Answer* column: *Option A*: Yes: captions could help me improving my vocabulary knowledge; *Option B*: Yes: captions could help me improving my pronunciation; *Option C*: Yes: captions could help me recovering words I missed/I did not understand for some reason; *Option D*: No: captions would only be redundant, I prefer listening to the professor; *Option E*: No: too many errors make live captions useless.

5.5 Discussion

In this section, we present our findings and address our research question. RQ2 and RQ3 will be discussed together as they focus on how the different display formats affect speech processing, comprehension, and attention by discussing both behavioral data and insights from L2 speakers.

RQ1. Do errors in automatic captions affect speech processing and content comprehension in L2 speakers of English?

Overall, the analysis of the performance of L2 speakers of English assessed that comprehension of the content was sufficient for most participants, and the average comprehension rates were similar in the four groups (Mean score = 6.4, SD = 0,625). However, results from the questionnaire administered after the comprehension test suggest that participants likely did not rely heavily on automatic captions, as errors in the text may have caused more drawbacks than benefits for L2 speakers.

Results indicated that most L2 speakers had trouble with speech perception and segmentation mainly because they were unfamiliar with some of the speakers' accents. These difficulties are common in L2 speakers of English and are well documented in the literature (e.g., Goh, 2000; Garcia Lecumberri *et al.*, 2010; Grey *et al*., 2018). Moreover, their unfamiliarity with the topic discussed in the clip has also contributed to their difficulties in understanding the content. In this context, the missing knowledge of technical lexical terms related to the topic of the video and the erroneous transcription of technical terms in the captions may have impeded L2 speakers from retuning their phonemic categories using their existing lexical knowledge (Romero-Rivas *et al*., 2015). In this case, the orthographic information did not help categorize ambiguous phonemes due to the high percentage of errors in the automatic captions. As the results from the questionnaire on viewing habits and supporting written content highlighted, it is evident that participants could not use the text in automatic captions as they usually do in their day-to-day life when they are dealing with unfamiliar accents to assist them with speech perception and segmentation. In fact, nearly 90% of participants did not find automatic captions helpful for understanding the content of the clip, stating that the high percentage of errors affected their listening experience and attention, making them feel confused and distracted. Some participants also said that once they realized that captions were not accurate enough to aid speech processing and content comprehension, they tried to ignore the text and focus on the speakers' speech. Therefore, it is plausible that the concurrent presence of accented speech and errors in the text of automatic captions increased both participants' listening effort and cognitive load (e.g., Byrd & Mintz, 2010; Mattys & Wiget, 2011; Van Engen & Peelle, 2014; Peelle, 2017; Porretta &

Tucker, 2019; Porretta *et al*., 2020). These results replicate the findings of previous research, which highlighted the influence of transcription errors on speech processing and content comprehension (Cao *et al*., 2018) and the fundamental role of accuracy in ASR-generated transcriptions to aid content comprehension and guarantee access to information to diverse users (e.g., Bain *et al*., 2002; Butler *et al*., 2019; Chan *et al*., 2019; Romero-Fresco & Fresno, 2023). There is also a probability that errors in the transcriptions may have affected the responses in the comprehension task. Upon closer inspection, it is possible that the answers participants selected for questions n° 1 and 8 in the comprehension test were influenced in part by the misrecognitions of the ASR system and in part by the difficulties with speech processing. The fact that the majority of the responses given for the two questions were those where some keywords were aligned with the text in the captions and which differed from those pronounced in the speech due to variability in pronunciation and phenomena linked to conversational speech (e.g., connected speech - Farnetani & Recasens, 1997).

The number of errors in the automatic captions also influenced L2 speakers' propensity to use ASR systems in class to aid speech processing, enhance content comprehension, and improve their knowledge of English during academic lectures (as some participants noted - since many of the technical terms related to the topic of the video lecture were transcribed erroneously). The main concern arising from the results of one of the questionnaires was indeed related to the number of errors in the text of captions. However, some participants also raised the concern that the presence of captions along with the text in the slides could potentially distract them, hinder content comprehension, and interfere with the learning process. Captions (but more in general, supporting written content) are indeed 'attention-grabber' elements on the screen (Gass *et al.*, 2019). Even if they are experienced users of supporting written content both for aiding speech processing, improving content comprehension, and learning, they are concerned that the additional written input would add a redundant element in an already rich multimodal input, distracting them from the primary task, which is understanding the content of the lecture, but also taking notes. This concern relates directly to the *redundancy effect* theorized by Sweller (2005) and the negative impact of bimodal input supported by some results in studies conducted in the *multimedia learning* framework (e.g., Mayer, 2002). Students may be aware of the quantity of cognitive resources required to process rich multimodal (especially attention, in this case).

In sum, even if their scores for the comprehension task were sufficient, L2 speakers stated that they did not rely on automatic captions as much as they do in their day-to-day lives due to the errors in the text of the transcriptions. They underlined the detrimental effects that errors in the text of automatic captions brought to speech processing, content comprehension, and learning, expressing great

concern for the use of these systems if not adequately tested for their robustness. If the accuracy of the transcriptions is too low, they would prefer not to use this tool since it can have detrimental effects rather than benefits.

RQ2. Do color-coded markups affect content comprehension and attention in L2 speakers of English, hindering speech processing?

RQ3. What opinions do L2 speakers of English hold on the usefulness of displaying the confidence level of the ASR system through different (color-coded) markups?

The results of the analysis aimed at determining if the different display formats affected the performance of L2 speakers revealed a medium, statistically non-significant (p = 0.821) effect of condition on comprehension score. The analysis also revealed an interesting trend in the performance of participants on the basis of the condition they were assigned to. The group of participants who watched the video with captions displayed in the V2 format achieved a higher mean comprehension score compared to the other groups, while the lowest average comprehension score was attested in the group who watched the video with captions in the classic format (baseline condition). On the other hand, participants assigned to the other two groups - namely, OG and V3 - scored a very similar mean comprehension score (6.25 and 6.50, respectively). Subsequent statistical analyses suggested a potential positive (although weak) statistically non-significant effect of the prototypes on comprehension scores (*OG*, p = 0.760; *V2*: p = 0.366; *V3*: p = 0.647) compared to the baseline condition - the classic display format. These results align with those of Berke and colleagues (2017), in which they did not find any statistically significant influence of markups on participants' content comprehension. Moreover, results collected from the questionnaire investigating the opinions and insights of participants on the markups highlighted that, in general, the display formats confused L2 speakers and seeing the changing graphical features distracted them (a similar result was found for L2 speakers assigned to the classic format; these participants would not have found it helpful to have some graphical features indicating the confidence level of the ASR system). Again, similar results were found by Berke and colleagues in their study conducted in 2017. Adding graphical features to the text can be seen as an additional layer of information to be processed by users that could increase the cognitive load in an already rich, multimodal, already cognitively-taxing setting - that is, attending a lecture delivered in English (the L2) on a highly specialized topic at the university, understanding its content, and taking notes.

Regarding the single markups - as we previously mentioned - some participants found one display format to be more informative than the others - that is, the *V2 markup*.

The trends in the results of both the descriptive and the inferential statistical analyses hint at a potential positive effect of the V2 display format on content comprehension higher than the rest of the display formats. Moreover, data collected through the questionnaire investigating the participants' opinions on the graphical features of the markups highlighted that L2 speakers (not only those assigned to this group) found this format informative, even if the changes in the graphical features distracted them. Participants found it helpful to see the markup and check if the ASR system was sure or not of its transcription: this information determined if they could trust the written input to support speech processing. Similarly, some of the participants assigned to watch the video lecture with captions in different conditions found the V2 markup the most promising display format.

Some participants preferred the graphical features contained in the *OG* markup. However, they found it difficult to read the text of the captions in this format, making it harder for them to assess the accuracy of the captions based on the display format. This fact may be because the OG format has three colors (white, grey, and red) compared to the V2 format (white, red): again, any additional information, also linked to the graphical features of the text (compared to only plain white text) may cause an increase in the load on participants' cognitive system, taking off resources that could be devoted to processing the speech in the L2.

The OG and V2 markups are also the display formats that L2 speakers are most inclined to test in class. This preference is due to their effectiveness in showing the confidence of the ASR system in its transcription.

Opposed to what we expected, the V3 display format was the most distracting and confusing for participants. This may be due to the confusion provoked by the underlined words, which, in our idea, indicate that the ASR system is uncertain about its transcription, but in general it may indicate relevant or important words[80].

In summary, color-coded markups overall distracted and confused L2 speakers potentially due to an overload of information in the graphical features of the text of the automatic captions. However, this did not hinder content comprehension; rather, they tended to improve the performance of participants in the comprehension test, especially the V2 format. This format is the most balanced among those developed for the study, as it provides information about three different confidence levels of the ASR system using two colors (white and red) and two graphical features (underlined and strikethrough).

[80] This hypothesis is partially confirmed by a comment made by a participant exactly on this graphical feature.

However, no statistically significant result was found when comparing the average comprehension scores across conditions, nor when assessing the potential effect of the different markups on comprehension scores. For these reasons, a follow-up investigation is needed to confirm or reject the results of this pilot study.

Lastly, participants most appreciated the V2 display format because it was informative regarding the confidence level of the ASR system in its transcription, thereby increasing the system's reliability.

5.6 Limitations

This pilot study has two limitations that need to be acknowledged.

First - as we previously stated - given the small number of participants in each group, we must cautiously interpret the results of the collected data. Future research should assess the optimal sample size by running a power analysis before starting data collection and recruit a higher number of participants to confirm or reject the results of this study.

Second, in the comprehension test, we used two slightly different automatically generated transcriptions for technical reasons (the generation of the markups themselves by the ASR system). One was used to display the captions in the classic and OG formats, while another was used to display the V2 and V3 markups (see §5.3.2.1 for details on the materials). The analysis we ran with *sclite* revealed a small difference in the transcriptions, but there were no apparent substantial differences in the most relevant information in the video. However, we did not conduct a detailed analysis of the transcriptions using the NER model (Romero-Fresco & Pérez, 2015) to determine if the different errors in the text had the same impact on the meaning of the utterances. If detected, potential differences in the text of the automatic captions could affect participants' comprehension and answers in the task. Therefore, using the same script across all conditions is essential to control for potential confoundings in the experiment.

5.7 Conclusions

Research has yet to fully assess the effects of automatic captions on speech processing and content comprehension in diverse individuals in educational settings. However, there is a general consensus that the errors in the transcriptions can negatively impact these linguistic processes.

Previous research has suggested that the development and implementation of specific graphical features (e.g., a color scheme) in the text of the captions to show how confident the system is with its output could be helpful to users in deciding whether to trust or not the automatic transcription (e.g., Wald & Bain, 2008). Results from these studies, however, are mixed, and a similar solution has not been tested yet in L2 speakers of English (e.g., Shiver & Wolfe, 2015; Berke *et al*., 2017).

This chapter reported on a study aimed to investigate 1) the effects of ASR-generated captions on speech processing and content comprehension in L2 speakers of English, and 2) the usefulness and potential benefits of displaying the degree of confidence the ASR system has in its transcription through graphical features (color-coded markups) in the text of captions. We also asked participants to provide their opinions and insights on automatic captions and markups to investigate the effects of the errors and different display formats on speech processing, content comprehension, and attention.

Results showed that the different markups did not affect the outcome of the comprehension test (similarly to Berke *et al*., 2017). However, participants expressed frustration over the high number of errors, which confused and distracted them. As a result, they concluded that automatic captions were not helpful in aiding speech processing or understanding the content of the video they watched. Results also indicated a positive, albeit not significant, trend for the V2 display format to improve content comprehension compared to the other markups, even though it made participants feel confused and distracted. L2 speakers also found this display format more informative and helpful compared to other markups when evaluating the reliability of the ASR system's transcription.

Data needs to be interpreted cautiously: conclusions are not final since there are no statistically significant results, mostly due to the low number of participants per condition. Therefore, results from this pilot study not only call to enlarge the sample size, but also prompt us to conduct this study in real-life settings by evaluating the usefulness of different markups in class using a user-centered design approach. This context could also help us determine if automatic captions could be helpful or redundant since the lectures in class usually have slides with images and text projected onto a screen.

6 General Discussion[81]

6.1 Discussion of findings

In this chapter, we will briefly review the findings from the three studies we conducted (for more details, see Chapters 3, 4, and 5) and consider their broader, applied implications. Next, we will discuss some limitations of this research project and suggest directions for future studies.

The main aims of this doctoral project were to:

1. Investigate the habits of use of audiovisual translation products (captions, interlingual, and intralingual subtitles) in L2 speakers of English;
2. Assess the effects of automatic captions on speech processing and content comprehension in L2 speakers of English;
3. Evaluate the performance of a speaker-independent ASR system in real-world applications by analyzing a corpus of automatically-generated transcriptions with various characteristics (main independent variables: topic and language of the speakers);
4. Assess the reliability of confidence scores as a metric to develop a series of graphical features (i.e., color-coded markups) to be implemented in the text of the captions to signal users the confidence the ASR system has in its transcription;

[81] **Disclaimer** - This chapter includes excerpts from the following publication: Pucci, M. (2023). Towards Universally Designed Communication: Opportunities and Challenges in the Use of Automatic Speech Recognition Systems to Support Access, Understanding and Use of Information in Communicative Settings. In *Design for Inclusion* (pp. 18–25). IOS Press. https://doi.org/10.3233/SHTI230394

5. Test the usefulness and effects the different display formats and markups have on content comprehension and attention in L2 speakers of English.

