# SUBTITLING AS MULTIMODAL REPRESENTATION:

# A CORPUS-BASED EXPERIMENTAL APPROACH

# TO TEXT-IMAGE RELATIONS

ZHUOJIA CHEN

PhD

The Hong Kong Polytechnic University

2024

The Hong Kong Polytechnic University

Department of Chinese and Bilingual Studies

# Subtitling as Multimodal Representation:

# A Corpus-based Experimental Approach to Text-image Relations

Zhuojia CHEN

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

August 2023

# CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

______________________ (Signature)

______Zhuojia CHEN______ (Name)

# ABSTRACT

Communication is intrinsically a multimodal act. In an audiovisual product, communication seldom involves mere language but incorporates a strong relation between language and image. One of the challenges in audiovisual translation (AVT) is the thin line between assisting and distracting the audience. While translators need to maximize viewers' comprehension of the linguistic content, they also need to minimize viewers' efforts in enjoying the audiovisual product as a multimodal ensemble. This project is composed of a corpus-based study and an eye-tracking experiment, attempting to explore: (a) how interlingual subtitles interact with the image in films, and (b) how the target subtitle's interactions with the image may affect viewer reception.

The corpus-based study and the experiment are underlined by two frameworks proposed by the author to theorize text-image relations. The first framework is put forward to analyze text-image relations in subtitled films. It is adapted from previous frameworks by Martinec and Salway (2005), Unsworth (2006, 2007), and Pastra (2008), with four major categories of text-image relations (4 Cs): *Concurrence* (text = image), *Complementarity* (text > image), *Condensation* (text < image), and *Contradiction* (text ≠ image). The other framework focuses on translation shifts in text-image relations specifically within the context of interlingual subtitling. This framework is formulated through a bottom-up approach, drawing on real-life subtitling cases. It comprises five major categories: *non-shifts*, *obligatory shifts*, *preferential shifts*, *strengthening shifts* (*4 Es*: *expansion*, *explicitation*, *enhancement*, and *elaboration*), and *weakening shifts* (*4 Ds*: *detachment*, *diminution*, *dilution*, and *decrement*). The first three types of shifts are believed to be language-induced, driven by the subtitler's linguistic consideration between the target and source languages. The latter two types are considered as image-induced shifts, arising from the subtitler's conscious or unconscious attention to the pictorial elements during the translation process. For a more systematic analysis, the two frameworks draw on transitivity systems from Systemic Functional Grammar (Halliday & Matthiessen, 2014) and Visual Grammar (Kress & Van Leeuwen, 2021) to delineate the basic comparative units for analyzing text-image relations, namely, the verbal and visual *participants*, *processes*, and *circumstances*.

To explore how the text interacts with the image in audiovisual products, the first main study employs a corpus-based approach. A multimodal corpus was compiled, consisting of 30 scenes sampled from 10 English films, each containing source English subtitles and target Chinese subtitles. The findings revealed that both the source and target texts were semantically more specific than the image, as the frequency of *complementarity* relations was over seven times that of *condensation*. Moreover, the meanings of the text and image tended to be closely associated, as evidenced by the substantial number of *concurrence* relations. In terms of translation shifts in text-image relations through subtitling, three patterns were observed. The target subtitles tended to (a) avoid mentioning the visual *participants* (e.g., people and things) and leave them implicit; (b) verbalize and explicate the co-occurring visual *processes* (e.g., actions and gestures); and (c) include additional linguistic *circumstances* (e.g., the manner of an action) to describe or modify the visual *processes*. The corpus findings shed light on the role of interlingual subtitling as multimodal representation, which may remove, reiterate, or reinforce the visual information in the multimodal film narrative.

To investigate the potential impact of the translation shifts observed in the corpus, the other main study of the project adopts an experimental approach. Drawing on eye-tracking data, comprehension tests, perception questionnaires, and semi-structured interviews, the experiment assigned 82 participants to either a Control Group or an Experimental Group and investigated the impact of translation shifts in text-image relations on the participants' visual attention, comprehension, perception of subtitle quality, and preferences for subtitling methods. The results showed that explicating the visual actions in the target subtitles induced significantly better comprehension and shorter gaze time to the subtitle-related visual information. Moreover, the viewers appeared to perceive lower subtitle quality when target subtitles contained less referential information related to the visual entities. However, this did not necessarily hinder the viewers' comprehension or affect their distribution of visual attention. In terms of the viewers' preferences for subtitling methods, both groups strongly favored the method of providing more descriptive information to modify the visual kinesics, which helped reinforce the traits or emotions of the portrayed character. The findings from the experiment highlight the potential influence of subtitlers in guiding, or even manipulating, the cognitive attention of target viewers towards the audiovisual content.

The research contributions of this thesis are threefold. Theoretically, two frameworks have been proposed to systematically (re)conceptualize subtitled products as multimodal ensembles by highlighting the text-image interplay, thus further moving AVT research in the direction of multimodality. Methodologically, this thesis represents one of the few attempts to combine the corpus-based approach and the experimental approach in a single project, which has demonstrated the feasibility of using multiple methods in empirical AVT studies. For translation praxis, the findings from the thesis can support audiovisual translators or trainees in making an informed decision when tackling non-verbal meaning-making elements in audiovisual products, so as to achieve a more comprehensible and multimodally immersive translation.

# ACKNOWLEDGEMENTS

The completion of this thesis has been a miraculous achievement for me. When I first embarked on my educational journey 23 years ago, at the age of seven, I could never have imagined that one day I would have the opportunity to pursue a PhD degree. My earliest memories of my education date back to when my mother would hold my hand and walk me to preschool in the mornings. Today, I still recall the peaceful afternoons spent at home with my mother, while my father was working in a distant city and my older siblings were attending school. During the afternoons, usually with a shard of sunlight seeping through the living room window, my mother would sit patiently beside me as I worked on my preschool homework. The homework was a no-brainer, but I would sometimes pretend to struggle and ask for her assistance, hoping that she would recognize my need for her presence and that she would not consider it a waste of her time to be with me. Now, approaching the age of 30, I wish to tell her that I have always been grateful for having her company and care in my life. I hope this thesis can serve as a humble testament to the fact that the time she dedicated to her little boy was never in vain.

The writing of this thesis owes a great deal to my chief supervisor, Dr. Zhiwei Wu. Throughout the entire process, he actively participated and provided guidance from the initial stages of topic selection and data coding to the final stages of data analysis and result interpretation. His meticulous attention to detail was awe-inspiring. His timely and invaluable feedback played a crucial role in refining my work to a better shape. His unwavering support and wealth of academic resources made me feel secure throughout the program, enabling me to push beyond my limits. Above all, he led by example and exemplified what it means to be a dedicated researcher. I am immensely grateful for his mentorship, kindness, and inspiration throughout this arduous journey.

My profound gratitude extends to my co-supervisor, Professor Dechao Li. He has been a constant source of encouragement not only for me but also for my fellow colleagues. His infectious positivity and cheerful energy have transformed our office, AG518, into a warm and supportive family. Through our weekly book club, he has motivated us to delve into extensive reading with a critical mindset, fostering discussions and intellectual growth. The impact of his guidance and uplifting spirit is immeasurable, and I am profoundly grateful for his presence in my academic journey.

I would also like to express my deepest appreciation for the guidance provided by Dr. Olli Philippe Lautenbacher, my academic supervisor during my six-month visit to the University of Helsinki in Finland. It was a fantastic opportunity to meet him and to share a part of my academic journey with him. Our regular meetings at his office to discuss my PhD project were enlightening, thought-provoking, and at times mind-blowing. He significantly expanded my understanding of the complexities of multimodal analysis. His critical insights have helped improve the framework of text-image relations in this thesis and propelled me to develop the other framework for analyzing translation shifts in these relations. Additionally, I was amazed by his diverse range of hobbies outside of academia, which served as a valuable example of how to achieve a work-life balance as a researcher.

The writing of this thesis has also been indebted to other esteemed scholars. I would like to extend my sincere gratitude to Professor Agnieszka Szarkowska and Professor Meifang Zhang, the external examiners of this thesis. Their astute insights and suggestions have been instrumental in improving the quality of this work. The encouraging words that Professor Szarkowska generously gave in the review report have bolstered my confidence in the importance of my research. For the oral examination, I was fortunate to have Dr. Sing Bik Cindy Ngai as the Chair of the Board of Examiners. Her kind support greatly smoothed the arrangement of the viva and helped alleviate my stress. My sincere thanks go to Dr. William Dezheng Feng and Dr. Yu-Yin Hsu for their invaluable feedback on my thesis proposal. Their expertise has helped shape the direction of my research. In particular, I am grateful to William for his inspiring course, New Media, which expanded my understanding of multimodality studies. I also drew inspiration from the talk delivered by Professor Jan-Luis Kruger at the 8th International Conference on Cognitive Research in Translation and Interpreting. It was a true honor to receive his invaluable comments on my short presentation about this project in the city of Chongqing. I would also like to express my gratitude to Dr. Henri Satokangas from the

University of Helsinki for his feedback on the written text of my corpus-based study at the Nordic Night Seminar. His fresh perspectives and the readings he shared with me have shed new light on the thesis.

I would like to express my gratitude to Raymond from the IT department at PolyU for his assistance in guiding me through the intricacies of the R language. I would also like to extend my heartfelt thanks to Albert, the language lab manager, for his prompt technical support at the eye-tracking lab. Furthermore, I am indebted to Anqi, a peer research student from the School of Optometry at PolyU, who generously shared her hands-on experience and helped me overcome some technical challenges related to eye tracking.

To my peer friends in Hong Kong, I am truly blessed to have Weilong as my fabulous sports partner and like-minded companion. Hanging out with him and letting off steam on the tennis court with other friends (especially Qingxiang) in the Sunday afternoons have etched unforgettable memories of my leisure time in Hong Kong. It is remarkable how our paths crossed due to the global pandemic. I first met Weilong as my roommate during the mandatory 14-day quarantine before starting my PhD program at PolyU. Befriending him turned out to be the silver lining during the challenging years of the pandemic. My special regards extend to my beloved friends Kangqun, Xifan, Zhuohui, and Dongxu, who have been kind enough to lend a listening ear and offer me emotional support during times of academic stress. They have been the incredible online source of joy during the three years of border restriction between Hong Kong and mainland China due to the pandemic. To my dear colleagues and friends at office AG518, I would like to express my sincere gratitude. Their peer support has been precious, greatly inspiring me in academics and lightening me in daily life. I am particularly grateful for Weixin’s kind help during my preparation for the oral examination.

My profoundly heartfelt thanks, undoubtedly, go to my parents, my sister, my brother, and my extended family. They have been my safe haven, always there to provide unconditional love and emotional support throughout every stage of my life. I am eternally grateful for their presence and the comfort they bring to my life.

Looking back on my educational journey, I am deeply grateful to the teachers and professors who have played a significant role in nurturing my love for writing in both Chinese and English. I would like to express my sincere appreciation to the following individuals: Professor Jichun Liu (my exceptional supervisor for my master's degree); Paul Briody (I truly miss his humor and his English literature course at I-Shou University. I cannot thank him enough for the boost of confidence he instilled in me regarding my writing abilities); Professor Jing Yang (I was honored to have her generous encouraging words on my final essay for her course on neurolinguistics at Guangwai); Chunmei Pan (I learned so much about English writing from her course on communicative English at Guangwai); Dongqing Wang (His kind words on my essays for the English writing course at Guangwai motivated me to push forward with my writing); James Campion (I love his detailed feedbacks on my essays for the English writing course at Guangwai); Kangsheng Zhang (my Chinese teacher at senior high school who made me believe that I may have a bit of talent for writing); Yuqiong Li (a great Chinese teacher at my junior high school who gave me the freedom and confidence to write what I felt); Shangqing Li (my Chinese teacher in third grade who provided me with early enlightenment about the essence of writing); Xiaobo Wei (my first-grade Chinese teacher who cultivated my early love for the Chinese language); and last but not the least, Qiaolin Liang (my first English teacher in first grade who laid the foundation for my English learning. I still vividly remember the first two English words I learned in her class: Kate and cake).

Returning to the thesis itself, this is a study on audiovisual translation, and more specifically, subtitling. My interest in audiovisual translation might be traced back to my internship seven years ago at NetEase company, where I worked as a subtitler and later as a translation proofreader. During that time, I was constantly challenged by the craft (or art) of subtitling, yet I was also drawn to its nature of conciseness. This tallies with the writing principles outlined in William Strunk Jr.'s influential work, *The Elements of Style*, which greatly influenced my notion of English writing. As the literary giant William Shakespeare once wrote, "brevity is the soul of wit." I believe that the brevity in subtitles also manifests the witty soul in the subtitler. Hopefully, as a study on subtitling, this thesis may also offer the potential readers a glimpse, if there is any, into the witty soul of the author.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AOI | Area of Interest |
| AVT | Audiovisual Translation |
| CG | Control Group |
| CGP | Control Group Participant |
| DT | Dwell Time |
| DT% | Dwell Time Percentage |
| EXG | Experimental Group |
| EXGP | Experimental Group Participant |
| FFLIM | First Fixation Latency to the Primary Image AOI |
| MFD | Mean Fixation Duration |
| MILF | Multimodal-integrated Language Framework |
| RQ | Research Question |
| SFG | Systemic Functional Grammar |
| SL | Source Language |
| ST | Source Text |
| TL | Target Language |
| TT | Target Text |
| VG | Visual Grammar |

# GLOSSARY OF KEY TERMS

**Area of Interest (AOI)** the spatial area where eye-movement data are extracted for further analysis

**Circumstance** one of the transitivity components in Systemic Functional Grammar or Visual Grammar, referring to the information such as time and place about a given action or a state of being

**Complementarity** a type of text-image relations, in which the text further modifies or explicates what is shown in the image (text > image)

**Concurrence** a type of text-image relations, where the text almost equally refers to what is presented in the image (text = image)

**Condensation** a type of text-image relations, where the information encoded in the text is less specific than that in the image (text < image)

**Contradiction** a type of text-image relations, in which the text shows the opposing meaning of the image (text ≠ image)

**Decrement** a subtype of weakening shifts (4 Ds) in text-image relation, in which the target subtitle changes the original text-image relation of *complementarity* into a *condensation* relation

**Detachment** a subtype of weakening shifts (4 Ds) in text-image relation, referring to the omission of an existing text-image relation from the source text

**Dilution** a subtype of weakening shifts (4 Ds) in text-image relation, in which the target subtitle alters the original *complementarity* relation into a *concurrence* relation

**Diminution** a subtype of weakening shifts (4 Ds) in text-image relation, where the target subtitle alters the original *concurrence* relation into a *condensation* relation

**Dwell Time** the total duration of all fixation and saccades within an AOI

**Dwell Time Percentage** the proportion of the dwell time on the subtitle or image AOI to the dwell time on the global AOI

**Elaboration** a subtype of strengthening shifts (4 Es) in text-image relation, where the target subtitle changes the original *concurrence* relation into a *complementarity* relation

**Enhancement** a subtype of strengthening shifts (4 Es) in text-image relation, in which the target subtitle transforms the *concurrence* relation into a *complementarity* relation

**Expansion** a subtype of strengthening shifts (4 Es) in text-image relation, pertaining to the addition of a new text-image relation in the target subtitle that is not present in the source text

**Explicitation** a subtype of strengthening shifts (4 Es) in text-image relation, in which the target subtitle changes the original *condensation* relation into a *concurrence* relation

**First Fixation Latency** the time elapsed before the first fixation on an AOI

**Fixation** a period during which the eyes are relatively motionless and the visual gaze remains at a specific location

**Global AOI** the area for eye-tracking data analysis, which encompasses the entire visual frame

**Mean Fixation Duration** the average duration of each individual fixation within an AOI

**Narrative** a story or a series of events represented by textual and/or visual elements

**Non-Shift** a type of translation shifts in text-image relation, in which no shift takes place between the source and the target subtitles

**Obligatory Shift** a language-induced type of translation shifts in text-image relation, caused by inherent differences between the source and the target language systems

**Participant** one of the transitivity components in Systemic Functional Grammar or Visual Grammar, such as people and things, either concrete or abstract

| | |
|---|---|
| **Preferential Shift** | a language-induced type of translation shifts in text-image relation, caused by the stylistic preferences between the source and the target language systems |
| **Primary Image AOI** | an area for eye-tracking data analysis, where the subtitle-related visual objects are presented |
| **Process** | one of the transitivity components in Systemic Functional Grammar or Visual Grammar, such as actions or the state of being of a participant |
| **Saccade** | the rapid eye movement between two consecutive fixations |
| **Secondary Image AOI** | an area for eye-tracking data analysis, covering the area apart from the subtitle AOI and the primary image AOI |
| **Strengthening Shift** | an image-induced type of translation shifts in text-image relation, by adding more specific or explicit information in the target subtitles in relation with the image |
| **Subtitle AOI** | an area for eye-tracking data analysis, where the subtitle texts are presented |
| **Text-Image Relation** | the interaction between the subtitle and the image in terms of their representational meanings |
| **Transitivity** | a concept in Systemic Functional Grammar or Visual Grammar, referring to how meaning is represented through three major components (participant, process, and circumstance) |
| **Weakening Shift** | an image-induced type of translation shifts in text-image relation, by making the information in the target subtitles more implicit or less specific compared with the image |

# CHAPTER 1  INTRODUCTION

## 1.1 Background

One of the challenges of audiovisual translation (AVT) is that the translator constantly walks a thin line between assisting and distracting the audience. While the translator needs to maximize the audience's comprehension of the linguistic content, they also need to minimize the audience's efforts in enjoying the audiovisual product as a multimodal ensemble, given that the audience "often have to be a reader, listener and viewer at the same time" (Taylor, 2012, p. 18). The multimodality of language in communication is a fundamental characteristic that has been present in all aspects of human lives (Matthiessen, 2007). Communication seldom involves mere language but other non-verbal meaning-making modalities. When it comes to the field of Translation Studies, particularly AVT, the meaning potential of an audiovisual text goes far beyond the linguistic content and thus its translation should pay attention to the various semiotic modalities that interact with the language (Taylor, 2003). For instance, an audiovisual product incorporates a strong relation between language and image. When viewing a film, our viewing experience can by no means be the same if we only read (or listen to) the (translated) language without looking at the images on the screen or vice versa. Our ultimate viewing experience primarily arises from the synergy of language and image, which are intertwined to create a coherent narrative. It is in this respect that Díaz Cintas and Remael (2021) recently called for "greater scholarly attention to the interplay between dialogue and the rest of semiotic layers that configure the audiovisual production", which "can only be a positive, if challenging, development for the discipline" of AVT (p. 3).