Results from the three studies we conducted highlight several important points that professionals and educators should remember when considering using ASR-generated captions to support speech processing and content comprehension in educational contexts. At the same time, this research project aims to enhance the understanding of the psycholinguistic mechanisms involved in speech processing (i.e., speech perception and segmentation).

First, the results of our studies highlight the importance of implementing user-centered designs (Vredenburg *et al.*, 2002) in similar research projects, following the steps of the *Liberated Learning Project* (e.g., Bain *et al*., 2002). For instance, in our first study, we collected opinions and insights from the target population (university students L2 speakers of English) on the use of automatic captions and the perceived usefulness of displaying confidence through color-coded markups with the aim of improving the reliability of a tool that still has flaws in real-world contexts (see Chapter 3; e.g., Dua *et al*., 2023). The results from the questionnaire revealed that the participants' primary concern was the accuracy of automatic captions. If the transcriptions were inaccurate, they stated that this tool could have drawbacks rather than benefits in their ability to process speech and comprehend the content of the lectures, confirming previous results of existing literature on the helpfulness of automatic captions in real-world contexts (e.g., Cao *et al*., 2018; Chan *et al*., 2019; Romero-Fresco & Fresno, 2023). Participants also felt that the varying colors used in the text of the automatic captions could be distracting and confusing, potentially hindering their ability to focus on the professor or, more generally, to engage with the lecture content. We took these results into account, as well as the findings from the corpus analysis we conducted in our second study (refer to Chapter 4). Together with our partner company, we then designed the second (V2) and third (V3) versions of the markup, which were tested in the third study alongside the other markups. Their experience with audiovisual translation products and the ASR technology, along with the knowledge of their needs to aid access to information and foster learning, provided great feedback in the development of the color-coded markup we tested in our third study, and in part was also the basis for our hypotheses. This experience showcases a productive collaboration between engineers, researchers, and end users to develop a solution together, always keeping user needs at the forefront.

In sum, this research project calls for a well-refined and ethically approved user-centered approach to be incorporated into the process of technology development, similar to what the *Liberated Learning Project* did almost 25 years ago (Bain *et al*., 2002; Vredenburg *et al*., 2002). Specifically,

- Different users should be encouraged throughout the development process of ASR systems to state their needs, share their doubts, and provide feedback to developers.
- At the same time, developers need to enhance the accuracy of ASR systems by increasing the robustness of speech recognition models (e.g., O'Shaughnessy, 2024).
- Researchers should continue expanding their knowledge of the cognitive mechanisms underlying the processing of bi- and multi-modal information by cooperating with persons with different needs.
- Institutions need to provide the necessary technological tools (e.g., unidirectional microphones) and lecture halls for tools like ASR systems to work in the best conditions to guarantee the support of diverse users (Bencini *et al*., 2018; Van Den Heuij *et al*., 2018; Bencini *et al*., 2021).

Taking a user-centered approach to technology advancement implies that individuals from diverse populations work together with researchers and developers to (I) contribute to an increase in knowledge about the mechanisms behind human cognition and (II) express their views, requirements, and feedback during the development phase of devices.

All around the world, individuals should actively cooperate with academics by participating in research projects that aim at expanding their knowledge of the cognitive mechanisms involved in accessing and processing multimodal information. This refined knowledge should then be transferred to developers for technology development since it will help them create or enhance systems that meet the characteristics of users. At the same time, developers should listen to diverse individuals and encourage them to share their requests, doubts, needs, and feedback during the entire development process of any technological tool. Research on the impact of ASR-generated captions and transcriptions on speech processing and content comprehension where feedback from participants was collected has already demonstrated how imperative it is to listen to diverse users (Ryba *et al*., 2006; Berke *et al*., 2017; Cao *et al*., 2018; Butler *et al*., 2019). Additionally, national and supranational organizations should continue funding projects that integrate basic research with technology advancement, promoting user-centered approaches (Vredenburg *et al*., 2002). It is critical to underline the importance of supporting these projects since progress in this kind of technology is linked to knowledge of the mechanisms behind human functioning.

The second point - as we already briefly mentioned and as the outcomes of our studies underline - concerns the importance of continuing to investigate the cognitive mechanisms involved in speech processing and content comprehension in multimodal contexts.

In our third study, most participants indicated that the accuracy of transcriptions is crucial for supporting their understanding and learning processes. Consequently, they found the automatic captions in the video lectures to be unhelpful, as these captions contained numerous errors that distracted and confused them. On the same page, L2 speakers of English in our first study also found it equally important for supporting speech processing and content comprehension both the close distance between speech and written text, as well as hearing the exact words they see on the screen. Lastly, as an overall trend, most L2 speakers of English in our studies stated that they use audiovisual translation products to become familiar with accents they are not frequently exposed to, helping their cognitive mechanisms retune and compensate for the natural variability that occurs among speakers. These results indicate that L2 speakers - consciously or not - are aware of the benefits that the bi- and multi-modal settings provide, but also what are their limits.

Basic research can also identify the essential characteristics of assistive tools such as automatic captions need to have in order to support speech processes effectively and inform engineers to optimize them for a diverse range of users (e.g., Nettelbeck, 2024). This section, therefore, also wants to stress the importance of considering *natural variability* and not underestimating its effects in both language production and comprehension (since, for instance, certain factors affect speech signals, sometimes determining it to be degraded and difficult to decode, as results from our third study highlighted).

We, as humans, are unique. Each of us has a set of characteristics that need to be considered since they may also affect how we speak and comprehend, which strategies we apply to overcome difficulties or learn, and so on. Therefore, variability in speech processing concerns not only differences in the basic cognitive mechanisms (e.g., working memory capacity - e.g., see Gass *et al*., 2019; attention distribution - e.g., Liao & Kruger, 2023), but also the strategies individuals employ to overcome the difficulties they encounter. Additionally, if individuals use tools to overcome these difficulties, it's essential to identify the key characteristics that these tools should possess to effectively support speech processing and content comprehension (see Chapter 7 for a brief overview on the relevant literature and on our work-in-progress study on the role of proficiency level in the L2 determines the helpfulness of captions).

In summary, observing the behavior of L2 speakers of English in applied settings like this one can provide insights into the fundamental mechanisms that govern speech processing and language comprehension. Some questions, however, still remain open. For instance, which errors in the automatic transcriptions affect speech processing and content comprehension the most? Is there an

interference effect due to the mismatching of written and aural inputs? For these reasons, further research on the processing of bimodal input is still needed.

There is a third and last point to discuss when considering the implementation of ASR in communicative settings.

Designers and instructors should remember that there is no single device or system that guarantees the same level of effective communication and access to information for everyone (and again, this is also related to natural variability among individuals). Therefore, professionals should regard ASR as *one* of the tools that can potentially be employed in real-world applications. As a large body of research has already highlighted, this is because users utilize technology based on their characteristics, needs, and strategies. For example, in contexts where automatic captions were provided, some deaf and hard-of-hearing individuals reported that they relied on ASR-generated transcriptions as an integration to the information received from sign language interpreters. In the same contexts, other individuals entirely focused on the interpreters to have access to information (Berke *et al*., 2017; Butler *et al*., 2019). We found similar results in our studies. For instance, some of the L2 speakers who participated in our first and third studies stated that they preferred not to have automatic captions in class since they felt that the captions would be too redundant, as the text from the slides was already available (redundancy effect - Sweller, 2005). On the contrary, other participants were positively inclined to use this type of audiovisual translation products since it helped them with speech processing and lowering the listening effort imposed by the L2. Therefore, to promote access to information and ensure efficient learning, educational institutions should provide a range of technological tools for students, ask them which devices they prefer, helping them find the ones most suitable for all. At the same time, educational institutions should focus on providing suitable infrastructures (i.e., lecture halls and classrooms - Van Den Heuij *et al*., 2018) that include good acoustics and well-equipped technological instruments (e.g., unidirectional microphones). Research has shown that the acoustic characteristics of the environment where ASR systems operate, as well as the quality of the devices used to capture speech, significantly influence the quality of transcriptions (Alharbi et al., 2021). The results of the analysis we conducted on the corpus of automatic transcriptions also align with existing research. Therefore, institutions must ensure that ASR systems, when adopted as a tool to improve access to information and facilitate communication, operate under optimal conditions. Last but not least, once ASR systems will have reached acceptable accuracy levels, local governments should also consider implementing ASR in public settings such as post offices, banks, hospitals, and municipal/national offices. ASR use will help improve

communication between officials and citizens, facilitating access to information. The implementation of this technology in public and private settings also requires policies aimed at guaranteeing funding to buy these systems and providing support to users. Developers and researchers should continue conversations with supranational institutions, local governments, and the public to stress the importance of access to information and efficient communication.

Technology and ASR can be active players in the progress of society, ensuring equity in communication and access to information in universally designed environments where a service is provided to all.

6.2 Limitations and future directions

Even if we adopted a user-centered approach for the design of this project, we only tested one population, that is L2 speakers of English (native speakers of Italian). Future research should consider developing projects where multiple population groups with diverse individual characteristics (e.g., linguistic background, level of proficiency in the L2) so as to see if these results can be replicated or if they differ due to the natural variability among L2 speakers. Specifically, since studies on the relationship between the use of written supporting content and language proficiency have yielded mixed results, and the majority of studies have tested intermediate and highly proficient L2 speakers, researchers should consider conducting studies with low-proficient speakers so as to investigate their strategies and preference in the use of audiovisual translation products.

As previously reported, research on live subtitles and automatic captions in educational contexts is scarce and inconclusive (e.g., Chan *et al*., 2019; van Gauwbergen *et al*., 2024). For this reason, researchers should design studies that mix more naturalistic experiments with more controlled laboratory-based studies. Specifically, more naturalistic studies can give us insights into the more natural behavior of L2 speakers in an authentic multimodal setting.

Lastly, research on the neural correlates behind the mechanisms involved in processing bi- and multi-modal inputs is still scarce (Montero Perez, 2022; Goh, 2023). Future research, therefore, should utilize online measures like EEG or fNIRS to further investigate this issue so as to enrich the knowledge on the cognitive mechanisms conducted by using the eye tracking methodology (e.g., Liao & Kruger, 2023).

Conclusion

This doctoral project aimed to 1) investigate the habits of using audiovisual translation products (captions, intralingual, and interlingual subtitles) to aid speech processing and content comprehension in L2 speakers of English in their day-to-day life, 2) test the robustness of a speaker-independent ASR system by analyzing a corpus of automatically-generated transcriptions from different domains (politics, academia) with various characteristics (audio quality, spoken language), 3) create a set of graphical features (i.e., color-coded markups) to be implemented in the text of automatic captions using confidence score to display how sure the ASR system is of its transcription, 4) assess the role of automatic captions in in L2 speakers of English in educational contexts, and 5) assess the usefulness and impact on cognitive processes of the different color-coded markups developed.

Chapter 1 and Chapter 2, respectively, covered the relevant literature on the architecture of automatic speech recognition (ASR) systems and the cognitive and psycholinguistic mechanisms behind speech, bimodal, and multimodal input, as well as the existing literature on the effects of audiovisual translation products on speech processing and content comprehension in L2 speakers.

Chapters 3, 4, and 5 reported on the three studies we conducted within the framework of this research project. Chapter 3 reported on the study investigating the habits of use of audiovisual translation products by university students, L2 speakers of English, their inclination to use ASR-generated captions in class during lectures delivered in their L2, and their opinions on the potential implementation of graphical features in the text of automatic captions to signal the confidence level of the system/the errors. Chapter 4 reported on the analysis of a corpus of ASR-generated transcriptions aimed to assess 1) the robustness of the system when dealing with speech with different characteristics (i.e., language, audio quality) recorded in various real-world applications, 2) evaluate the reliability of the confidence scores as a metric to build color-coded markups to be implemented in the text of captions and 3) create two new sets of graphical features. Chapter 5 reported on the

study where we tested the effects of the developed display formats on content comprehension and attention in L2 speakers of English while watching a seminar-style video lecture in English. Finally, in Chapter 6, we discussed the broader implications of the findings from the empirical research we conducted.

Results from the three studies highlighted that:

- Italian university students L2 speakers of English regularly use audiovisual translation products (especially captions and interlingual subtitles) when watching audiovisual products in English, especially to support perceptual retuning in adverse conditions (i.e., when listening to speakers talking with an unfamiliar accent) and content comprehension. They would be inclined to use ASR-generated captions in class during academic lectures delivered in English, but only if the text is accurate enough not to distract or confuse them. They did not find it helpful to use automatic color-coded markups to display the confidence level of the system or transcription errors due to their worry about being distracted while attending lectures;
- The ASR system performed well under all conditions; however, some of the external variables were considered to have affected the transcriptions' accuracy and confidence scores more. The corpus analysis showed a medium correlation between confidence scores and transcription type (correct or incorrect). This suggests that using confidence scores as a metric for creating color-coded markups may not be entirely reliable, as they are heavily influenced by the characteristics of the speech signal being analyzed.
- L2 speakers of English found the markups overall confusing and distracting, but they viewed the V2 markup as informative and helpful for assessing the reliability of the ASR system's transcription, and were keen to test it in class during lectures delivered in English. Content comprehension was not affected by the graphical features of the display formats. Finally, participants found the errors in the transcriptions to be excessive, making automatic captions unhelpful for supporting speech processing.

In conclusion, these findings deepen our understanding of why and when L2 speakers use captions to assist speech processing and content comprehension in diverse contexts, provide insights into if and how this technology can aid information accessibility, and what engineers and (educational) institutions can do to make sure that this tool offers adequate access to information for all.