The increasing attention to non-verbal semiotic modalities in Translation Studies has been decentralizing the role of language in conveying meaning (e.g., Chen & Wang, 2019; Desilla, 2014; Huang & Wang, 2023). In the study that focused on implicit messages in films, Desilla (2014) found that the audience consistently processed non-verbal information to understand the implied meanings in a film. Similarly, the research by Chen & Wang (2019) suggested that integrating semiotic information into target subtitles could greatly improve semiotic cohesion in subtitled films. A more recent experiment conducted by Huang & Wang (2023) further showed the positive impact of non-verbal elements on the subtitling process. It was found that non-verbal input in films could facilitate translators' understanding of subtitles and significantly reduced their cognitive load on subtitling. These findings have extensively indicated how non-verbal information can function as a scaffold for viewers to make better sense of the audiovisual content.

Although previous literature has asserted the importance of non-verbal elements in subtitling, few studies have systematically looked further into the mechanism of textual and non-verbal interactions in AVT. Some critical questions have not yet been fully answered, such as "What non-verbal elements matter most in AVT?", "How frequently are the spoken (original or translated) texts intersected by other non-verbal resources in an audiovisual product?", and "to what extent may the relation between the text and the non-verbal elements be altered through AVT and thus affect the audience's viewing experience?" To address these issues requires a more systematic analysis of authentic AVT materials and real-life audience responses, for which corpus-based and experimental approaches appear to be favorable.

### 1.2 Objectives and research questions

The present research project synthesizes research insights from AVT, systemic functional linguistics, multimodal discourse analysis (text-image relations), multimodal corpora, and experimental eye-tracking studies on audience reception. It will examine the interrelations among text-image relations in films, subtitling strategies, and audience reception. The purpose of the study is manifold and it seeks to:

- Theorize evidence-based frameworks for conceptualizing text-image relations in subtitled audiovisual products;
- Examine what and how text-image relations in the audiovisual content are altered through interlingual subtitling;

- Ascertain the extent to which translation shifts of text-image relations in interlingual subtitling affect target viewers' comprehension, distribution of visual attention, and perception of translation quality.

To address previous research gaps, the research objectives will be guided by the following research questions (RQs):

RQ1: What are the possible text-image relations in subtitled audiovisual products?

RQ2: To what extent are text-image relations preserved or altered in interlingual subtitling?

RQ3: How do translation shifts in text-image relations through interlingual subtitling affect viewers' distribution of visual attention, comprehension, and perception of translation quality?

## 1.3 Methodological overview

This project is one of the few attempts in Translation Studies to combine the corpus-based approach and the experimental approach (e.g., Neumann et al., 2022). Figure 1.1 presents the overall methodological design of the research.

To answer RQ1 and RQ2, a self-built multimodal corpus is compiled, which has been proved to be a useful tool for AVT studies (Soffritti, 2019). Translated subtitles are more than just the written counterpart of the spoken mode in a foreign film (Gambier, 2006, Taylor, 2016; Ramos Pinto, 2018). It is often necessary for the audiovisual translator to take into account the non-verbal information, who is more as "a communicator addressing the receptor language audience" (Gutt, 2000, p. 199). In this respect, a multimodal corpus is built for the analysis of text-image relations in ten subtitled films. Different types of text-image relations and translation shifts in the films, based on the two theoretical frameworks proposed in this project, are manually annotated in the video annotation software ELAN (Wittenburg et al., 2006). Then, a descriptive approach is applied to analyze the annotations, examining which types of verbal and visual interactions are more commonly found in films and which types of text-image relations are shifted through interlingual subtitling.

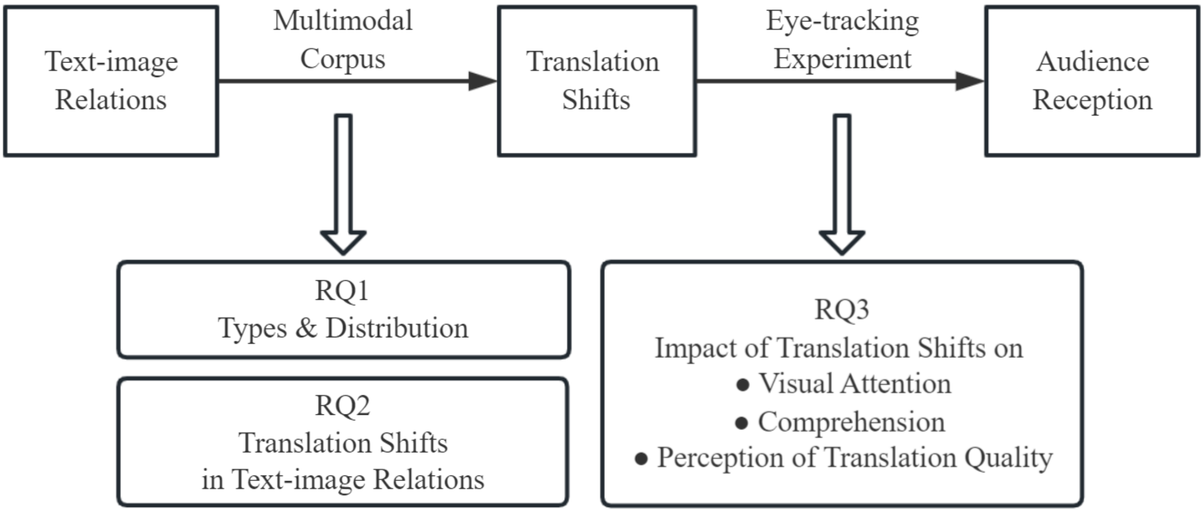


**Figure 1.1** Methodological overview of the study

Based on the corpus-based findings, an eye tracking experiment on audience reception is carried out to answer RQ3. The experimental data can explain the impact of the patterns observed in the multimodal corpus. As Lautenbacher (2012) once remarked, "translation can alter the semiotic reading of a scene" by changing the original text-image link in films (p. 151). Thus, it is reasonable to investigate how the change of text-image relations through subtitling may affect the audience's viewing behavior and experience. The independent variable in the experiment is the version of subtitles (no translation shifts in text-image relations vs. translation shifts in text-image relations). The main dependent variables examined in this study are the participants' visual attention, comprehension performance, and perception of subtitle quality, which will be assessed by eye-tracking data, comprehension tests, and perception questionnaires. The participant's visual attention is examined by eye-tracking measures, including *dwell time*, *dwell time percentage*, *mean fixation duration*, and *first fixation latency* (Black, 2022; Holmqvist et al., 2011; Kruger et al., 2014). The comprehension performance is tested by open-ended questions and viewers' responses are quantified from a 2-point comprehension scale adapted from Desilla (2014). The viewers' perception of subtitle quality is measured with close-ended questions in the form of six-point Likert scale, which is based on Künzli's (2021) CIA model of subtitle quality, concerning the target subtitle's *correspondence*, *intelligibility*, and *linguistic authenticity*. In addition, viewers' preferences for different subtitling methods concerning text-image interplay are investigated in the experiment. From semi-structured interviews, viewers' individual preferences are evaluated and serve as supplementary resources to both the experimental results and the corpus findings in this project.

The combined use of corpora and experimental methods in this research project can contribute "to a holistic understanding of translation" (Neumann et al., 2022, p. 99).

### 1.4 Structure of the thesis

This thesis comprises five major chapters. Chapter 2 presents the literature review to contextualize the research project. The first part of the section touches upon previous theories of subtitling. As this study synthesizes research insights from AVT, multimodality, and reception studies, the review of subtitling theories focuses on research with an audience-oriented or multimodal perspective. The other part of the review section examines the research methods applied in previous studies, with the emphasis on corpus-based and experimental approaches. Current research gaps will then be critically discussed.

Chapter 3 introduces two theoretical frameworks for the thesis. The chapter first examines some previous theoretical attempts by Martinec and Salway (2005), Unsworth (2006, 2007), and Pastra (2008) to sort out the interwoven relations between text and image. Then, it provides an account of structures in text and image that constitute the multimodal representation of a narrative, exploring the basic analytical units in text-image relations. The discussion draws on theoretical insights from Halliday's (Halliday & Matthiessen, 2014) systemic functional grammar and Kress and Van Leeuwen's (2021) visual grammar. In the third section of the chapter (Section 3.3), a new theoretical framework of text-image relations is proposed to account for the dynamic interplay between subtitles and image in AVT. In the last part of the chapter (Section 3.4), the other framework is proposed, which focuses on the translation shifts in text-image relations through interlingual subtitling. The two theoretical frameworks serve as the basis for the corpus analysis.

Chapter 4 presents a corpus-based study, starting with an introduction to methodological issues such as the selection of corpus materials, the corpus size, and the annotation scheme for the corpus analysis. The chapter then proceeds to present both quantitative and qualitative results of the corpus study, followed by a discussion of the findings.

Chapter 5 delves into the eye-tracking experimental study. The chapter begins with a brief description of the design of the pilot experiment and the insights gained from it. Then, the design of the main experiment is presented, covering stimulus materials, participants, treatment conditions, variables, instruments, data processing, and procedure. The chapter concludes with a quantitative and qualitative analysis, followed by a discussion.

Chapter 6 serves as the conclusive section, summarizing the major findings from both the corpus-based study and the eye-tracking experiment. It also highlights the potential contributions of the project to Translation Studies, addresses its limitations, and suggests future avenues for empirical research on subtitling from a multimodal perspective and on the reception of subtitled audiovisual products.

# CHAPTER 2  LITERATURE REVIEW

The literature review chapter consists of two major parts, critically examining previous conceptions of and methodological approaches to the studies of subtitling. The first part (Section 2.1) concerns previous conceptions of subtitling. It starts with a discussion on the language-centered research on subtitling, pivoting around the traditional text-based subtitling strategies proposed by AVT researchers. It later looks into the audience-oriented research and scrutinizes the effects of subtitling from the perspective of real-life viewers. The discussion then moves on to the multimodal consideration for subtitling, which has been attracting growing interests in AVT research. The section ends with a review of previous conceptualization of translation shifts in subtitling studies.