References

Bibliography

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 520–529. https://doi.org/10.1037/a0013552

Adesope, O. O., & Nesbit, J. C. (2012). Verbal redundancy in multimedia learning environments: A meta-analysis. *Journal of Educational Psychology*, *104*(1), 250–263. https://doi.org/10.1037/a0026147

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, *9*, 131858–131876. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3112535

Alm, A. (2019). Piloting Netflix for intra-formal language learning. *CALL and Complexity - Short Papers from EUROCALL 2019*. EUROCALL 2019.

Anderson, J.R. (1995). Cognitive Psychology and its Implications, 4th Edition. *Freeman*, New York.

Ayres, P., & Sweller, J. (2014). The Split-Attention Principle in Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning* (Cambridge University Press, pp. 206–226).

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *The Annual Review of Psychology*, *63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Bain, K., Basson, S. H., & Wald, M. (2002). Speech recognition in university classrooms: Liberated learning project. *Proceedings of the Fifth International ACM Conference on Assistive Technologies*, 192–196. https://doi.org/10.1145/638249.638284

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, *46*(3), 252–267. https://doi.org/10.1016/j.specom.2005.02.016

Baranowska, K. (2020). Learning most with least effort: Subtitles and cognitive load. *ELT Journal*, *74*(2), 105–115. https://doi.org/10.1093/elt/ccz060

Bencini, G., Arenghi, A., & Garofolo, I. (2021). Is My University Inclusive? Towards a Multi-Domain Instrument for Sustainable Environments in Higher Education. In *Universal Design 2021: From Special to Mainstream Solutions* (pp. 137–143). IOS press.

Bencini, G. M., Garofolo, I., & Arenghi, A. (2018). Implementing universal design and the ICF in higher education: Towards a model that achieves quality higher education for all. In *Transforming our world through design, diversity and education* (pp. 464–472). IOS Press.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Berke, L. (2017). Displaying confidence from imperfect automatic speech recognition for captioning. *SIGACCESS Access. Comput.*, *117*, 14–18. https://doi.org/10.1145/3051519.3051522

Berke, L., Caulfield, C., & Huenerfauth, M. (2017). Deaf and hard-of-hearing perspectives on imperfect automatic speech recognition for captioning one-on-one meetings. *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 155–164.

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. https://doi.org/10.1017/S0142716402004022

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Birulés-Muntané, J., & Soto-Faraco, S. (2016). Watching Subtitled Films Can Help Learning Foreign Languages. *PLOS ONE*, *11*(6), e0158409. https://doi.org/10.1371/journal.pone.0158409

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387. https://doi.org/10.1016/0010-0285(86)90004-6

Boland, J. E., Kaan, E., Kroff, J. V., & Wulff, S. (2016). Psycholinguistics and variation in language processing. *Linguistics Vanguard*, *2*(s1). https://doi.org/10.1515/lingvan-2016-0064

Bolaños García-Escribano, A., Talaván, N., & Fernández-Costales, A. (2024). Audiovisual translation and media accessibility in language education. *Parallèles*, *36*(1), 1–233.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Brown, V. A., McLaughlin, D. J., Strand, J. F., & Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology*, *73*(9), 1431–1443. https://doi.org/10.1177/1747021820916726

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. Journal of Memory and Language, 109, 104047.

Butler, J., Trager, B., & Behm, B. (2019). Exploration of Automatic Speech Recognition for Deaf and Hard of Hearing Students in Higher Education Classes. *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 32–42. https://doi.org/10.1145/3308561.3353772

Byrd, D., & Mintz, T. H. (2010). *Discovering speech, words, and mind*. Wiley-Blackwell.

Cao, X., Yamashita, N., & Ishida, T. (2018). Effects of Automated Transcripts on Non-Native Speakers' Listening Comprehension. *IEICE TRANSACTIONS on Information and Systems*, *E101-D*(3), 730–739.

Carter, B. T., & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology*, *155*, 49–62. https://doi.org/10.1016/j.ijpsycho.2020.05.010

Chan, W. S., Kruger, J.-L., & Doherty, S. (2019). Comparing the impact of automatically generated and corrected subtitles on cognitive load and learning in a first-and second-language educational context. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *18*, 237-272.

Chan, W. S., Kruger, J.-L., & Doherty, S. (2022). An investigation of subtitles as learning support in university education. *Journal of Specialised Translation*, *38*, 155–179.

Charles, T., & Trenkic, D. (2015). Speech segmentation in a second language: The role of bimodal input. *Subtitles and Language Learning: Principles, Strategies and Practical Experiences*, 173–198.

Cintas, J. D., Orero, P., & Remael, A. (2007). *Media for All*. Brill. https://brill.com/display/title/27746

Clinton-Lisell, V. (2023). Does Reading while Listening to Text Improve Comprehension Compared to Reading Only? A Systematic Review and Meta-Analysis. *Educational Research: Theory and Practice*, 34(3), 133-155.

Conklin, K., Alotaibi, S., Pellicer-Sanchez, A., Vilkaite-Lozdiene, L. (2020). What eye-tracking tells us about reading-only and reading-while-listening in a first and second language. *Second Language Research*, 36(3), 257-276.

Cutler, A. & Clifton, C. Jr. (1999). Comprehending spoken language: a blueprint of the listener. In *The Neurocognition of Language*, edited by Brown, C. M. and Hagoort, P. Oxford University Press.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Mit Press.

Danan, M. (2004). Captioning and Subtitling: Undervalued Language Learning Strategies. *Meta: Journal Des Traducteurs / Meta: Translators' Journal*, *49*(1), 67–77. https://doi.org/10.7202/009021ar

Danan, M. (2016). Enhancing listening with captions and transcripts: Exploring learner differences. *Applied Language Learning*, *26*(2), 1–24.

DeGraff, M. (Spring 2017). Lesson 1. Do "Pidgins" exist? Do creoles come from pidgin? Creole Languages and Caribbean Identities - MIT OpenCourseWare 24.908.

Del Rosso, G. A., & Brambilla, S. (2022). L'accuratezza della trascrizione ASR sul parlato non-standard. L'italiano nell'OH Portal. *Fare Linguistica Applicata Con Le Digital Humanities*, 99–116.

Dhanjal, A. S., & Singh, W. (2024). A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications*, *83*(8), 23367–23412. https://doi.org/10.1007/s11042-023-16438-y

Diaz Cintas, J., & Remael, A. (2014). *Audiovisual Translation: Subtitling*. Routledge.

Dizon, G. (2016). Online Video Streaming in the L2 Classroom: Japanese Students' Opinions Towards Netflix and Subtitles. *Osaka JALT Journal*, *3*, 70–87.

Dizon, G., & Thanyawatpokin, B. (2021). Language Learning with Netflix: Exploring the Effects of Dual Subtitles on Vocabulary Learning and Listening Comprehension. *Computer-Assisted Language Learning Electronic Journal*, *22*(3), Article 3.

Dowding, S., Gutwin, C., & Cockburn, A. (2024). User speech rates and preferences for system speech rates. *International Journal of Human-Computer Studies*, *184*, 103222. https://doi.org/10.1016/j.ijhcs.2024.103222

Dua, M., Akanksha, & Dua, S. (2023). Noise robust automatic speech recognition: Review and analysis. *International Journal of Speech Technology*, *26*(2), 475–519. https://doi.org/10.1007/s10772-023-10033-0

Emara, I. F., & Shaker, N. H. (2024). The impact of non-native English speakers' phonological and prosodic features on automatic speech recognition accuracy. *Speech Communication*, *157*, 103038. https://doi.org/10.1016/j.specom.2024.103038

Farnetani, E., & Recasens, D. (1997). Coarticulation and connected speech processes. *The Handbook of Phonetic Sciences*, *371*, 404.

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying Bias in Automatic Speech Recognition (arXiv:2103.15122). arXiv. https://doi.org/10.48550/arXiv.2103.15122

Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. *Computer Speech & Language*, *84*, 101567. https://doi.org/10.1016/j.csl.2023.101567

Floccia, C., Butler, J., Goslin, J., & Ellis, L. (2009). Regional and Foreign Accent Processing in English: Can Listeners Adapt? *Journal of Psycholinguistic Research*, *38*(4), 379–412. https://doi.org/10.1007/s10936-008-9097-8

Furui, S. (2003). Recent Advances in Spontaneous Speech Recognition and Understanding. In *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (SSPR), Tokyo: 1–6.

Garcia Lecumberri, M. L., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52*(11), 864–886. https://doi.org/10.1016/j.specom.2010.08.014

Gass, S., Winke, P., Isbell, D. R., & Ahn, J. (2019). *How captions help people learn languages: A working-memory, eye-tracking study*. http://hdl.handle.net/10125/44684

Gernsbacher, M. A. (2015). Video captions benefit everyone. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 195–202.

Gervain, J., & Mehler, J. (2010). Speech Perception and Language Acquisition in the First Year of Life. *Annual Review of Psychology*, *61*(1), 191–218. https://doi.org/10.1146/annurev.psych.093008.100408

Ghia, E. (2012). *Subtitling matters. New perspectives on subtitling and foreign language learning.* Peter Lang.

Gillespie, B. W., & Atlas, L. E. (2002). Acoustic diversity for improved speech recognition in reverberant environments. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, I-557-I–560. https://doi.org/10.1109/ICASSP.2002.5743778

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

Goh, C. C. M. (2023). Learners' cognitive processing problems during comprehension as a basis for L2 listening research. *System*, *119*, 103164. https://doi.org/10.1016/j.system.2023.103164

Grey, S., Schubel, L. C., McQueen, J. M., & Hell, J. G. V. (2019). Processing foreign-accented speech in a second language: Evidence from ERPs during sentence comprehension in bilinguals. *Bilingualism: Language and Cognition*, *22*(5), 912–929. https://doi.org/10.1017/S1366728918000937

Hamada, Y. (n.d.). Aural Decoding and Comprehension in L2 Listening. *International Journal of Applied Linguistics*, *n/a*(n/a). https://doi.org/10.1111/ijal.12747

Hoey, M. (2012). *Lexical priming: A new theory of words and language*. Routledge.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674. doi:10.1198/106186006X133933

Hothorn T, Buehlmann P, Dudoit S, Molinaro A, Van Der Laan M (2006). "Survival Ensembles." *Biostatistics*, **7**(3), 355–373. doi:10.1093/biostatistics/kxj011

Hosogoshi, K. (2024). Effects of captions and subtitles on the listening process: Insights from EFL learners' listening strategies. *The JALT CALL Journal*, *12*(3), Article 3. https://doi.org/10.29140/jaltcall.v12n3.j206

Hulstijn, J. H. (2003). Connectionist Models of Language Processing and the Training of Listening Skills with the Aid of Multimedia Software. *Computer Assisted Language Learning*, *16*(5), 413–425. https://doi.org/10.1076/call.16.5.413.29488

Inceoglu, S., Chen, W.-H., & Lim, H. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. *ReCALL*, *35*(1), 89–104. https://doi.org/10.1017/S0958344022000192

Ito, A., & Pickering, M. J. (2021). Chapter 2. Automaticity and prediction in non-native language comprehension. In E. Kaan & T. Grüter (Eds.), *Prediction in Second Language Processing and Learning* (pp. 25–46). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.12.02ito

Ivanchenko, S. (2023). The Effects of Captioning on Content Comprehension. [Unpublished master's thesis]. Ca' Foscari University of Venice. http://hdl.handle.net/10579/25176

Ivarsson, J. (2009). The history of subtitles in Europe. *Dubbing and Subtitling in a World Context*, 3–12.

Jia, C., & Hew, K. F. (2023). Meeting the challenges of decoding training in English as a foreign/second language listening education: Current status and opportunities for technology-assisted decoding training. *Computer Assisted Language Learning*, *36*(5–6), 1116–1145. https://doi.org/10.1080/09588221.2021.1974051

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, *21*(1), 47–77. https://doi.org/10.1093/applin/21.1.47

Juang, B.-H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, *1*, 67.

Juffs, A. (2013). Second language acquisition of the lexicon. In Ritchie, W. and Bhatia, T. (eds.). *The New Handbook of Second Language Acquisition*. Amsterdam, The Netherlands: Elsevier.

Jurafsky, D., Martin, J. H. (2025). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released January 12, 2025. https://web.stanford.edu/~jurafsky/slp3.

Kalyuga, S., Chandler, P., & Sweller, J. (2004). When Redundant On-Screen Text in Multimedia Technical Instruction Can Interfere With Learning. *Human Factors*, *46*(3), 567–581. https://doi.org/10.1518/hfes.46.3.567.50405

Kalyuga, S., & Sweller, J. (2014). 10 The Redundancy Principle in Multimedia Learning. *The Cambridge Handbook of Multimedia Learning*, 247.

Karpagavalli, S., & Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *9*(4), 393–404.