The second part of the review (Section 2.2) looks into the methods applied in previous subtitling research. It first outlines some findings of subtitling research based on text (monomodal) corpora and multimodal corpora. Then, it reviews empirical studies on subtitling that use eye-tracking technology. The review section serves to sort out the nature of subtitling and, in particular, to highlight how a multimodal perspective and empirical evidence can contribute to the understanding of subtitling and thus pave the path for the theoretical frameworks and methodologies applied in this study.

## 2.1 Previous conceptions of subtitling and the multimodal consideration

### *2.1.1 The language-centered research on subtitling*

Traditionally, subtitling is deemed as a condensed written translation of the spoken source (Nedergaard-Larsen, 1993). It is not possible, and seldom required, to give a complete literal rendition of the original in subtitling. The spatial and temporal constraints on subtitling oblige the translator to make the subtitles as concise as possible, as they are commonly fixed at the bottom of the screen with a maximum of two lines and drop out of sight in seconds. In a practical sense, shorter subtitles with omission seem more likely to be the case. Díaz Cintas (2003, cited in Díaz Cintas & Remael, 2021, p. 149) found a reduction of 40% English content in the Spanish subtitles in Woody Allen's *Manhattan Murder Mystery*. Han & Wang (2014), who examined the subtitling of swearwords in eight episodes of the Australian reality TV series *The Family*, found an omission of 31% of swearwords from English into Chinese. However, the strategy of omission may sometimes cause adverse impacts. For instance, the reduction of information in the English subtitles could prevent British viewers from understanding the original cultural values in Chinese films (Chen, 2018). Although a longer subtitle may distract viewers' attention from the visual content on the screen, more textual information in the target subtitles has been proved beneficial by some researchers (Caffrey, 2008; Zheng & Xie, 2018).

When it comes to the subtitling strategies of the source text, over the past three decades, there has been an extensive number of terms proposed by different researchers in regard to adding, omitting, or preserving textual information in subtitling. The strategy of preserving is the least disputable means by which the subtitler exactly reproduces or even merely transfers the original content in the target subtitles. This strategy is named "transfer" and "imitation" by Gottlieb (1992, p. 166), overlapping with what Nedergaard-Larsen (1993, p. 219) calls "transfer/loan" and "direct translation" and what Díaz Cintas and Remael (2021, p. 207) call "loan" and "calque". Pedersen (2011, p. 77) similarly names this strategy "retention" and "direct translation".

Merely omitting some elements from the source language, on the other hand, is a way the subtitler can resort to in order to tackle spatial and temporal restraints. This strategy has different names, including "omission" by many researchers, "resignation, deletion, decimation, condensation, dislocation" by (Gottlieb, 1992, p. 166), "reduction" by Georgakopoulou (2009, p. 30), "reduction-based explicitation" by Perego (2003, p. 82), "judicious reduction" by Taylor (2003, p. 204), and "condensation/reformulation" by Díaz Cintas and Remael (2021, p. 151). Although these terms are sometimes defined a bit differently and used for various specific problems in interlingual subtitling, they still refer to the same type of subtitling strategy.

When it comes to the strategy of adding, the terms for it are more distinct from one another. Gottlieb (1992, p. 166) puts forth the word “expansion” regarding the added information in the target subtitles for facilitating viewers’ understanding. And for the same reference, Pedersen (2011, p. 79) applies the name “specification” and Perego (2003, p. 76-79) turns to the terms “cultural explicitation” and “channel-based explicitation”.

The subtitling strategies proposed in previous research are problem-oriented, serving to tackle specific obstacles in subtitling such as culture-specific references and humor. However, most of the strategies pivot around the language problems in the source content, paying little attention to other non-verbal elements in the audiovisual products. Gambier (2006) once criticized the language-centered study of subtitling, stating that it was merely paradoxical to investigate subtitling with no regard for the visual content given that researchers had acknowledged the interplay between the verbal and the visual. Moreover, few researchers have tested their strategies on real viewers. This is where audience-oriented research comes in, which will be reviewed in the next section.

### *2.1.2 The audience-oriented research on subtitling*

Reception studies have been gathering growing momentum in AVT research (Orrego-Carmona, 2019). As Božović (2022) recently put forwards, knowing viewers’ expectations and needs “can improve the positive reception, placement, and usability of the [audiovisual] product” (p. 2). For the measures of audience reception, Gambier (2006, 2018) once formulated his top-down 3-Rs model that distinguished three types of reception of AVT: *response* on the behavioral level (e.g., viewers’ spontaneous gazing patterns), *reaction* on the cognitive level (e.g., understanding of the content), and *repercussion* on the attitudinal and cultural sense (e.g., viewing habits). Although not all reception studies on subtitling investigate viewers’ reception from these aspects, the majority of reception variables examined in previous research fall into these tripartite domains.

As for the research methods, major findings of AVT reception studies have been mainly explored through experiments (Orrego-Carmona, 2019). In the recent systematic review of experimental research in AVT, Wu and Chen (2021) categorized three themes in AVT reception studies, i.e., *pedagogy*, *process*, and *product*, among which *pedagogy* and *product* are directly related to the audience.

The *pedagogy* theme concerns the impact of subtitles on language learning. There has been an extensive body of literature recognizing the value of interlingual subtitling for language acquisition. Researchers have proved that subtitles are beneficial for learning second or foreign languages by promoting viewers' acquisition of vocabulary (Bisson et al. 2014; Danan 1992; Marzban & Zamanian, 2015), sentence structure (d'Ydewalle & Van de Poel, 1999), and grammar rules (Van Lommel et al., 2006) as well as improving their listening skills (Ghoneam 2015), writing skills (Talaván & Rodríguez-Arancón, 2014), and pragmatic awareness in written productions (Lertola & Mariotti, 2017).

*Product* is the most investigated theme in the literature (Wu & Chen, 2021), which alludes to the impact of the subtitle per se. One of the most measured impacts of subtitling in previous studies has been the audience's comprehension, which corresponds to *reaction* in Gambier's (2006, 2018) model. Other reception factors that have been found to be affected by subtitling include perception of subtitle quality (Kuscu-Ozbudak, 2022), enjoyment (Szarkowska & Gerber-Morón, 2018), cognitive load (Perego et al., 2016), and visual attention (see Section 2.2.2 for a more detailed review of eye tracking studies).

Despite the significant development of reception studies on subtitling over the years, most experiments have been restricted to one single mode of audiovisual content, focusing on the pure textual information of the dialogues. By addressing the lack of studies on multimodal interplay in audiovisual content, Desilla (2014) conducted a reception study to investigate the comprehensibility of messages that are visually implied in films. The results showed that the viewers, either native or foreign to the language of the subtitled film, did not always understand the verbal or non-verbal implicit meaning in the way the filmmakers intended them to. The findings suggested the potential of subtitling to have a closer interaction with the visual information in audiovisual content.

A more recent theorization of viewers' reception and cognitive processing of subtitled products from a multimodal perspective is the multimodal-integrated language framework

(MILF) proposed by Jan-Louis Kruger and his colleagues (Kruger & Liao, 2022; Liao et al., 2021; Liao et al., 2022). According to MILF, viewers usually construct a mental representation to comprehend audiovisual texts. This mental representation integrates information from two distinct processing systems: the auditory system and the visual system. Through the two systems, viewers perceive meanings from the verbal (written or spoken words), visual, and/or aural modes, and ultimately form their working memory of the multimodal narrative. It was found that verbal-visual redundancy, by facilitating viewers' "parallel processing" of "multiple representations" (Kruger & Liao, 2022, p. 30), could benefit their comprehension of audiovisual content (Liao et al., 2021). The MILF model provides a clear and systematic roadmap for unraveling the black box of viewers' cognitive processing of audiovisual content with subtitles.

Given the current scarcity of empirical evidence on the impact of different subtitle-image relations on viewers' reception, it appears worthwhile to assess how the subtitle's interplay with the visual mode may affect audience reception.

### *2.1.3 The multimodal consideration for subtitling*

Research into multimodality is not a new phenomenon. According to Kress and Van Leeuwen (2001, p. 20), multimodality is "the use of several semiotic modes in the design of a semiotic product or event". Multimodal texts can be defined as "texts which combine and integrate the meaning-making resources of more than one semiotic modality – for example, language, gesture, movement, visual images, sound and so on – in order to produce a text-specific meaning" (Thibault, 2000, p. 311).

Notwithstanding the current marginal status that translation studies occupy in the field of multimodality (Taylor, 2013; Dicerto, 2018), recent years have witnessed a growing body of research looking into AVT from a multimodal perspective. Early studies involving multimodality and subtitling were mainly prescriptive. For example, Taylor (2003, 2004) and Chuang (2006) proposed possible subtitling strategies based on multimodal theories. Chaume (2004) presented the signifying codes in film language (e.g., the linguistic code, the paralinguistic code, and the iconographic code) that primarily affect subtitling. Caffrey (2008) examined pop-up notes that can appear in any position on the screen to additionally explain items in other semiotic channels (e.g., the images or sounds). Although researchers during this period did not aim to pin down a systematic framework for multimodal analysis of translation, they did sensitize subtitlers to the entire semiotic impact of a multimodal audiovisual product. It is nevertheless worth noting that some linguistic explorations into the interrelation of semiotic resources such as words, images, and music (e.g., Thibault, 2000; Baldry & Thibault, 2006; Pastra, 2008) has provided the important groundwork to develop ideas on multimodal analyses for subtitling. For example, Thibault's (2000) work on multimodal transcription provided Taylor (2003, 2004) with the basis for investigating how the integration of semiotic modalities in a film could assist the subtitler.

A more recent and systematic attempt to connect subtitles with visual elements in audiovisual products is carried out by Chen (2019). In her work that integrates theories of multimodality, systemic functional linguistics, and semiotic translation, Chen proposed a multimodal framework for subtitling non-verbal information in films. The framework identifies three metafunctions of subtitle-image interactions, i.e., representational, compositional, and interactive. It is found that subtitling is often influenced by the visual elements on the screen, such as human faces, body movements, and surrounding settings. While the study has managed to underscore the role of semiotic interaction in subtitling, it initiates more potential for further research. Firstly, as Chen's framework focuses on the functions of subtitle-image interactions, it does not provide a refined framework to categorize different types of subtitle-image relations. Secondly, as admitted by Chen, her qualitative study is based on a small data bank of subtitles randomly selected from ten films, which calls for a larger dataset for the sake of representativeness. Additionally, the potential effects of the semiotic interplay between subtitles and image can be further examined from the real audience's viewing experience rather than the researcher's (subjective) perspective.

To date, there is still an appeal for a more systematic study of the contribution of non-verbal modes to the overall meaning of AVT (Pérez-González, 2014a). More applicable frameworks for analyzing the subtitle-image interplay are still needed. Moreover, there has been little agreement about how subtitling may alter the multimodal relations in the audiovisual content and how such alterations may influence real-life viewing experience. More empirical evidence is still needed to ascertain the extent to which non-verbal elements matter for subtitling and audience reception.

### *2.1.4 The conceptualization of translation shifts in subtitling*

The concept of *shift* has been extensively investigated in the field of Translation Studies, so as in AVT studies. While the theoretical interests in categorizing linguistic changes through translation can be traced back to the notable taxonomy by Vinay and Darbelnet (1958), the term *translation shifts*, as Munday (1998, 2016) suggested, was first introduced to the field of Translation Studies in Catford's work, *A Linguistic Theory of Translation* (1965). According to Catford's definition, translation shifts refer to "departures from formal correspondence in the process of going from the SL to the TL" (1965, p. 73). By this narrowly linguistic definition, shift analysis in the early years can be regarded as "structure-oriented" (Zhang & Pan, 2009, p. 352), revolving around the formal divergences between a source and a target text caused by the systemic differences between the two.

In its later theoretical development, the concept was emancipated from the sole linguistic focus and encompassed factors beyond language structures, such as text styles (Miko, 1970), text types and functions (Reiss, 1977), and translational norms in target cultures (Toury, 1980). A more significant development in the conceptualization of translation shifts occurred with Van Leuven-Zwart's (1989, 1990) detailed model. The model recognized shifts as a "phenomenon inherent to translation" rather than viewing them as "mistranslations" or "deviations of the norm" implied in previous models (1990b, p. 228, as cited in Cyrus, 2009. p. 89). Van Leuven-Zwart's approach steered the prescriptive undertone in previous research to a more neutral attitude towards the concept. Translation shifts were no longer deemed as somewhat undesirable but as justifiable techniques, which "makes them a suitable object of investigation within descriptive translation studies and the empirical corpus-based approach" (Cyrus, 2009, p. 89).

In AVT research, although the term “translation shifts” is usually not explicitly mentioned, researchers have dedicated their efforts to studying changes between a source and a target subtitle, proposing various terms related to subtitling strategies (see Section 2.1.1). However, their approaches to studying shifts have often been similar to those applied in previous translation research, primarily examining shifts from linguistic or cultural perspectives. As Pérez-González (2014b) argued, a significant amount of AVT research still pivoted around “elaborating taxonomies of different types of equivalence between short, decontextualized stretches of dialogue in the source and target language, with little or no attention to the interplay between dialogue and visual semiotic resources” (p. 185).

A recent theoretical and multimodal exploration of shifts in subtitling research was conducted by Qian and Feng (2020). They applied the term “intersemiotic shifts” and categorized five major types of shifts, namely, addition, omission, addition + omission, compensation, and typographic transformation. Their research revealed that intersemiotic shifts in TV drama tended to occur when the subtitles involved forms of address, modal particles, repeated words, and inner monologues. While their effort to expand translation shift theories is commendable by considering extra-linguistic elements, a more comprehensive model is still required to systematically describe the complex changes in text-image relations observed in films. The framework proposed later in this study (see Section 3.4) goes beyond Qian and Feng’s (2020) model by incorporating more refined categories and considering a wider range of linguistic and extra-linguistic elements in the analysis of translation shifts.

## 2.2 Corpus-based and eye-tracking studies on subtitling

To investigate viewers’ reception of subtitling, previous research has applied various empirical approaches such as surveys with questionnaires (Aleksandrowicz, 2019; Wu, 2017), interviews (Božović, 2019), and direct observation (Lee et al., 2013). Considering methodological relevance to the present project, the following sections will focus on corpus-based and eye tracking studies, respectively, although it should be noted that these methods are usually used in tandem with others to collect “more comprehensive data” (Orrego-Carmona, 2019, p. 369).

### *2.2.1 Corpus-based studies on subtitling*

Corpora, defined as "a collection of naturally-occurring language texts" that are stored and processed electronically (Sinclair 1991, p. 172), have been increasingly applied as an empirical method to AVT research in recent years (Bruti, 2020). For AVT studies, the methodological strength of corpora is "to provide reliable generalizations and achieve descriptive adequacy by moving beyond the limited scope of single case studies" (Pavesi, 2019, p. 315). Previous corpus-based studies on subtitling derived findings from either monomodal corpora or multimodal corpora.

Monomodal corpora in subtitling studies were compiled from pure texts of subtitles, such as the parallel English-Galician corpus of film subtitles (Sotelo Dios, 2011) and a similar bilingual corpus of film subtitles in Italian and Dutch (Caniato et al., 2015). Those corpora were mainly used to identify certain textual characteristics of subtitles, which were summarized by Pavesi (2019) as register-specificity (e.g., naturalness), translation tendencies (e.g., norms and universals), and translation strategies (e.g., translation shifts). For example, Tirkkonen-Condit and Mäkisalo (2007) built a subtitle corpus totaling around 100 million Finnish words and found that translated subtitles were more concise and colloquial than other types of translated texts and non-translated texts. The results suggested that the language of subtitles had its own register-specificity independent from other language genres. While findings from these monomodal corpora have deepened the understanding of subtitling features, they did not acknowledge the non-verbal elements in the audiovisual content. As Gambier (2006, p. 7) criticizes, it is "a contradiction to set up a database or a corpus of film dialogues and their subtitles, with no pictures, and still pretend to study screen translation". In this respect, multimodal corpora appear to be a new favorable method to "include a fully-fledged semiotic account of communication" in AVT (Bruti, 2020, p. 390).

Multimodal corpora are defined as "collections of 'data' in which distinct semiotic modes are *presumed* (as a research hypothesis) to be at work" (Bateman, 2014a, p. 241, emphasis in original). Unlike monomodal corpora, multimodal corpora of subtitles not only compare the source-target textual pairs but also take into account the visual and/or acoustic meaning-making resources. Although they are more time-consuming, complex and difficult to build and smaller in size compared with their monomodal counterparts (Soffritti, 2019), multimodal corpora contain quantitative information that can reveal how subtitle translation works in multimodal communications.

The research by Mattsson (2009) and Ramos Pinto and Mubaraki (2020) are two of the few studies that apply multimodal corpora to examine subtitle-image interplay. In her doctoral dissertation, Mattsson (2009) built a multimodal corpus of ten US films to investigate the subtitling of discourse particles (i.e., *well*, *you know*, *I mean*, and *like*) from English into Swedish. The author used several parameters for analyzing the interplay between the subtitles and other non-verbal elements such as intonation, pauses, and body language of the speakers. It was found that the interpersonal function of the discourse particles was often lost by the use of the non-translation strategy. Mattsson reasoned that film dialogues usually provided a non-verbal context (e.g., intonation and body language) to hint at the interpersonal communication and thus the interpersonal function was deemed unnecessary for translation. In a more recent study, Ramos Pinto and Mubaraki (2020, p. 26) conducted a multimodal corpus analysis on the subtitling of non-standard language varieties in films. From a corpus of three English films with Portuguese target subtitles, the authors found that the original linguistic varieties (e.g., dialects) were very often standardized in the target subtitles. It was assumed, though, that the loss of the original linguistic features did not necessarily lead to the reduction of meaning because visual information could also be used by the viewers as meaning-making resources. From the corpus analysis, Ramos Pinto and Mubaraki (2020) asserted that "a comprehensive analysis of subtitling cannot remain focused on the verbal mode alone and needs to account for the intermodal network in which subtitling participates".

Previous studies have demonstrated the feasibility and methodological power of corpora for subtitling research. By taking non-verbal resources into consideration, multimodal corpora have extended the line of monomodal corpus research. However, and also interestingly, most of the discussions in multimodal corpus research centered on some exclusively specific linguistic elements such as discourse particles and language varieties. Few studies have looked into more common and comprehensive elements in language use. Furthermore, patterns discovered from previous multimodal corpora were rarely examined from the audience's point of view. Little is known about how the subtitle-image interaction may affect viewers' allocation of visual attention on the screen.