Ke, Z., & Wang, Y. (2022). Exploring the relationship between aural decoding and listening comprehension among L2 learners of English. *System*, *104*, 102688. https://doi.org/10.1016/j.system.2021.102688

Kruger, J.-L. (2013). Subtitles in the classroom: Balancing the benefits of dual coding with the cost of increased cognitive load. *Journal for Language Teaching*, *47*(1), 29–53. https://doi.org/10.10520/EJC143069

Kruger, J.-L. (2016). Psycholinguistics and audiovisual translation. *Target. International Journal of Translation Studies*, *28*(2), 276–287. https://doi.org/10.1075/target.28.2.08kru

Kruger, J.-L., & Liao, S. (2022). Establishing a theoretical framework for AVT research: The importance of cognitive models. *Translation Spaces*, *11*(1), 12–37. https://doi.org/10.1075/ts.21024.kru

Kruger, J.-L., Wisniewska, N., & Liao, S. (2022). Why subtitle speed matters: Evidence from word skipping and rereading. *Applied Psycholinguistics*, *43*(1), 211–236. https://doi.org/10.1017/S0142716421000503

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the Accuracy of Automatic Speech Recognition Solutions. *ACM Trans. Access. Comput.*, *16*(4), 25:1-25:23. https://doi.org/10.1145/3636513

Kuo, F. (2004). Open and closed: Captioning technology as a means to equality. *The John Marshall Journal of Computer & Information Law*, *23*(1), 159–207.

Laurent, A., Meignier, S., & Deléglise, P. (2014). Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. *Computer Speech & Language*, *28*(4), 979–996. https://doi.org/10.1016/j.csl.2014.02.006

Levshina, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.

Levshina, N. (2020). Conditional Inference Trees and Random Forests. In M. Paquot & S. Th. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 611–643). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_25

Li, Q. (2018). *Confidence score for speech processing* [University of Cambridge]. Retrieved May 14, 2023, from https://mi.eng.cam.ac.uk/~ql264/doc/MEng_thesis.pdf

Liao, S., Yu, L., Reichle, E., D. & and Kruger, J.-L. (2021). Using Eye Movements to Study the Reading of Subtitles in Video. *Scientific Studies of Reading*, *25*(5), 417–435. https://doi.org/10.1080/10888438.2020.1823986

Liao, S., Yu, L., Kruger, J.-L., & Reichle, E. D. (2022). The impact of audio on the reading of intralingual versus interlingual subtitles: Evidence from eye movements. *Applied Psycholinguistics*, *43*(1), 237–269. https://doi.org/10.1017/S0142716421000527

Liao, S., & Kruger, J.-L. (2023). Cognitive processing of subtitles: Charting the future by mapping the past. In *The Routledge Handbook of Translation, Interpreting and Bilingualism*. Routledge.

Matielo, R., D'Ely, R. C. S. F., & Baretta, L. (2015). The effects of interlingual and intralingual subtitles on second language learning/acquisition: A state-of-the-art review. *Trabalhos Em Linguística Aplicada*, *54*, 161–182. https://doi.org/10.1590/0103-18134456147091

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, *65*(2), 145–160. https://doi.org/10.1016/j.jml.2011.04.004

Mayer, R. E. (2002). Multimedia learning. In *Psychology of Learning and Motivation* (Vol. 41, pp. 85–139). Academic Press. https://doi.org/10.1016/S0079-7421(02)80005-6

Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. Cambridge university press.

Mayer, R. E. (2024). The Past, Present, and Future of the Cognitive Theory of Multimedia Learning. *Educational Psychology Review*, *36*(1), 8. https://doi.org/10.1007/s10648-023-09842-1

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

McClelland, J. L. & Elman, J. L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology*, 18, 1-86.

Mitterer, H., & McQueen, J. M. (2009). Foreign Subtitles Help but Native-Language Subtitles Harm Foreign Speech Perception. *PLOS ONE*, *4*(11), e7785. https://doi.org/10.1371/journal.pone.0007785

Montero Perez, M. (2020). Incidental Vocabulary Learning Through Viewing Video: The Role of Vocabulary Knowledge and Working Memory. *Studies in Second Language Acquisition*, *42*(4), 749–773. https://doi.org/10.1017/S0272263119000706

Montero Perez, M. (2022). Second or foreign language learning through watching audio-visual input and the role of on-screen text. *Language Teaching*, *55*(2), 163–192.

Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology*, *18*(1), 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

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology*, *94*(1), 156–163. https://doi.org/10.1037/0022-0663.94.1.156

Muñoz, C. (2022). Audiovisual input in L2 learning. *Language, Interaction and Acquisition*, *13*(1), 125–143. https://doi.org/10.1075/lia.22001.mun

Munro, M. J., & Derwing, T. M. (1995). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Nagels, K. (2012). 'Those funny subtitles': Silent film intertitles in exhibition and discourse. *Early Popular Visual Culture*, *10*(4), 367–382. https://doi.org/10.1080/17460654.2012.724570

Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, *22*(2), 171–184. https://doi.org/10.1016/j.csl.2007.07.003

Nettelbeck, H. W. (n.d.). Are subtitling norms evidence-based? A narrative comparison of guidelines and subtitling advice with empirical evidence. *Perspectives*, *0*(0), 1–17. https://doi.org/10.1080/0907676X.2024.2421773

Ohala, D. K. (2008). Phonological acquisition in a first language. In *Phonology and Second Language Acquisition*, edited by Hansen Edwards, J. G. and Zampini, M. L. John Benjamins Publishing Company.

O'Shaughnessy, D. (2024). Trends and developments in automatic speech recognition research. *Computer Speech & Language*, *83*, 101538. https://doi.org/10.1016/j.csl.2023.101538

Paivio, A. (1986). *Mental Representations: A Dual-Coding Approach.* (Oxford University Press).

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *45*(3), 255.

Pal, S., Badhe, V., & Dasgupta, C. (2023). Investigating the Role of Closed Captioning and Live Transcription on DHH Students' Perception of Inclusivity in a Face-to-Face Classroom Environment.

2023 IEEE International Conference on Advanced Learning Technologies (ICALT), 218–220. https://doi.org/10.1109/ICALT58122.2023.00070

Palmerini, M., & Savy, R. (2014). Gli errori di un sistema di riconoscimento automatico del parlato: Analisi linguistica e primi risultati di una ricerca interdisciplinare. *Proceedings of the First Italian Conference on Computational Linguistics CLiC-It 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa*, 281–285.

Pattemore, A., & Muñoz, C. (2020). Learning L2 constructions from captioned audio-visual exposure: The effect of learner-related factors. *System*, *93*, 102303. https://doi.org/10.1016/j.system.2020.102303

Peelle, J. E. (2018). Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear and Hearing*, *39*(2), 204. https://doi.org/10.1097/AUD.0000000000000494

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*, 51, 195-203. 10.3758/s13428-018-01193-y

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. https://doi.org/10.1017/S0140525X04000056

Piquard-Kipffer, A., Mella, O., Miranda, J., Jouvet, D., & Orosanu, L. (2015). Qualitative investigation of the display of speech recognition results for communication with deaf people. *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 36–41. https://aclanthology.org/W15-5107.pdf

Pisoni, D. B. (2017). Speech Perception: Research, Theory, and Clinical Application. In *The Handbook of Psycholinguistics* (pp. 193–212). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118829516.ch9

Porretta, V., Buchanan, L., & Järvikivi, J. (2020). When processing costs impact predictive processing: The case of foreign-accented speech and accent experience. *Attention, Perception, & Psychophysics*, *82*(4), 1558–1565. https://doi.org/10.3758/s13414-019-01946-7

Porretta, V., & Tucker, B. V. (2019). Eyes Wide Open: Pupillary Response to a Foreign Accent Varying in Intelligibility. *Frontiers in Communication*, *4*. https://doi.org/10.3389/fcomm.2019.00008

Pucci, M. (2023). Towards Universally Designed Communication: Opportunities and Challenges in the Use of Automatic Speech Recognition Systems to Support Access, Understanding and Use of Information in Communicative Settings. In *Design for Inclusion* (pp. 18–25). IOS Press. https://doi.org/10.3233/SHTI230394

Pujadas, G., & Muñoz, C. (2024). When to switch captions off? Exploring the effects of L2 proficiency and vocabulary knowledge on comprehension of captioned and uncaptioned TV. *Studies in Second Language Learning and Teaching*, *14*(3), Article 3. https://doi.org/10.14746/ssllt.38036

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rajendran, D. J., Duchowski, A. T., Orero, P., Martínez, J., & Romero-Fresco, P. (2013). Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, *21*(1), 5–21. https://doi.org/10.1080/0907676X.2012.722651

Reichle, E. D. (2021). *Computational models of reading: A handbook*. Oxford University Press.

Robert, I. S., De Meulder, A., & Schrijver, I. (2021). Live subtitling for access to education: A pilot study of university students' reception of intralingual live subtitles. *JoSTrans: The Journal of Specialised Translation. London, 2003, Currens*, *36*, 53–78.

Romero-Fresco, P., Amigo, Ó. A., & Bacigalupe, L. A. (2024). The use of artificial intelligence in the assessment of live subtitling quality: The NER Buddy. *Tradumàtica Tecnologies de La Traducció*, *22*, Article 22. https://doi.org/10.5565/rev/tradumatica.408

Romero-Fresco, P., & Fresno, N. (2023). The accuracy of automatic and human live captions in English. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, *22*. https://doi.org/10.52034/lans-tts.v22i.774

Romero-Fresco, P., & Pérez, J. M. (2015). Accuracy Rate in Live Subtitling: The NER Model. In R. B. Piñero & J. D. Cintas (Eds.), *Audiovisual Translation in a Global Context: Mapping an Ever-changing Landscape* (pp. 28–50). Palgrave Macmillan UK. https://doi.org/10.1057/9781137552891_3

Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00167

Ryba, K., McIvor, T., Shakir, M., & Paez, D. (2006). Liberated Learning: Analysis of University Students' Perceptions and Experiences with Continuous Automated Speech Recognition. *E-Journal of Instructional Science and Technology*, *9*(1), n1.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saraclar, M., Riley, M., Bocchieri, E., & Goffin, V. (2002). Towards automatic closed captioning: Low latency real time broadcast news transcription. *Interspeech*, 1741–1744. https://www.academia.edu/download/50297684/icslp02.pdf

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, *49*(5), 336–347. https://doi.org/10.1016/j.specom.2007.01.009

Shimogori, N., Ikeda, T., & Tsuboi, S. (2010). Automatically generated captions: Will they help non-native speakers communicate in english? *Proceedings of the 3rd International Conference on Intercultural Collaboration*, 79–86.

Shiver, B. N., & Wolfe, R. J. (2015). Evaluating Alternatives for Better Deaf Accessibility to Selected Web-Based Multimedia. *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility - ASSETS '15*, 231–238. https://doi.org/10.1145/2700648.2809857

Strobl C, Boulesteix A, Zeileis A, Hothorn T (2007). "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics*, **8**(25). doi:10.1186/1471-2105-8-25.

Strobl C, Boulesteix A, Kneib T, Augustin T, Zeileis A (2008). "Conditional Variable Importance for Random Forests." *BMC Bioinformatics*, **9**(307). doi:10.1186/1471-2105-9-307.

Swarup, P., Maas, R., Garimella, S., Mallidi, S. H., & Hoffmeister, B. (2019). Improving ASR Confidence Scores for Alexa Using Acoustic and Hypothesis Embeddings. *Interspeech 2019*, 2175–2179. https://doi.org/10.21437/Interspeech.2019-1241

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J. (2004). Instructional Design Consequences of an Analogy between Evolution by Natural Selection and Human Cognitive Architecture. *Instructional Science*, *32*(1), 9–31. https://doi.org/10.1023/B:TRUC.0000021808.72598.4d

Sweller, J. (2005). The Redundancy Principle in Multimedia Learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 159–168). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.011

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3617–3632. https://doi.org/10.1098/rstb.2009.0107

Sydorenko, T. (2010). *Modality of input and vocabulary acquisition*. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44214/1/14_02_sydorenko.pdf

Szarkowska, A., & Gerber-Morón, O. (2018). Viewers can keep up with fast subtitles: Evidence from eye movements. *PLOS ONE*, *13*(6), e0199331. https://doi.org/10.1371/journal.pone.0199331

Szarkowska, A., & Gerber-Morón, O. (2019). Two or three lines: A mixed-methods study on subtitle processing and preferences. *Perspectives*, *27*(1), 144–164. https://doi.org/10.1080/0907676X.2018.1520267

Szarkowska, A., Krejtz, I., Klyszejko, Z., & Wieczorek, A. (2011). Verbatim, Standard, or Edited? Reading Patterns of Different Captioning Styles Among Deaf, Hard of Hearing, and Hearing Viewers. *American Annals of the Deaf*, *156*(4), 363–378.

Szarkowska, A., Ragni, V., Szkriba, S., Black, S., Kruger, J.-L., & Orrego-Carmona, D. (2024). 'That's not what they said!' The impact of incongruities between the dialogue and intralingual subtitles on viewer experience. *Perspectives*, 1–20. https://doi.org/10.1080/0907676X.2024.2386040

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, *24*(2), 135–178. https://doi.org/10.1017/S0954394512000129

Tauroza, S., & Allison, D. (1990). Speech Rates in British English. *Applied Linguistics*, *11*(1), 90–105. https://doi.org/10.1093/applin/11.1.90

Taylor, G. (2005). Perceived Processing Strategies of Students Watching Captioned Video. *Foreign Language Annals*, *38*(3), 422–427. https://doi.org/10.1111/j.1944-9720.2005.tb02228.x

Teng, M. F. (2022). Incidental L2 vocabulary learning from viewing captioned videos: Effects of learner-related factors. *System*, *105*, 102736. https://doi.org/10.1016/j.system.2022.102736

To, C. (2024). Are subtitles useful for language learners? *Journal of Language Teaching*, *4*(2), 1–6.

Tomasello, M. (2000). Do young children have adult syntactic competence? Cognition, 74, 209-253.