### *2.2.2 Eye-tracking studies on subtitling*

The earliest interests in applying eye-tracking technology to the study of subtitling emerged from Géry d'Ydewalle and his colleagues in the 1980s. In their line of pioneering research at the Katholieke Universiteit Leuven in Belgium, they used eye trackers to explore the impact of the information redundancy of subtitling (e.g., d'Ydewalle et al., 1987; d'Ydewalle et al., 1991; d'Ydewalle & Gielen, 1992), One of the important findings from those investigations is the automaticity of subtitle reading. In other words, even when the viewers knew the spoken language of the audiovisual program and they did not need the support of subtitles to comprehend the content, they were still inclined to allocate their visual attention to the subtitles. Such automaticity was reasoned by d'Ydewalle et al. (1991) with two explanations. The first one was that reading the subtitles was a more efficient way to follow and understand the audiovisual content. The other reasoning was that the visual modality was dominant in viewers' processing of information instead of the auditory channel.

These early explorations, proving the significant role of subtitles in altering viewers' visual attention, laid the foundation for subsequent experimental research on subtitling. However, due to the technological limitations in the early years, the findings from these studies were somewhat complex and challenging to generalize. For example, the eye tracking device mostly used in d'Ydewalle's experiments before the 2000s was the "eye-movement-registration system (DEBIC 80)" (d'Ydewalle et al., 1991, p. 655), with a sampling frequency of 50 Hz. In other words, the device recorded the gaze data 50 times per second, which is now considered relatively slow compared to the widely accepted benchmark rate of over 250 Hz for high-quality data collection (Holmqvist et al., 2011).

With the increasing availability of eye tracking devices and inspired by the initial works of d'Ydewalle and his colleagues, more academic efforts have been engaged to examine the dynamic nature of subtitling and its multi-faceted impact on the viewer's gaze patterns. Over the past decade, eye tracking studies on subtitling have encompassed a wide range of factors related to subtitle processing. Regarding the diverse subtitle-related factors that can potentially affect reception, Gambier (2018, p. 57) identified three major types of variables: 1) "the space–time characteristics of subtitles" such as position, font size, and presentation rate, 2) "textual parameters" such as lexical frequency and text segmentation, and 3) "para-textual features" such as punctuation. In addition, subtitle reading patterns could also be influenced by the characteristics of the viewers themselves such as their age and language proficiency, which, according to Chesterman (1998), can be identified as the sociological variables. Chesterman (1998) also identified the audiovisual variables, which refer to the non-verbal elements of the audiovisual product such as the genre of the program and the image-language interplay. With Gambier's (2018) and Chesterman's (1998) taxonomies combined, the variables that have been investigated in the previous eye tracking experiments can be sorted out as: 1) **subtitle-textual variables**, 2) **subtitle-presentation variables**, and 3) **viewer-sociological variables**.

**The subtitle-textual variables**, concerning the linguistic aspect of subtitling, have attracted most researchers' attention in the past decade. Those examined linguistic factors in subtitling have encompassed translation strategies (Ghia, 2012; Künzli & Ehrensberger-Dow, 2011; Ragni, 2020; Zheng & Xie, 2018), availability of L1 and/or L2 subtitles (Kruger et al., 2014; Liao et al., 2020; Wang & Pellicer-Sánchez, 2022), availability of punctuation marks (Cui et al., 2023), non-conventional subtitles (Secară, 2011; Fernández et al., 2014), and subtitling source (Hu et al., 2020; Matthew, 2021; Orrego-Carmona, 2016). **The subtitle-presentation variables** refer to the way the subtitle is shown in the audiovisual content. Previous research interests have ranged from the standard issue of presentation rate (Kruger et al., 2022; Szarkowska & Gerber-Morón, 2018), text segmentation (Gerber-Mormon & Szarkowska, 2018; Gerber-Morón et al., 2018; Perego et al., 2010; Rajendran et al., 2013), the number of lines (Hefer. 2013a; Hefer, 2013b) to some more innovative presentation styles like integrated subtitles (Black, 2022). **The viewer-sociological variables** concern the characteristics of the subtitle viewers such as their linguistic background (Gerber-Morón et al., 2018) and previous experience with subtitling (Gerber-Mormon & Szarkowska, 2018). These examined variables have testified to an upsurge of subtitling studies resorting to eye tracking technology.

To provide an overview of the research interests over the past decades, Table 2.1 summarizes the variables and eye-tracking measures examined in the reviewed studies in this section. Two points are noteworthy from a close examination of the table. The first point is the strikingly diverse variables charted in the field of subtitling. The most investigated factor in previous studies has been the subtitle-textual variable. In other words, subtitling researchers are very much interested in how the linguistic content of the subtitles may affect viewers' gaze patterns. However, few studies have considered how the non-verbal resources may interact with the subtitle texts and thus affect viewers' visual attention, as the subtitle is only "part of the bigger whole" in meaning construction along with other semiotic resources in an audiovisual product (Jewitt et al., 2016, p. 24). Additionally, given that different variables have been proved to be effective on viewers' visual attention, future researchers should keep this issue in mind and have adequate control of all the possible confounding variables, otherwise it might hinder the generalizability of the findings obtained from the eye tracking experiment.

The other critical point is the wide array of eye-tracking measures for analysis. As illustrated in Table 1, numerous eye-tracking measures have been examined in previous studies on subtitling, with fixation duration and fixation counts being the most frequently explored (see Section 5.2.5 for definitions of the eye-tracking measures examined in this thesis). This large pool of possible options has provided an important road map for research comparability and research continuity in the field of subtitling. For future eye tracking studies on subtitling, researchers are suggested to follow some of the eye-tracking measures that have been commonly examined so as to ensure research comparability. Nevertheless, it is also essential for researchers to judiciously choose the measures that align with the specific focus of their research inquiries. which will yield fresh empirical insights in subtitling studies.

**Table 2.1** Eye-tracking measures reported in previous subtitling studies on different types of independent variables

| **Source** | **fixation duration** | **fixation count** | **time to first fixation** | **skipped subtitle** | **regression** | **deflection** | **Dwell time** | **Dwell count** | **Revisit** | **saccade length/ amplitude** | **RIDT*** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cui et al. (2023) | 🟧+ | 🟧+ | | | | | | | | | |
| Black (2022) | 🟦+ | 🟦+ | 🟦+ | 🟦+ | | 🟦+ | | | | | |
| Kruger et al. (2022) | | | | 🟦+ | | | | | 🟦+ | | |
| Wang & Pellicer-Sánchez (2022) | 🟧+ | 🟧+ | | 🟧+ | | | | | 🟧+ | | |
| Matthew (2021) | | | | | | | | | | | 🟧+ |
| Hu et al. (2020) | 🟧+ | 🟧+ | | | | | 🟧+ | 🟧+ | | | |
| Negi (2020) | 🟧+ | 🟧+ | | | | | | | | | |
| Liao et al. (2020) | 🟧+ | | | | | | 🟧+ | | | | |
| Ragni (2020) | 🟧+ | 🟧+ | | | | | | | | | |
| Gerber-Morón et al. (2018) | 🟦+ 🟩+ | 🟦+ 🟩+ | | | | | 🟦+ 🟩+ | | 🟦+ 🟩+ | | |
| Gerber-Morón & Szarkowska (2018) | 🟦+ 🟩+ | | | | | | 🟦+ 🟩+ | | 🟦+ 🟩+ | | |
| Szarkowska & Gerber-Morón (2018) | 🟧+ 🟦+ 🟩+ | 🟧+ 🟦+ 🟩+ | | | | | 🟧+ 🟦+ 🟩+ | | 🟧+ 🟦+ 🟩+ | | |
| Zheng & Xie (2018) | 🟧+ | 🟧+ | 🟧+ | | | | | | | | |
| Orrego-Carmona (2016) | 🟧+ | 🟧+ | | 🟧+ | | 🟧+ | | | | | |
| Fernández et al. (2014) | 🟧+ | 🟧+ | | | | | | | | | |
| Kruger et al. (2014) | | | | | | | 🟧+ | | | | 🟧+ |
| Hefer (2013a) | 🟧+ 🟦+ | | | | | | 🟧+ 🟦+ | | | | |
| Hefer (2013b) | 🟧+ 🟦+ | 🟧+ 🟦+ | | | | | 🟧+ 🟦+ | | | | |
| Rajendran et al. (2013) | 🟦+ | 🟦+ | | | | 🟦+ | | | | | |
| Ghia (2012) | | 🟧+ | | | 🟧+ | 🟧+ | | | | | |
| Künzli & Ehrensberger-Dow (2011) | 🟧+ | 🟧+ | | | | | | | | | |
| Secară (2011) | 🟧+ | 🟧+ | | | 🟧+ | 🟧+ | | | | | |
| Perego et al. (2010) | 🟦+ | 🟦+ | | | | 🟦+ | | | 🟦+ | 🟦+ | |
| Caffrey (2008) | | 🟧▲* | | | | | | | | | |

**Notes**: different types of independent variables assessed by eye tracking metrics are coded by the four colors below. When two or more types of variables are simultaneously measured in one study, the colored cell in the table is split accordingly into two or more colors.

🟧 : subtitle-textual variables; 🟦 : subtitle-presentation variables; 🟩 : viewer-sociological variables;

* RIDT refers to the Reading Index for Dynamic Texts proposed by Kruger and Steyn (2014).

* The triangle "▲" indicates the eye-tracking metric is examined with consideration of subtitle-image interplay.

### 2.3 Summary

The theoretical endeavor and methodological exploration in previous subtitling research have extensively advanced understanding of the nature of subtitling. As noted by Zhang and Feng (2020), adopting a multimodal perspective in the study of translation "enables a better understanding of intersemiotic translation and widens the scope of translation studies" (p. 9). Contrary to the traditional language-centered viewpoint, the recently rising multimodal analysis of subtitling has acknowledged that subtitling is not a process of simply translating textual information but involves transferring non-verbal meaning of the audiovisual products. However, there is still a dearth of refined frameworks to systematically conceptualize the interplay between subtitles and other semiotic resources. Moreover, few researchers have tested multimodal subtitling strategies on real viewers. Little is known about how a multimodal consideration for subtitling may affect the audience's viewing experience. In this respect, the growing application of multimodal corpora and eye-tracking technology in the past decade have opened up more opportunities for empirically examining the nature of subtitling as multimodal representation and its impact on audience reception.