Van Den Heuij, K. M. L., Neijenhuis, K., & Coene, M. (2018). Acoustic environments that support equally accessible oral higher education as a human right. *International Journal of Speech-Language Pathology*, *20*(1), 108–114. https://doi.org/10.1080/17549507.2017.1413136

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00577

Van Gauwbergen, Y., Robert, I. S., & Schrijver, I. (2024). The effect of intralingual live subtitling on students' performance and perception in EMI lectures in Flanders: A pilot study. *Journal of English for Academic Purposes*, *72*, 101445. https://doi.org/10.1016/j.jeap.2024.101445

Van Rossum, G. (2007, June). Python Programming Language. In USENIX annual technical conference (Vol. 41, No. 1, pp. 1-36).

Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching*, *43*(1), 1–37. https://doi.org/10.1017/S0261444809990267

Vanderplank, R. (2016). 'Effects of' and 'effects with' captions: How exactly does watching a TV programme with same-language subtitles make a difference to language learners? *Language Teaching*, *49*(2), 235–250. https://doi.org/10.1017/S0261444813000207

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. https://doi.org/10.1080/09571736.2019.1610033

Venturini, S. (2022). Do captions benefit everyone? Comparing the effects of automatic versus human captions on learner comprehension. [Closed access master's thesis]. Ca' Foscari University of Venice. http://hdl.handle.net/10579/21607

Venturini, S., Vann, M. M., Pucci, M., & Bencini, G. M. (2022). Towards a More Inclusive Learning Environment: The Importance of Providing Captions That Are Suited to Learners' Language Proficiency in the UDL Classroom. In *Transforming our World through Universal Design for Human Development* (pp. 533–540). IOS Press.

Vertanen, K., & Kristensson, P. O. (2008). On the benefits of confidence visualization in speech recognition. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1497–1500. https://doi.org/10.1145/1357054.1357288

Vredenburg, K., Mao, J.-Y., Smith, P. W., & Carey, T. (2002). *A Survey of User-Centered Design Practice*. *1*.

Wald, M. (2006a). An exploration of the potential of Automatic Speech Recognition to assist and enable receptive communication in higher education. *ALT-J*, *14*(1), 9–20.

Wald, M. (2006b). Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education*.

Wald, M. (2007). A research agenda for transforming pedagogy and enhancing inclusive learning through synchronised multimedia captioned using speech recognition. *EdMedia+ Innovate Learning*, 4479–4485.

Wald, M., & Bain, K. (2008). Universal access to communication and learning: The role of automatic speech recognition. *Universal Access in the Information Society*, *6*(4), 435–447.

Wang, Y., & Daghigh, A. J. (2024). Two Decades of Audiovisual Translation Studies: A Bibliometric Literature Review. *SAGE Open*, *14*(3), 21582440241274575. https://doi.org/10.1177/21582440241274575

Warren, P. (2013). *Introducing Psycholinguistics*. Cambridge University Press.

Williams, G. (1998). *A Study of the Use and Evaluation of Confidence Measures in Automatic Speech Recognition*.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

Wisniewska, N., & Mora, J. C. (2020). Can Captioned Video Benefit Second Language Pronunciation? *Studies in Second Language Acquisition*, *42*(3), 599–624. https://doi.org/10.1017/S0272263120000029

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

Yu, D., & Deng, L. (2015). *Automatic Speech Recognition: A Deep Learning Approach*. Springer. https://doi.org/10.1007/978-1-4471-5779-3

Yuan, J., Liberman, M., & Cieri, C. (2006, September 17). Towards an integrated understanding of speaking rate in conversation. *Interspeech 2006*. Interspeech 2006, ISCA. https://doi.org/10.21437/interspeech.2006-204

Zeileis A, Hothorn T, Hornik K (2008). "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**(2), 492–514. doi:10.1198/106186008X319331

Webography

Python | PoS Tagging and Lemmatization using spaCy. (2023). GeeksforGeeks. Retrieved January 4, 2025, from https://www.geeksforgeeks.org/python-pos-tagging-and-lemmatization-using-spacy/.

2024 State of Captioning Report. (2024). Retrieved February 18, 2025, from https://go.3playmedia.com/soc-2024.

Definition and overview | Centre for Excellence in Universal Design. (2022). Universaldesign.Ie. Retrieved December 16, 2022, from https://universaldesign.ie/what-is-universal-design/definition-and-overview/.

Digital Education Action Plan (2021-2027) | European Education Area. (n.d.). Retrieved January 8, 2023, from https://education.ec.europa.eu/node/1518.

English (USA) Timed Text Style Guide. (n.d.). Netflix | Partner Help Center. Retrieved January 15, 2025, from https://partnerhelp.netflixstudios.com/hc/en-us/articles/217350977-English-USA-Timed-Text-Style-Guide.

Grammarly: Free AI Writing Assistance. (n.d.). Retrieved March 12, 2025, from https://www.grammarly.com/.

ICT for Inclusion. (n.d.). European Agency for Special Needs and Inclusive Education. Retrieved January 8, 2023, from https://www.european-agency.org/activities/ict4i.

Intelligenza artificiale (AI). (n.d.). Università Ca' Foscari Venezia. Retrieved March 12, 2025, from https://www.unive.it/pag/49804.

Lewis, E. (2019, July 19). *Verizon Media and Publicis Media Find Viewers Want Captions*. 3Play Media. https://www.3playmedia.com/blog/verizon-media-and-publicis-media-find-viewers-want-captions/.

National Institute of Standards and Technology. (2025, February 5). [Text]. NIST. https://www.nist.gov/.

PNSD – Scuoladigitale. (n.d.). Retrieved December 16, 2022, from https://scuoladigitale.istruzione.it/pnsd/.

Survey: Why America Is Obsessed with Subtitles. (2022, June 17). https://preply.com/en/blog/americas-subtitles-use/.

The 7 Principles | Centre for Excellence in Universal Design. (n.d.). Retrieved December 16, 2022, from https://universaldesign.ie/what-is-universal-design/the-7-principles/.

UDL: Offer alternatives for auditory information. (n.d.). Retrieved January 8, 2023, from https://udlguidelines.cast.org/representation/perception/alternatives-auditory.

UDL: The UDL Guidelines. (n.d.). Retrieved December 16, 2022, from https://udlguidelines.cast.org/.

Usnistgov/SCTK. (2025). [C]. National Institute of Standards and Technology. https://github.com/usnistgov/SCTK (Original work published 2016).

Wallace, J. (2023, March 13). *Captions vs. Subtitles: Breaking Down the Differences*. 3Play Media. https://www.3playmedia.com/blog/closed-captioning-vs-subtitles/.

Why America is (Still) Obsessed with Subtitles. (2023, July 17). https://preply.com/en/blog/america-still-obsessed-subtitles/.

Twenty-First Century Communications and Video Accessibility Act | Federal Communications Commission. (2010). Retrieved March 15, 2025, from https://www.fcc.gov/cvaa

Appendix

A. Questionnaire on participants' viewing habits and use of audiovisual translation products while watching audiovisual products in English

I. First version of the questionnaire

In this part you will be asked to complete a questionnaire on your habits of viewing audiovisual content in English and captions use.

1. Do you watch audiovisual content (e.g., movies, TV series, short clips, etc.) in English at home, university, etc.?
    a. Yes
    b. No
2. How often do you watch audiovisual content in English in your day-to-day life (e.g., movies, TV series, social media, etc.)?
    a. Never
    b. Rarely
    c. A few times a week
    d. Every day
3. What kind of audiovisual content in English do you watch? Choose all that apply to you.
    a. TV news online/on YouTube
    b. Documentaries on streaming platforms
    c. Short clips on TikTok/Instagram/other video platforms
    d. Lectures on YouTube/Coursera/edx/other websites
    e. TV series
    f. Movies
    g. Vloggers on YouTube

h. Learning videos from textbooks of English

i. I do not watch videos in English

j. Other (please, write down your answer)

4. When you watch audiovisual content in English, you mostly watch it…

   a. On a desktop computer/PC

   b. On a laptop

   c. On a smartphone

   d. On a tablet

   e. I do not watch audiovisual content in English

5. Do you use captions, subtitles, etc. when watching audiovisual content in English?

   a. Yes

   b. No

6. Do you know the difference between **captions** and **subtitles**?

   Captions are transcripts (written copies of spoken language) of speakers' original speech. The language of the text on the screen coincides with the language in the audio. **Subtitles** refer to text representing the content of the audio in an audiovisual product (e.g., movies, TV series, etc.) in the same or in a different language of oral speech.

   Do you prefer using captions or subtitles?

   a. Captions

   b. Subtitles

   c. I don’t use captions/subtitles

7. Could you please motivate your previous answer? You can choose more than one answer.

   a. To understand where words end by reading the written transcript

   b. To make sure I understand the content of what I am watching/listening

   c. To learn new words

   d. To improve/learn words pronunciation

e. I do not turn captions/subtitles on because I prefer listening to what is being said rather than reading captions/subtitles

f. I find it hard to concentrate if I have to pay attention both to the audio and the text

g. I do not turn captions/subtitles on because I try to improve my listening skills, and I don't think that written support helps

h. I do not turn captions/subtitles on because I don't need them

i. Other (please, write down your answer)

8. How often do you use captions/subtitles when watching audiovisual content in English?

    a. Never

    b. Rarely

    c. A few times a week

    d. Every day

9. In which language do you prefer subtitles/captions?

    a. I do not use captions/subtitles

    b. In my native language (subtitles)

    c. In the same language of the oral speech (captions)

    d. In the same language of the oral speech (subtitles)

10. If you turn captions/subtitles on, how important is it for you to read the same words that you are hearing when watching audiovisual content in English?

    a. Absolutely not important

    b. Not important

    c. Pretty much important

    d. Absolutely important

11. If you turn captions/subtitles on, how important is it for you to read the same words that you are hearing appearing at the same time on the screen when watching audiovisual content in English?

a. Absolutely not important

b. Not important

c. Pretty much important

d. Absolutely important

II. Second version of the questionnaire

We would like you to complete **a questionnaire**. The questionnaire is divided into two parts:

The first part aims at investigating your habits when using captions/subtitles while watching audiovisual content in English and your preferences regarding the formats in which captions/subtitles are displayed, and the role of captions/subtitles in learning/improving your knowledge of English.

(The second part of the instructions can be found in Appendix B.II).

1. Do you watch audiovisual content (e.g., movies, TV series, short clips, etc.) in English in your day-to-day life?
    a. Yes
    b. No
2. How often do you watch audiovisual content in English in your day-to-day life (e.g., movies, TV series, etc.)?
    a. Never
    b. Rarely
    c. A few times a week
    d. Every day
3. What kind of audiovisual content in English do you watch? Choose all that apply to you.
    a. TV news online/on YouTube
    b. Documentaries on streaming platforms
    c. Short clips on TikTok/Instagram/other video platforms

d. Lectures on YouTube/Coursera/edx/other websites

e. TV series

f. Movies

g. Vloggers on YouTube

h. Learning videos from textbooks of English

i. I do not watch audiovisual content in English

4. When you watch audiovisual content in English, you mostly watch it…

   a. On a desktop computer/PC

   b. On a laptop

   c. On a smartphone

   d. On a tablet

   e. I do not watch audiovisual content in English

5. It is common to use supporting written content (captions, subtitles, etc.) when watching audiovisual content (e.g., on a streaming platform while watching a movie). Do you know the difference between **captions** and **subtitles**?

   Captions are transcripts of speakers' original speech. The language of the text on the screen coincides with the language in the audio.

   Subtitling is a translation practice. To create **subtitles**, translators adapt the original dialogue to include information also from the image and the soundtrack (Diaz Cintas & Remael, 2014). Subtitles can be in the same or in a different language of the oral speech (usually, your native language).

   Do you use captions/subtitles when watching audiovisual content in English?

   a. Yes

   b. No

6. Which type of supporting written content (captions, subtitles, etc.) do you prefer using when watching audiovisual content in English?

a. Captions (written transcript of the audio)

b. Subtitles (same language of the audio)

c. Subtitles (translation of the audio in your native language)

d. I don’t use captions/subtitles

e. It depends (please, specify by selecting the option 'Other' below and writing your answer there)

f. Other (please, write down your answer).

7. Why do you prefer that type of supporting written content (captions, subtitles, etc.)? Could you please motivate your previous answer? You can choose more than one answer.

    a. I find it hard to concentrate if I have to listen to the audio and read the text at the same time

    b. I don't use captions/subtitles because I want to improve my listening skills, and I don't think that written support helps

    c. I don't use captions/subtitles because I prefer listening to what is being said rather than reading the text.

    d. I don't use captions/subtitles because I don't need them.

    e. I need to identify where words begin and end by reading the written transcript

    f. I want to learn new words.

    g. I want to improve/learn the pronunciation of words.

    h. I prefer reading the exact word pronounced by the speakers

    i. I want to make sure I understand the content of what I am watching/listening to.

    j. I want to know what’s the translation in my native language of the words in the speech.

    k. I want to know what's the meaning of a word in my native language

    l. I prefer reading each word as soon as it is pronounced by speakers

    m. I prefer reading the text in my native language

    n. It requires less effort to read the text in my native language

o. Other (please, write down your answer).

8. How often do you use these formats when watching audiovisual content in English? *[one answer for each type of written supporting content, i.e. captions, intralingual subtitles, interlingual subtitles]*

    a. Never

    b. Rarely

    c. Frequently

    d. Always

9. If you turn captions/subtitles on, how important is it for you to read **the same words that you are hearing** when watching audiovisual content in English?

    a. Absolutely not important

    b. Not important

    c. Important

    d. Absolutely important

10. If you turn captions/subtitles on, how important is it for you to read **the same words you are hearing appear at the same time on the screen** when watching audiovisual content in English?

    a. Absolutely not important

    b. Not important

    c. Important

    d. Absolutely important

11. You are given the possibility to choose how text will be displayed on screen (on two lines, word-by-word, etc.) when watching a video in English. Which format would you use?

    a. Text on one line and stays on screen for a few seconds

    b. Text on two lines and stays on screen for a few seconds

    c. Text on one line that changes color/is highlighted when each word is pronounced by speakers

d. Text that appears word by word, synchronized with speech

e. I won’t use captions/subtitles

f. It depends (please, specify by selecting the 'Altro' option and writing your answer there)

g. Other (please, write down your answer)

12. Why would you choose that format? Could you please motivate your previous answer? Select all that apply to you.

    a. I need to identify where words begin and end by reading the written transcript

    b. I want to make sure I understand the content of what I am watching/listening

    c. I prefer reading each word as soon as they are pronounced by speakers

    d. I don’t use captions/subtitles because I prefer listening to what is being said rather than reading the text

    e. I find it hard to concentrate if I have to listen to the audio and read the text at the same time

    f. I don’t use captions/subtitles because I want to improve my listening skills, and I don't think that written support helps

    g. I don’t use captions/subtitles because I don't need them

    h. Other (please, write down your answer)

13. Reading captions/subtitles, listening to speech and watching the images on screen at the same time can be difficult. Which of the following statements are true for you? Choose all that apply to you.

    a. It’s difficult for me to read captions/subtitles when watching audiovisual content in English.

    b. I need to concentrate really hard to read captions/subtitles and listen to speakers talking.

    c. Sometimes I feel annoyed by captions/subtitles appearing on screen.

    d. Captions/subtitles easily attract my attention – I cannot concentrate as much as I want on the image nor the speakers’ speech.

e. It's not particularly difficult for me reading captions/subtitles and listening to speakers while watching audiovisual content in English.

f. I feel reassured if captions/subtitles are on – I rely on text occasionally if I miss some words while listening.

g. If I find it hard or too overwhelming just to listen to the speakers' talking, I rely on captions/subtitles.

h. Captions/subtitles help me understanding the content of the video – I'm used to watching audiovisual content in English with captions/subtitles on.

i. It depends (please, specify by selecting the 'Altro' option and writing your answer there)

j. Other (please, write down your answer)

14. Think about the characteristics the text of captions/subtitles can have on screen. In your opinion, which of these characteristics help you the most in understanding speech in English when watching audiovisual content? Choose all that apply to you.

    a. Font

    b. Color of the font

    c. The format of the captions/subtitles (two or more lines, word-by-word, etc.)

    d. Size of the text

    e. Size of the captions/subtitles box

    f. Color of the background if the captions/subtitles box is present

    g. Misalignment (captions/subtitles appear much later/earlier than the speech you hear)

    h. Errors in captions/subtitles

    i. Other (please, write down your answer)

15. Why do you think these characteristics are the most relevant for you to understand speech in English when watching audiovisual content? Please, write your answer below *[open question]*.

16. Did you use to watch audiovisual content in English to improve your knowledge of the language **when your proficiency in English was lower**?

a. Yes

b. No

17. Did you use captions/subtitles when watching audiovisual content in English to improve your knowledge of the language **when your proficiency in English was lower**?

    a. Yes

    b. No

18. Why did you use captions/subtitles to learn English **when your proficiency was lower**? Choose all that apply to you.

    a. I could easily spot words I didn't know and learn them when the translation was provided in the subtitles

    b. I could spot words I didn't know and learn them in English

    c. I could learn the meaning of a word by reading the supporting written text

    d. I could learn the pronunciation of a word

    e. It helped me becoming acquainted with the English accents I wasn't familiar with

    f. I could learn how a word was written in English

    g. The text in English helped identify where words pronounced by speakers began and ended

    h. I have never used captions/subtitles to improve my proficiency in English

    i. Other (please, write down your answer)

19. Did you enjoying watching audiovisual content in English and using captions/subtitles to learn the language when your proficiency in English was lower? Why? Choose all that apply to you.

    a. Yes: it was fun to watch movies, TV series, etc. in English and read the translation in my native language

    b. Yes: I realized it was easier for me to learn new words and/or the language in general when I was watching audiovisual content in English outside the scholastic environment

c. Yes: it helped me improve my pronunciation. I could repeat in real time what I was listening to with the help of the written transcription

d. Yes: it motivated me to improve my knowledge of English in a more relaxed environment

e. No: I found it too hard to keep up with text and speech delivered at the same time - I preferred watching audiovisual content in my native language

f. No: I found too hard to make sense of the speech and read the text in English

g. No: I used to prefer studying English only at school

h. Other (please, write down your answer)

20. How frequently did you use supporting written text (captions, subtitles, etc.) as a mean to learn English **when your proficiency in English was lower**? *[one answer for each type of written supporting content, i.e. captions, intralingual subtitles, interlingual subtitles]*

    a. Never

    b. Rarely

    c. Frequently

    d. Always

21. In your opinion, how much did **captions** (written transcript of the audio) help you improve your language skills **when your knowledge of English was lower**?

    [For each ability (reading, speaking, listening, writing, and vocabulary), participants needed to specify the quantity by selecting one of the following options]:

    a. Not much

    b. Fairly

    c. Very much

    d. I didn't use captions to improve this language skill

22. In your opinion, how much did **subtitles** (same language of the audio) help you improve your language skills **when your knowledge of English was lower**?

[For each ability (reading, speaking, listening, writing, and vocabulary), participants needed to specify the quantity by selecting one of the following options]:

a. Not much

b. Fairly

c. Very much

d. I didn't use captions to improve this language skill

23. In your opinion, how much did **subtitles** (translation of the audio) help you improve your language skills **when your knowledge of English was lower**?

[For each ability (reading, speaking, listening, writing, and vocabulary), participants needed to specify the quantity by selecting one of the following options]:

a. Not much

b. Fairly

c. Very much

d. I didn't use captions to improve this language skill

24. Do you watch audiovisual content in English to improve your knowledge of the language **in the present day**?

a. Yes

b. No

25. Do you use captions/subtitles to improve your knowledge of English **in the present day**?

a. Yes

b. No

26. Why do you use captions/subtitles to learn English in the present day? Choose all that apply to you.

a. I can easily spot words I don’t know and learn them when the translation is provided in the subtitles

b. I can spot words I don't know and learn them in English

c. I can learn the meaning of a word by reading the supporting written text

d. I can learn the pronunciation of a word

e. It helps me becoming acquainted with English accents I am not familiar with

f. I can learn how a word is written in English

g. The text in English helped identifying where words pronounced by speakers begin and ends

h. I don't use captions/subtitles to improve my proficiency in English

i. Other (write down your answer)

27. Do you enjoying watching audiovisual content in English and using captions/subtitles to learn the language in the present day? Why? Choose all that apply to you.

    a. Yes: it is fun to watch movies, TV series, etc. in English and read the translation in my native language

    b. Yes: it's easier for me to learn new words and/or the language in general when I am watching audiovisual content in English outside the academic environment

    c. Yes: it helps me improve my pronunciation. I can repeat in real time what I am listening to with the help of the written transcription

    d. Yes: it motivates me to improve my knowledge of English in a more relaxed environment

    e. No: I find it too hard to keep up with text and speech delivered at the same time - I prefer watching audiovisual content in my native language

    f. No: I find too hard to make sense of the speech and read the text in English

    g. No: I prefer studying English only at the University during lectures

    h. Other (write down your answer)

28. How frequently are you using supporting written text (captions, subtitles, etc.) as a mean to learn English **in the present day**? *[one answer for each type of written supporting content, i.e. captions, intralingual subtitles, interlingual subtitles]*

    a. Never

b. Rarely

c. Frequently

d. Always

29. In your opinion, how much are **captions** (written transcript of the audio) helping you improve your language skills **in the present day**?

    [For each ability (reading, speaking, listening, writing, and vocabulary), participants needed to specify the quantity by selecting one of the following options]:

    a. Not much

    b. Fairly

    c. Very much

    d. I didn't use captions to improve this language skill

30. In your opinion, how much are **subtitles** (same language of the audio) helping you improve your language skills **in the present day**?

    [For each ability (reading, speaking, listening, writing, and vocabulary), participants needed to specify the quantity by selecting one of the following options]:

    e. Not much

    f. Fairly

    a. Very much

    b. I didn't use captions to improve this language skill

31. In your opinion, how much are **subtitles** (translation of the audio) helping you improve your language skills **in the present day**?

    [For each ability (reading, speaking, listening, writing, and vocabulary), participants needed to specify the quantity by selecting one of the following options]:

    a. Not much

    b. Fairly

    c. Very much

d. I didn't use captions to improve this language skill

32. Do you want to add something about how and when you use captions and/or subtitles? If not, just ignore this question *[open, optional question]*.

33. Do you want to add something about the characteristics of captions and/or subtitles you prefer? If not, just ignore this question *[open, optional question]*.

B. Questionnaire on the potential use of live automatic captions in educational settings and investigation on the implementation of a color-coded markup to display confidence levels

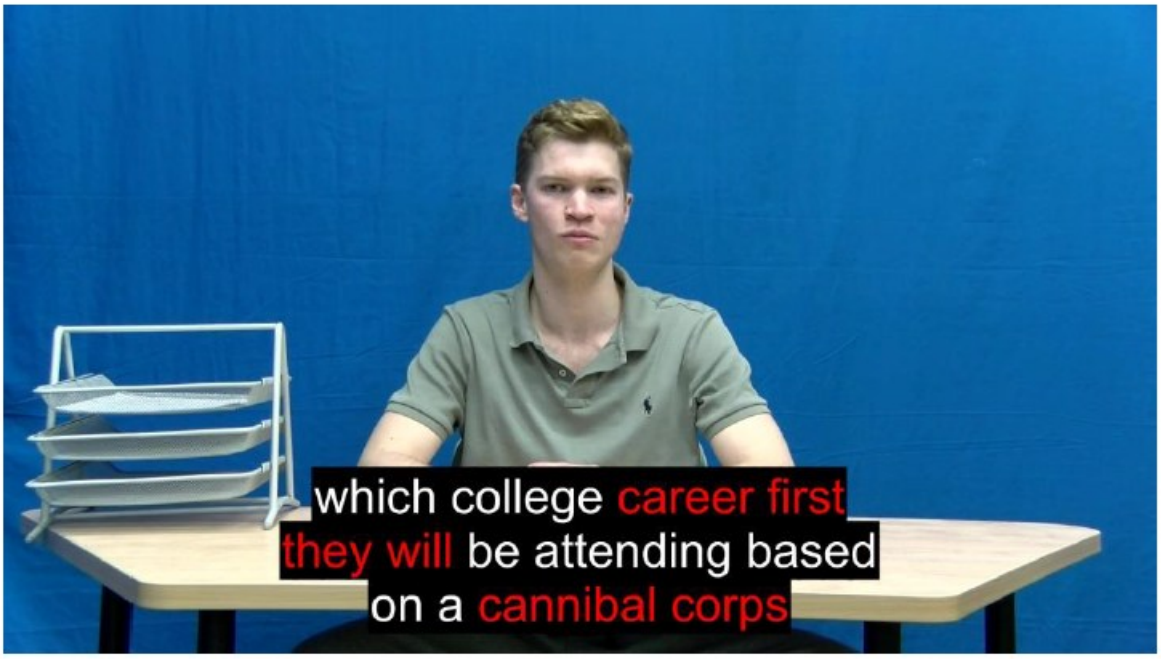
I. First version of the questionnaire

We would like to ask your opinion about using live captions in educational environments, such as during lessons at the University. Please, answer the following questions.

1. Live captions (transcription of speech in real-time, while a speaker is talking) can be generated using an automatic speech recognition (ASR) system. Would you consider using it during class if the University provides this service?
    a. Yes
    b. No
2. If live captions were available in class, which display format would you prefer?
    a. Captions appearing on two lines.
    b. Word-by-word captions appearing simultaneously with speech.
    c. Captions (transcript of spoken language) together with subtitles (translation of spoken language).
    d. I think I would be distracted by captions.
3. In your opinion, which characteristic(s) of the captions would be crucial to support comprehension in class?
    a. Number of lines on which the text is displayed
    b. Speed of appearance of the text
    c. Speech-captions alignment
    d. Accuracy of the automatic transcription
    e. Font type
    f. Font size
    g. I don’t know

4. Do you think your knowledge of the L2 would get better if live captions were provided? Why? You can select more than one answer.
    a. Yes: captions could help me improve my vocabulary knowledge
    b. Yes: captions could help me improve my pronunciation
    c. Yes: captions could help me recover words I missed/I did not understand for some reason
    d. No: captions would only be redundant, since I prefer listening to the professor/follow the sign language interpreter
    e. Yes: other reason (please, write down your answer)
    f. No: other reason (please, write down your answer)
    g. Maybe (please, write your answer down)
    h. Other (write your answer down)
5. ASR systems are not always accurate in their transcriptions due to various factors (e.g., environmental noise, low-quality microphone, etc.). Would you find live captions useful even though there were some errors in them?
    a. Yes
    b. No
    c. It depends on the quantity/quality of errors
6. Why? Please, motivate your previous answer. You can select more than one choice.
    a. Too many errors could prevent me from understanding the content of the lesson
    b. Errors could unnecessarily distract me
    c. Errors could confuse me
    d. I would ignore errors and just focus on the speaker's speech
    e. Other (write your answer down)
7. In your opinion, would it be helpful if a specific display format signaled transcription errors? Below you can see an example (Berke, 2017). The system signals users how confident it is that the transcribed words match the speaker's words based on the color of the text. The ASR

system displays words in red if it isn't sure they correspond to the speaker's words. On the other hand, the system shows the words in white if it is confident that they correspond to the speaker's words.



a. Yes

b. No

c. I’m not sure

8. Why? Please, motivate your previous answer. You can select more than one option.

a. I would find it distracting to see the different display format (i.e., change of colors) within the captions

b. It would help me to know how confident the system is with the transcription

c. Other (write your answer down)

II. Second version of the questionnaire

We would like to ask your opinion about using live captions in educational environments, such as during lessons of courses taught in English at the University. Please, answer the following questions.