# CHAPTER 3 THEORETICAL FRAMEWORK

Based on the research gaps addressed in the last literature review chapter, this chapter introduces two theoretical frameworks for further empirical analyses. Section 3.1 first examines some previous theoretical attempts by Martinec and Salway (2005), Unsworth (2006, 2007), and Pastra (2008) to sort out the interwoven relations between text and image. Section 3.2 then explores the structures in text and image that constitute the multimodal representation of a narrative, exploring the basic analytical units in text-image relations. The exploration draws on theoretical insights from Halliday's (Halliday & Matthiessen, 2014) systemic functional grammar and Kress and Van Leeuwen's (2021) visual grammar.

In Section 3.3, a new theoretical framework of text-image relations is proposed to account for the dynamic interplay between subtitles and image in AVT. The framework categorizes four types of text-image relations in audiovisual products termed as *4 Cs*, including *concurrence* (text = image), *complementarity* (text > image), *condensation* (text < image) and *contradiction* (text ≠ image).

Based on the framework of text-image relations and real-life subtitling cases, in Section 3.4, the other framework is proposed, which focuses on the translation shifts in text-image relations through interlingual subtitling. The framework classifies five major types of shifts: *non-shifts*, *obligatory shifts*, *preferential shifts*, *strengthening shifts* (*4 Es*: *expansion*, *explicitation*, *enhancement*, and *elaboration*), and *weakening shifts* (*4 Ds*: *detachment*, *diminution*, *dilution*, and *decrement*). The two frameworks serve as the theoretical foundation for the corpus study and the eye-tracking experiment conducted for this thesis.

## 3.1 Previous frameworks of text-image relations

The previous chapter reviews how researchers have conceptualized and investigated subtitling over the past decades. In a subtitled audiovisual product, from a multimodal perspective, the interplay between the subtitles and other mise-en-scène elements can be complex (Chen & Wang, 2019). To identify and sort out such interwoven relations between text and image, either in static or dynamic media, attempts have been made by researchers, most notably Martinec and Salway (2005), Unsworth (2006, 2007) and Pastra (2008). Their frameworks of text-image relations are discussed and compared in the following subsections.

### *3.1.1 Martinec and Salway's (2005) generalized system of text-image relations*

Martinec and Salway's (2005) system for analyzing text-image relations is considered as a fine set of criteria (Bateman, 2014b). The system is driven by two main motivations, that is, to classify text-image relations in both old and new genres of multimodal discourse, and to specify text-image relations adequate for practical analysis. To this end, the authors combine previous theoretical explorations by Halliday (1985, 1994, cited in Martinec & Salway, 2005) and Barthes (1977a, 1977b, cited in Martinec & Salway, 2005) and present the generalized system of *status* and *logico–semantics*. These two subsystems, independently combined, make up Martinec and Salway's (2005) generalized system as a whole.

The subsystem of *status* refers to the relative position between text and images. Under this subsystem are two major text-image relations: *equal* status and *unequal* status. In the *equal* status, a text and an image can be either *independent* when they do not modify each other or *complementary* when they are equally joined and one of the modes modifies the other. It should be noted here, nevertheless, that the *independent* and *complementary* relations are identified on the sense of processes in transitivity analyses rather than of semantic connection. In other words, the status between the text and the image is analyzed in terms of their participant-process-circumstance configuration. On the other hand, when it comes to the *unequal* status, a text or an image cannot stand on its own and relates to only part of the other, with *image subordinate to text* or *text subordinate to image*. Briefly put, the *equal* status can be understood as a whole-to-whole relation, while the *unequal* status is a part-to-whole relation.

The subsystem of *logico-semantics* concerns the semantic interaction between text and images. It is also composed of two major relations: *expansion*, and *projection*. *Expansion*, based on Halliday's classification, is further categorized into three subtypes, that is, *elaboration*, *extension*, and *enhancement*. For the first subtype, *elaboration*, it is further divided into two kinds: *exposition* and *exemplification*. If the text and the image are of the same level of generality, they are identified as a relation of *exposition*. When their generality levels are different, on the other hand, the text-image relation is recognized as *exemplification*, with either the text or the image more general. As for the second subtype, *extension,* it is the interaction between a text and an image where either one adds new relevant information to the other. In the relation of *enhancement,* the text or the image qualifies the other circumstantially by demonstrating the *time*, *place*, or *reason/purpose*. The other major relation, *projection*, is divided into two subtypes: *locution* and *idea,* in which *locution* is the projection of the exact wording, and *idea* the projection of the approximate meaning. The *projection* relation mostly occurs in two text-image contexts: in comic strips and in textbooks where texts and diagrams are closely integrated.

One of the contributions of Martinec and Salway's framework is their brief yet important discussion on the *units* in texts and images. Since both texts and images are built up by smaller components, it is problematic to analyze text-image relations without determining which components of the text and the image are being related. As the authors clarify, the largest unit in a text related to an image is the paragraph, whereas smaller units can also be ground into clauses and even words for analysis. Regarding the unit in images, as can be inferred from the authors, a visual unit refers to one cohesive set of processes, participants, and circumstances. The issue of determining units in text and image is further discussed in Section 3.2.

### *3.1.2 Unsworth's (2006, 2007) framework of text-image ideational meanings*

Unsworth's (2006) angle on text-image relations is different from Martinec and Salway's (2005). Focusing on educational materials, Unsworth (2006) adapts a systemic functional semiotic approach to multimodal texts. He discusses the meaning-making resources of text-image interaction from three perspectives: ideational, interpersonal, and textual. Although each of the perspectives is explained and exemplified by Unsworth, he puts most of his efforts in developing the framework of ideational meanings. In his later paper presented to the 33rd International Systemic Functional Congress, Unsworth (2007) further refines his framework of text-image relations in the construction of ideational meanings.

According to Unsworth (2007), the ideational meanings created by a text-image combination are classified into two major types: *expansion* and *projection*.

The *expansion* relation is subdivided into three types: *concurrence*, *complementarity*, and *enhancement*. In a *concurrence* relation, the meanings conveyed in the text and the image are ideationally equivalent. The *concurrence* relation can be further subcategorized into four kinds: c*larification*, where the text/image explains the meaning of the alternative mode; *exposition*, in which the meanings of the text/image are re-expressed by the other mode; *exemplification*, where the image exemplifies the text or the text instantiates what is depicted more generally in the image; and *homospatiality*, in which the text and the image are combined together to form a homogenous entity (i.e., a word-picture). For the *complementarity* relation, it refers to the case where the meanings of the text and the image are complementary to each other and they combine to create a fuller meaning. *Complementarity* consists of two subtypes: *augmentation* and *divergence*. *Augmentation* is where the text/image provides new, consistent information for the other mode. *Divergence*, on the other hand, involves a text/image adding opposing content to the other mode. Regarding the *enhancement* relation, it can be understood as an adverbial connection of images and text. In such a relation, the text/image may modify the other mode in terms of manner, condition, place, time, or reason.

The second major type, *projection*, refers to the situation where the text and the image together show the juxtaposition of a quoted speech/thought and the participant. A *projection* can be considered either as a *verbal* subtype when the quoting is about articulation of words, or a *mental* subtype when the quoting is in regard to thoughts.

Unsworth's framework can be seen as an extension of Martinec and Salway's (2005) (Bateman, 2014b). Many of the terminologies and taxonomies employed in Unsworth's framework are directly borrowed from Martinec and Salway's (2005), such as *expansion* and *projection*. Moreover, although some terms in Unsworth's system are different from Martinec and Salway's, they refer to the identical meaning. For example, concerning the quoting of words and thoughts in the *projection* relation, Unsworth adopts the phrases *verbal* and *mental*, whereas Martinec and Salway choose the terms *locution* and *idea*. That being said, Unsworth's contributions to theorizing the text-image interplay should not be discounted. First of all, he has extended and supplemented the system by Martinec and Salway, supporting a closer exploration of the complex interaction between texts and images. For example, he is probably the first researcher who identifies the contrasting meanings conveyed between texts and images, which he conceptualizes as *divergence*. Furthermore, instead of being ambitious to build a system for generic purposes, Unsworth confines his exploration to education, particular to the study of multiliteracies. Such a pivotal focus sets a useful precedent for future multimodal research with specific purposes (e.g., multimedia translation).

### *3.1.3 Pastra's (2008) COSMOROE*

Focusing on dynamic multimedia discourse, Pastra (2008) examines text-image relations at a more granular level. Her proposed framework, COSMOROE (cross-media interaction relations), analyzes the semantic interaction between "images, language and body movements" (p. 306). The framework categorizes multimodal relations into three major types: *equivalence*, *complementarity*, and *independence*. Each of the three types is further divided into several subtypes.

*Equivalence* is the relation where the information conveyed in different modes is equivalent in meaning and refers to the identical entity. This relation is further categorized into four subtypes: *token-token*, *type-token*, *metonymy*, and *metaphor*. The former two relate to literal equivalence and the latter two to figurative. *Token-token* means that different modes point to the same exact entity (e.g., a boy's name and the corresponding image of the boy). *Type-token*, on the other hand, is where one mode indicates the class of the entity conveyed through another mode (e.g., a boy presented in the image and the text saying "human beings"). *Metonymy* refers to the case where each of the modes indicates a different entity but the two entities are semantically equal (e.g., the image showing a boy and the text saying "innocence"; the boy here acting as the symbol of innocence). *Metaphor*, just as the term implies, means one mode indicates a similarity of an entity and the entity is also conveyed through another mode (e.g., the image presenting a boy and the text saying "the angel of the family").