1. Automatic speech recognition (ASR) is a technology that processes human speech and transcribes it into readable text.

   Do you use automatic speech recognition (ASR) technology in your day-to-day life?

   a. Yes

   b. No

2. Live captions (transcription of speech in real-time, while a speaker is talking) can be generated using an automatic speech recognition (ASR) system. Would you consider using it in class for courses taught in English if your university provided this service?

   a. Yes

   b. No

3. If live captions were available in class for courses taught in English, which display format would you prefer?

   a. Captions appearing on two lines.

   b. Word-by-word captions appearing simultaneously with speech.

   c. Captions (transcript of spoken language) together with subtitles (translation of spoken language).

   d. I think I would be distracted by captions.

4. In your opinion, which characteristic(s) of the captions would be crucial to support comprehension of lectures taught in English at the university?

   a. Number of lines on which the text is displayed

   b. Speed of appearance of the text

   c. Speech-captions alignment

   d. Accuracy of the automatic transcription

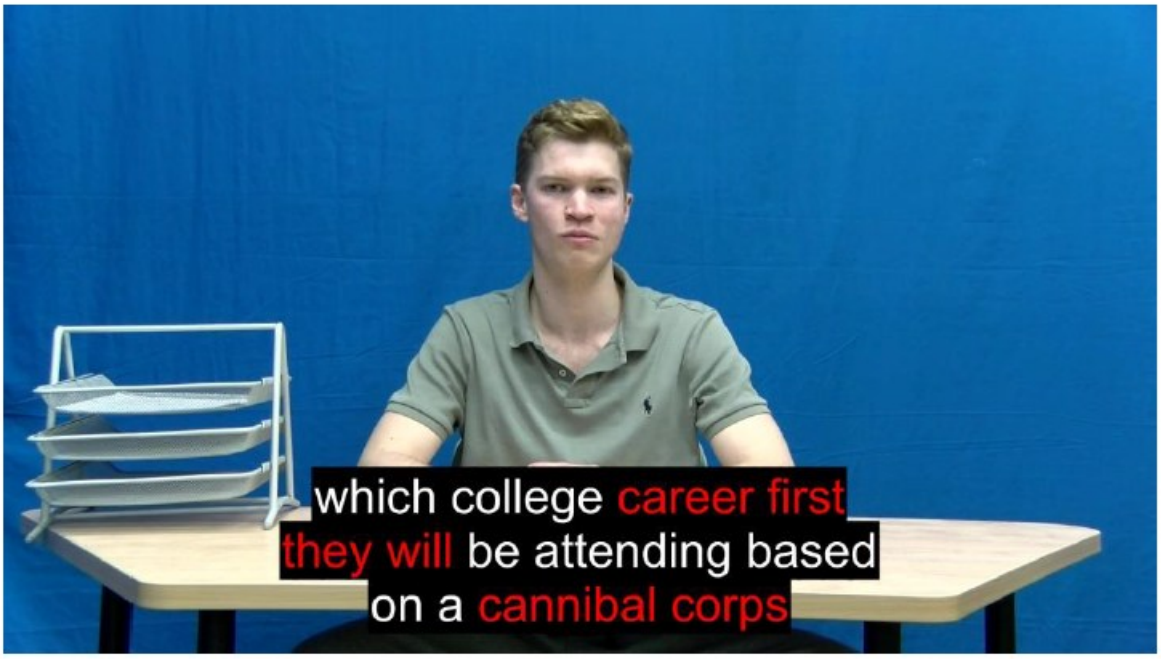
   e. Font type

   f. Font size

   g. I don’t know

5. Do you think your knowledge of the L2 would get better if live captions were provided? Why? You can select more than one answer.
    a. Yes: captions could help me improve my vocabulary knowledge
    b. Yes: captions could help me improve my pronunciation
    c. Yes: captions could help me recover words I missed/I did not understand for some reason
    d. No: captions would only be redundant, since I prefer listening to the professor/follow the sign language interpreter
    e. Yes: other reason (please, write down your answer)
    f. No: other reason (please, write down your answer)
    g. Maybe (please, write your answer down)
    h. Other (write your answer down)
6. ASR systems are not always accurate in their transcriptions due to various factors (e.g., environmental noise, low-quality microphone, etc.). Would you find live captions useful even though there were some errors in them?
    a. Yes
    b. No
    c. It depends on the quantity/quality of errors
7. Why? Please, motivate your previous answer. You can select more than one choice.
    a. Too many errors could prevent me from understanding the content of the lesson
    b. Errors could unnecessarily distract me
    c. Errors could confuse me
    d. I would ignore errors and just focus on the speaker's speech
    e. Other (write your answer down)
8. In your opinion, would it be helpful if a specific display format signaled how confident the system was of its transcription? Below you can see an example of display format (Berke, 2017). The system signals users how confident it is that the transcribed words match the

speaker's words based on the color of the text. The ASR system displays words in red if it isn't sure they correspond to the speaker's words. On the other hand, the system shows the words in white if it is confident that they correspond to the speaker's words.



a. Yes

b. No

c. I’m not sure

9. Why? Please, motivate your previous answer. You can select more than one option.

    a. I would find it distracting to see the different display format (i.e., change of colors) within the captions

    b. It would help me to know how confident the system is with the transcription

    c. Other (write your answer down)

C. Questionnaire investigating the preferences of students regarding the various color-coded markups to display confidence levels of the ASR system in the automatic captions

Answer the following questions related to the video and the characteristics of the captions you saw.

Q1. Think about the content of the clip you just watched and answer the following questions.

Did you already know something about the topic (Pidgin and Creole languages) before watching the video?

a. Yes

b. No

Q2. Did you have a hard time understanding the content of the clip?

a. Yes

b. No

c. Sometimes

Q3. Why? Could you motivate your previous answer? Choose all that apply to you. If you want to add an answer, please select the box Altro and write your answer down.

a. I was not familiar with the topic

b. I had trouble recognizing some words

c. I did not know enough vocabulary

d. It was not difficult for me

e. I had to pay close attention to the speakers' speech

f. It was hard to follow the speakers' interactions

g. I wasn't familiar with the speakers' accents

h. The audio quality of the clip was bad

i. Speakers were talking too fast - I couldn't keep up with the conversation

j. I have some knowledge of the topic discussed in the clip

k. Other

Q4. Did the captions help you understand the content of the clip?

a. Yes, the captions helped me

b. No, captions did not help me

Q5. Why? Please, motivate your previous answer. Choose all that apply to you. If you want to add an answer, please select the box Altro and write your answer down.

a. I could rely on captions to make sure I understood the content of the clip

b. I could rely on captions to understand the speakers since I was not familiar with the different accents

c. The captions had too many errors - I couldn't rely on captions to improve comprehension

d. I could rely on captions to recover words I missed for some reason

e. I could rely on captions since the quality of the audio was bad

f. Captions helped me pay more attention to the discussion

g. The errors in the captions distracted/annoyed me

h. I think my knowledge of English was too low to fully understand the content of the clip

i. I think my knowledge of English was high enough to fully understand the content of the clip without captions/subtitles

j. Captions were redundant

k. I look at speakers' faces while they speak to follow the discussion, I don't need captions

l. I prefer to listen to speakers talking

m. The font of the text (e.g., underlined, in a strike-through format, etc.) of the captions distracted me

n. I could rely on captions to recognize words while the speakers were talking

o. The format with which captions were displayed annoyed me

p. I could not concentrate on the speakers' voices, the captions kept distracting me

q. I prefer reading the text in the slides (if provided in this setting) rather than reading the captions

r. Other

Q6. It is common that captions generated by ASR systems - like the ones in the video you previously watched - contain errors in the transcription of speech. Did finding errors in the captions affect your listening experience?

a. Yes

b. No

c. I did not notice the errors in the captions

Q7. How? Please, motivate your previous answer. Choose all that apply to you. If you want to add an answer, please select the box Altro and write your answer down.

a. Errors confused me

b. Errors distracted me from listening to the speakers

c. I did not pay attention to the errors

d. I focused on listening rather than reading the captions

e. Other

Q8. Did the errors in the captions have an impact on your attention? Choose all that apply to you. If you want to add an answer, please select the box Altro and write your answer down.

a. Yes, the errors confused me

b. Yes, the errors distracted me from listening to the speakers

c. No, I didn't pay attention to the errors

d. No, I focused on listening rather than reading the captions

e. Other

Q9. Did the errors in the captions have an impact on the comprehension of the content of the video? Choose all that apply to you. If you want to add an answer, please select the box Altro and write your answer down.

a. Yes, the errors confused me

b. Yes, the errors distracted me from listening to the speakers

c. No, I didn't pay attention to the errors

d. No, I focused on listening rather than reading the captions

e. Other

Answer the following questions related to the characteristics of the captions in the video.

Q10. Think about the display format of the captions (see the image below) and say if you agree or not with the following statements[82]:

Captions were easy to read.

Captions helped me understand the content of the video.

It was easy for me to tell how accurate the captions were by looking at the display format.

I like the display format of the captions shown in the video[83].

Q11. Overall, did you find this type of display format useful to support comprehension of the content of the video?

a. Yes

b. No

Q12. Why? Please, motivate your previous answer. You can select more than one option.

a. Seeing the different display format (i.e., words underlined, etc.) within the captions distracted me

b. It helped me to know how confident the system is with the transcription

c. I didn't find it useful; it made me feel confused

d. It made me feel more confident with my listening skills

[82] Participants needed to select one option among the five available: strongly disagree, disagree, neutral, agree, strongly agree. Statements were formulated following Shiver & Wolfe (2015).

[83] A frame containing the captions shown in each condition was included in the question.

e. I couldn't focus on the text nor speech since there was too much information in the video (text, colors, audio, etc.) - I felt overwhelmed

f. Other

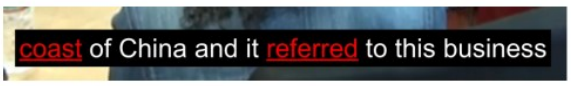
Questions 13-14 were included only in the "Classic captions" condition:

Q13. The captions shown in the clip had a standard format (i.e., it is the same one used on YouTube or other streaming services). However, current ASR systems can also display information regarding how confident the system is that the transcribed words match the speakers' words. For example, the ASR system could generate caption text in different markups (e.g., words in different colors, underlined or in a strikethrough format - see some examples below) if it isn't sure the words correspond to the speakers' words. On the other hand, the system could display the words in white if it is confident that they correspond to the speakers' words.

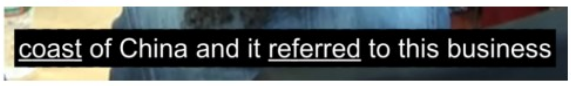
If captions included in the video, you watched had had a markup that conveyed such information, would you have found them useful?

1

2



3



a. Yes

b. No

Q14. Why? Could you motivate your previous answer? Choose all that apply to you.

a. I would have found the different display format (i.e., change of colors) within the captions distracting

b. It would have helped me to know how confident the system was with the transcription

c. I wouldn't have found it useful; it would have made me feel confused

d. I would like to try it before making my decision

e. I wouldn't be able to concentrate on the text nor speech since there would be too much information in the video (text, colors, audio, etc.)

f. It would have made me feel more confident with my listening skills

g. Other

Questions n° 15-18 were included the "OG", "V2" and "V3" questionnaires:

Q15. Think about the **markup** implemented in the captions. Did you find it useful that it signaled how confident the system was with its transcription?

a. Yes

b. No

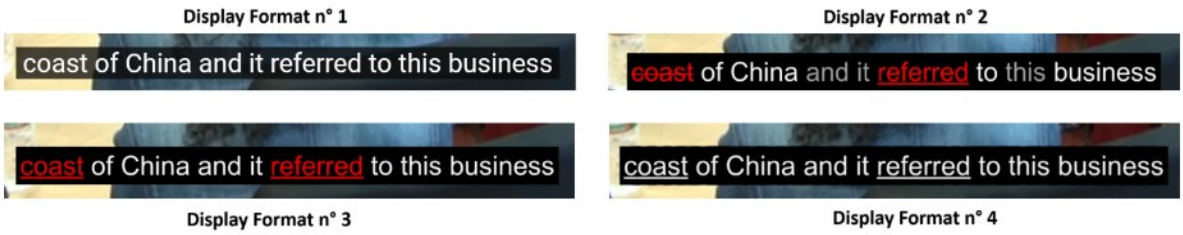
Q16. Why? Could you please briefly motivate your previous answer? Choose all that apply to you.

a. It is not relevant/useful information to me

b. The display format helps me determine if I can trust the text in the captions.

c. Other

Q17. If you were given the possibility to watch the video again and choose how captions would be displayed on screen, which format would you choose? The markup you saw in the video at the beginning of this questionnaire is n° 1/2/3/4.



a. Markup n° 1

b. Markup n° 2

c. Markup n° 3

d. Markup n° 4

e. I wouldn't use captions

f. I would use the same markup used in the video

g. Other

Q18. Do you think the markup implemented in the captions had an impact on your attention? Choose all that apply to you. If you want to add something, please select the option Altro and write your answer there.

a. Yes, the markup confused me

b. Yes, the markup distracted me from listening to the speakers

c. No, I didn't pay attention to the markup

d. No, I focused on listening rather than reading the captions

e. Other

Q19. In general, did you like the markup implemented in the captions?

a. Yes

b. No

Q20. Why? Could you motivate your previous answer? Choose all that apply to you.

a. I found it really informative

b. It distracted me

c. It helped me to understand the content of the video

d. It highlighted some words I didn't know

e. Other

Q21. Would you like to add some thoughts or opinions regarding the markup used in the captions? If not, click on *Avanti [optional, open-ended question]*.

Answer the following questions related to the potential characteristics of captions generated by automatic speech recognition (ASR) systems and the use of live captions (that is, generated by ASR systems) in educational environments, such as during lectures at the University.

Q22. Would you consider using live captions (generated by an ASR system) during courses taught in English if the University provided this service?

a. Yes

b. No

c. I don't know

Q23. Why? Could you motivate your previous answer? Choose all that apply to you.

a. I would find it distracting to see the additional text on screen

b. It would help me to know how confident the system is with the transcription

c. I would like to try it in class before making my decision

d. My listening comprehension would benefit by the presence of captions

e. Other

Q24. Would you consider using live captions (generated by an ASR system) with the display format you saw in the video during courses taught in English if the University provided this service?[84]

a. Yes

b. No

c. I don't know

Q25. Why? Could you motivate your previous answer? Choose all that apply to you.

a. I would find it distracting to see the different display formats (e.g., words underlined, etc.) within the captions

b. It would help me to know how confident the system is with the transcription

c. I would like to try it in class before making my decision

d. I would prefer to see the standard format for captions

e. Other

Q26. Would you consider using live captions (generated by an ASR system) with other types of display format during courses taught in English if the University provided this service?

[84] A screenshot of the captioned video was included in this question.

a. Yes

b. No

c. I don't know

Q27. Why? Could you motivate your previous answer? Choose all that apply to you.

a. I would find it distracting to see the different display formats (e.g., words underlined, etc.) within the captions

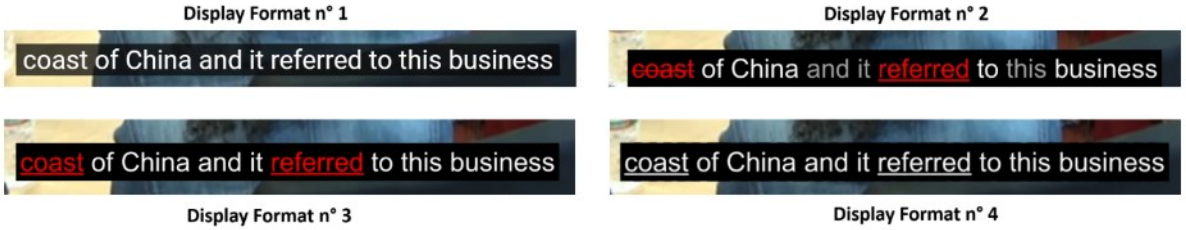
b. It would help me to know how confident the system is with the transcription

c. I would like to try it in class before making my decision

d. I would prefer to see the standard format for captions

e. Other

Q28. If your University offered live captioning during courses taught in English, which captioning style would you consider using? Choose one style/answer from the following list. The display format you saw in the video at the beginning of this questionnaire is n° 1/2/3/4.



a. Display format n° 1

b. Display format n° 2

c. Display format n° 3

d. Display format n° 4

e. I wouldn't use captions in class

f. I don't know

g. Other

Q29. Why? Could you motivate your previous answer? Choose all that apply to you.

a. I would find it distracting to see the different display formats (e.g., words underlined, etc.) within the captions

b. It would help me to know how confident the system is with the transcription

c. I would like to try it in class before making my decision

d. I would prefer to see the standard format for captions

e. Other

Q30. Do you think your knowledge of English would get better if live captions were provided in class for courses taught in English? Why? Choose all that apply to you.

a. Yes: captions could help me improve my vocabulary knowledge

b. Yes: captions could help me improve my pronunciation

c. Yes: captions could help me recover words I missed/I did not understand for some reason

d. No: captions would only be redundant; I prefer listening to the professor

e. No: too many errors make live captions useless

f. Yes: other reason (please, select the option "Altro" and write your answer there)

g. No: other reason (please, select the option "Altro" and write your answer there)

h. Maybe: other reason (please, select the option "Altro" and write your answer there)

i. Other

D. Python syntax used to calculate the Word Error Rate (WER) score

The following Python syntax was written by the author of this doctoral thesis based on the syntax provided by the community on the *Hugging Face* platform[85] (2023 version).

1. Install the relevant packages to carry out the calculations.

```
!pip install jiwer
!pip install evaluate
```

2. Load the two texts (*predictions* = HYP; *references* = REF) and compute (wer.compute) the *word error rate* (WER) metric (wer_score).

```
wer = load("wer")
predictions = ["TEXT HYP"]
references = ["TEXT REF"]
wer_score = wer.compute(predictions=predictions, references=references)
print(wer_score)
```

[85] *Evaluation metrics for ASR - Hugging Face Audio Course*. (n.d.). Retrieved January 4, 2025, from https://huggingface.co/learn/audio-course/chapter5/evaluation. Current syntax on the website slightly varies from the syntax I used to compute the *word error rate* metric in 2023 and 2024.

E. Python syntax to carry out PoS tagging

The following Python syntax was written by the author of this doctoral thesis based on the slides provided for the course "Natural Language Processing with Python" (Methods in Language Sciences Summer School – Ghent, Belgium) by instructors P. Singh and A. Tezcan (July 2023) and by consulting other sources on the Web[86]. The following Python syntax was corrected and optimized with the help of ChatGPT[87] (OpenAI, 2023).

I. English

3. Install SpaCy and large English model.

```
!pip install spacy
!python -m spacy download en_core_web_lg
```

2. Import useful libraries.

```
import spacy
import csv
from google.colab import files
```

3. Load large English model.

```
nlp = spacy.load("en_core_web_lg")
```

4. Open the REF input file saved in a .txt format and perform tokenization and PoS tagging using SpaCy. Define the output file by creating a new .csv file and print the message that you're done with the necessary steps.

```
def pos_tagging(input_filename):
    # Read input text file
    with open(input_filename, 'r') as f:
            text = f.read()

        # Tokenize and POS-tag the text using SpaCy
        doc = nlp(text)

```

[86] *Python | PoS Tagging and Lemmatization using spaCy*. (2023). GeeksforGeeks. Retrieved January 4, 2025 from https://www.geeksforgeeks.org/python-pos-tagging-and-lemmatization-using-spacy/
[87] OpenAI. (2023). *ChatGPT* (2023 version) [Large language model]. https://chat.openai.com/chat

```
        # Save token and POS-tag pairs to a list
        pos_tags = [(token.text, token.pos_) for token in doc]

        # Write token and POS-tag pairs to output text file (CSV
format)
        with open("input_filename.txt", 'w', newline = '') as f:
            writer = csv.writer(f)
            writer.writerow(['Token', 'POS Tag'])
            writer.writerows(pos_tags)

        print(f"POS-tagging completed. Output saved to
{output_filename}.")
```

5. Upload the input file.

```
    # Upload input text file
    uploaded = files.upload()
    input_filename = list(uploaded.keys())[0]
```

6. Perform PoS tagging on the text in the input file and save the results on a new .csv file. Print the message in cell 4 if the output file is correctly created.

```
    # Perform POS-tagging
    output_filename = "output_filename.csv"
    pos_tagging(input_filename)
```

7. Download the output file onto the PC.

```
    # Download output text file
    files.download(output_filename)
```

II. Italian

1. Install SpaCy and large English model.

```
!pip install spacy
!python -m spacy download it_core_news_lg
```

2. Import useful libraries.

```
3. import spacy
4. import csv
5. from google.colab import files
```

3. Load large Italian model.

```
6. nlp = spacy.load("it_core_news_lg")
```

4. Open the REF input file saved in a .txt format and perform tokenization and PoS tagging using SpaCy. Define the output file by creating a new .csv file and print the message that you're done with the necessary steps.

```
7. def pos_tagging(input_filename):
8.     # Read input text file
9.     with open(input_filename, 'r') as f:
10.             text = f.read()
11.
12.         # Tokenize and POS-tag the text using SpaCy
13.         doc = nlp(text)
14.
15.         # Save token and POS-tag pairs to a list
16.         pos_tags = [(token.text, token.pos_) for token in doc]
17.
18.         # Write token and POS-tag pairs to output text file (CSV
    format)
19.         with open("input_filename.txt", 'w', newline = '') as f:
20.             writer = csv.writer(f)
21.             writer.writerow(['Token', 'POS Tag'])
22.             writer.writerows(pos_tags)
23.
24.         print(f"POS-tagging completed. Output saved to
    {output_filename}.")
```

5. Upload the input file.

```
25.     # Upload input text file
26.     uploaded = files.upload()
27.     input_filename = list(uploaded.keys())[0]
```

6. Perform PoS tagging on the text in the input file and save the results on a new .csv file. Print the message in cell 4 if the output file is correctly created.

```
28.     # Perform POS-tagging
29.     output_filename = "output_filename.csv"
```

```
30.      pos_tagging(input_filename)
```

7. Download the output file onto the PC.

```
31.      # Download output text file
32.      files.download(output_filename)
```

F. Questions on the comprehension task

1. What are Creole languages?

   a. Creoles are the first stage of Pigeon languages
   b. Creoles are native languages derived from Pigeons
   c. Creoles are the first stage of Pidgin languages
   d. Creoles are native languages derived from Pidgins

2. What are Pidgin languages?

   a. A system of conditions created to cross business barriers
   b. A system of commissions created to cross business barriers
   c. A system of communication created for business purposes
   d. A system of conditions created for business purposes

3. What is the original meaning of "Pidgin"?

   a. A Chinese dialect
   b. Code switching in Chinese
   c. Peasants' language in Chinese
   d. Business language in Chinese

4. Who is the famous Haitian leader known to speak more than five languages?

   a. Louise Veracio
   b. Toussaint L'Ouverture
   c. Jean-Pierre Boye
   d. Philippe Guerrier

5. What characterizes Pidgin languages?

    a. They are non-native languages with a particularly rich structure
    b. They are native languages with a particularly rich structure
    c. They are native languages with a particularly reduced structure
    d. They are non-native languages with a particularly reduced structure

6. Does Haitian creole history present evidence support Bickerton's theory?

    a. Current evidence falsifies the hypothesis
    b. Current evidence is insufficient
    c. Current evidence supports the hypothesis
    d. Current evidence is not available

7. According to Bickerton, what is missing in the structure of Pidgin languages?

    a. Infixes
    b. Suffixes
    c. Prefixes
    d. Affixes

8. Following Bickerton, these languages seem:

    a. Possible, un-language-like
    b. Impossible, un-language-like
    c. Possible, language-like
    d. Impossible, language-like

9. Is it appropriate to classify Spanglish as a Pidgin?

   a. Yes, Pidgin languages and Spanglish are both a matter of code-switching
   b. No, because their sociolinguistic environment and historical background are different
   c. Yes, because they both are a mix of different non-native languages
   d. No, because Spanglish is linguistically richer than Pidgin languages

10. In the first stages of French Creole, what elements were missing?

   a. Prefixes
   b. Suffixes
   c. Infixes
   d. Affixes

Estratto per riassunto della tesi di dottorato

L'estratto (max. 1000 battute) deve essere redatto sia in lingua italiana che in lingua inglese e nella lingua straniera eventualmente indicata dal Collegio dei docenti.

L'estratto va firmato e rilegato come ultimo foglio della tesi.

Studente: Martina Pucci matricola: 989044

Dottorato: Lingue, Culture e Società Moderne e Scienze del Linguaggio, curriculum Scienze del Linguaggio

Ciclo: XXXVII

Titolo della tesi[1] : Multi-modality For All. Tecniche di sottotitolazione e trascrizione automatica in approccio human-centered

Abstract:

La tesi dottorale ha avuto come obiettivi quelli di 1) investigare le abitudini e preferenze di utilizzo dei sottotitoli generati da sistemi di riconoscimento automatico del parlato (ASR) da parte di studenti universitari parlanti di inglese L2/lingua franca per supportare la comprensione orale, 2) valutare la performance di un sistema ASR attraverso l'analisi qualitativa di un corpus di trascrizioni raccolte in contesti reali, 3) creare ed implementare nei sottotitoli degli elementi grafici (schema di colori) che comunicassero il grado di sicurezza del sistema ASR del suo output all'utente finale.

This doctoral thesis investigated 1) the preferences and opinions of university students who speak English as an L2/lingua franca on captions generated by automatic speech recognition (ASR) systems to aid speech processing and content comprehension, 2) the performance of an ASR system by evaluating its accuracy and confidence of a corpus of transcriptions collected from real-life contexts, 3) the development of graphical features (specifically, a color-coded scheme) to be implemented in the text of automatic captions to communicate to users how confident the ASR system is of its output.

Firma dello studente

VENEZIA, 24/05/2025

Martina Pucci

[1] Il titolo deve essere quello definitivo, uguale a quello che risulta stampato sulla copertina dell'elaborato consegnato.