The second major type of relation, *complementarity*, refers to the case where the information contained in one mode is complementary to that conveyed in another mode. Of this relation, four subtypes are normally identified: *exophora*, *agent-object*, *apposition*, and *adjunct*. According to Pastra, the former three subtypes can be either essential or non-essential based on their level of complementarity, whereas the last subtype can only be non-essential. *Exophora* is the relation in which one mode resolves the reference made by another (e.g., the reference of the word "it" in the sentence "the boy is eating it" is resolved from the image which shows a boy is eating a cake). Regarding an *agent-object* relation, one mode indicates the subject/object of a movement/event conveyed in another mode (e.g., the text stating "the boy is eating" and the image revealing that it is the cake that the boy is eating). In an *apposition* relation, one mode provides another mode with additional descriptive information, which is not necessarily valid or objective (e.g., the image showing a boy and the text saying "the smartest genius"). An *adjunct* relation is identified when one mode functions as an adverbial modification to the information in another mode (e.g., the text saying "the boy goes to school" and the image showing that the boy is taking a bus; here the image reveals the manner used to go to school).

The last major type of relation, *independence*, occurs when each mode stands independently with a coherent (or incoherent) message but can be combined to produce a larger message. This relation is composed of three subtypes: *contradiction*, *symbiosis* and *meta-information*. In a *contradiction* relation, one mode presents the opposite or incompatible meaning of another mode (e.g., the image showing a girl and the text stating "a boy"). *Symbiosis* is a relation where one mode provides some thematically related but noncomplementary information for another mode (e.g., a text stating "the boy is having a nice dream" and the image simply depicting a night sky; here the image is simply a visual filler that comes along with the verbal message). On the other hand, *meta-information* means one mode conveys semantic message that is inherently independent from pieces of messages conveyed in another mode (e.g., the font or special effect of texts in a post revealing additional information on the visual content).

What makes Pastra's COSMOROE framework distinct from prior ones is its applicability to multimedia corpora analysis, which has been endorsed by empirical evidence provided by other researchers (e.g., Ramos Pinto & Mubaraki, 2020; Zlatintsi et al., 2017). Pastra employed the framework to annotate the meaning-making relations in a corpus built on two TV programs. After comparing the annotations by an expert and a trainee annotator, it was observed a high inter-annotator agreement. Although the author admits several problems of using the framework in practice (e.g., demanding cognitive effort and time in annotation, indirectness of identifying certain COSMOROE relations), the framework is believed to be useful given its "descriptive power" and "computational applicability" (Pastra, 2008, p. 300).

### *3.1.4 Comparison and summary of previous frameworks of text-image relations*

The preceding frameworks of text-image relations have markedly advanced our understanding of multimodal texts. To better discern the similarities and differences between these frameworks, we can draw analogies to text-image relations from mathematics and regard the information conveyed in texts and the images as numeric entities. As O'Hallaron (2015, p. 69) aptly pointed out, "Mathematical symbolic notation evolved to encode mathematical relations in the most economical way possible, providing a precise, robust and flexible tool for capturing and rearranging mathematical relations without ambiguity."

As Table 3.1 shows, the verbal meaning can be equal to (=), approximately equal to (≈), less than (<), greater than (>), opposing to (≠), or simply juxtaposed with (:) the visual meaning (see the notes below Table 3.1 for brief definition of the terms). Based on these mathematical relationships, the concepts and terms proposed in the previous frameworks are sorted out accordingly. Generally, one categorized term for a text-image relation only goes to one single mathematical relationship. If the proposed term passes over two or more relations, it can be assumed that the term may have a blurred definition or boundary. Such issues, as can be seen from Table 2, are observed in all the three frameworks.

**Table 3.1** Comparison of preceding frameworks of text-image relations from a mathematic perspective

| **Mathematical relation** | **Martinec & Salway (2005)** | **Unsworth (2007)** | **Pastra (2008)** |
|---|---|---|---|
| text = image | equal status (independent)<br>elaboration (exposition) | concurrence (clarification)<br>concurrence (exposition)<br>concurrence (homospatiality) | equivalence (token-token) |
| text ≈ image | equal status (complementary)<br>expansion (extension)<br>expansion (temporal enhancement)<br>expansion (spatial enhancement)<br>expansion (causal enhancement) | complementarity (augmentation)<br>enhancement (manner)<br>enhancement (condition)<br>enhancement (spatial)<br>enhancement (temporal)<br>enhancement (causal) | complementarity (apposition)<br>complementarity (exophora)<br>complementarity (agent-object)<br>complementarity (adjunct) |
| text >/< image | elaboration (exemplification)<br>unequal status (subordinate) | concurrence (exemplification) | equivalence (type-token) |
| text ≠ image | / | complementarity (divergence) | independence (contradiction) |
| text : image | projection (locution)<br>projection (idea) | projection (verbal)<br>projection (mental) | equivalence (metonymy)<br>equivalence (metaphor)<br>independence (symbiosis)<br>independence (meta-information) |

Notes:

"=" (equal to): two modes equally refer to the same meaning;

"≈" (approximately equal to): two modes have some overlapping meanings but complement each other and combine to form a larger meaning;

">/<" (less or greater than): either mode is semantically more general;

"≠" (opposing to): two modes contradict each other by presenting opposite meanings;

":" (juxtaposed with): two modes are thematically put side by side;

The terms within parentheses are the subtypes of text-image relations.

In Martinec and Salway's (2005) system, one major relation, *elaboration,* cuts across both the "=" and ">/<" relationship, with its subtypes scattered into two distinct mathematical categories. In Unsworth's (2007) framework, two major types, i.e., *concurrence* and *complementarity,* also pass through different mathematical relationships. Similar cases go to Pastra's (2008) framework, as the major type of *equivalence* goes over the different mathematical categories. These overleapt terms indicate that a proposed term in these frameworks can sometimes have different or even contradictory meanings. For example, the term *concurrence* used by Unsworth refers to, assumedly, the equal ideational meaning between texts and images, but one of its subtypes, *exemplification,* actually alludes to the subordinate relation where the text or the image carries more general meaning. In other words, the meaning of the term *concurrence* here is somewhat ambiguous.

Apart from the naming and defining issues, more critical issues may occur when researchers try to apply the frameworks to practical analysis. Indeterminacy and confusion could come up. For example, in the Academy Award-winning film *Coco* (2017), a scene depicts a boy talking to a stray dog (with its first appearance in the film) and the boy saying, "Hey, Dante". Here, suppose a researcher employs Martinec and Salway's (2005) system to analyze the text-image relation. As the word "Dante" refers to the name of the dog shown in the image, the researcher's intuitive response may be that the text and the image form an *exposition* relation because they equally point to the same entity on the same level of generality. However, if the researcher makes a second thought, the text-image relation here can also be understood as *extension*, because both the text and the image add new relevant information to each other. In other words, the text provides the name of what is shown in the image (i.e., Dante), and the image shows what the name in the text refers to (i.e., a dog). If the viewers only read the text, they cannot know it is a name for a dog; if they only watch the image, they are not able to know the dog's name. It is by combining the extensive intermodal information that the viewers obtain the whole message — "this is a dog whose name is Dante". The same problem goes to Unsworth's (2007) framework. An analyst can either recognize the text-image relation here as *exposition* (with the same reason just mentioned) or view it as an *augmentation* relation given that the text and the image both provide new information for each other.

Similar ambiguity is observed in Pastra's (2008) framework. In an example the author provides in her discussion, the text states an utterance "hold…" while the image shows a man reaching out his hand and intended to give a microphone to the other man. Here, Pastra believes the text and the image form an *essential agent-object* relation, as the microphone in the image represents the object of the verb “hold” in the text. This analysis, undoubtedly, is reasonable and proper. However, as the body movement “hold” is also seen from the image and equally corresponds to the verb “hold”, the viewers also have reasons to claim that the text and the image constitute an *equivalence* relation, or put it more precisely, a *type-token* relation.

Notwithstanding the potential problem mentioned above, the previous frameworks have provided detailed theoretical basis for analyzing the text-image interplay in various genres and media. Except Unsworth (2006, 2007) who focuses on educational textbooks, Martinec and Salway (2005) and Pastra (2008) have explored various forms of new media. Martinec and Salway (2005) have investigated static media, both traditional and new, including textbooks, digital encyclopedias, and websites for news and art. Pastra (2008), on the other hand, further turns to investigate dynamic multimedia, closely exploring the text-image relations on TV programs. All these works raise the possibility to systematically examine the complex nature of text-image interactions in audiovisual texts.

## 3.2 Multimodal representation: *processes*, *participants*, and *circumstances*

The analysis of text-image relations in audiovisual texts entails a thorough examination of how the text and image respectively represent the multimodal world. In other words, before delving into the various types of text-image interplay, it is imperative to first discern the textual and visual constituents employed to construct a *narrative*, which is defined in this study as a story or a series of events conveyed through textual and/or visual elements. For example, who or what is displayed in the text or image? What activities or qualities of the participants are depicted? What circumstances are associated with them? To systematically identify these elements within both the language and image structure, the analysis in this study draws on theoretical insights from Halliday’s (Halliday & Matthiessen, 2014) systemic functional grammar and Kress and Van Leeuwen’s (2021) visual grammar, focusing on how textual and visual components represent the *participants*, *processes*, and *circumstances* in a multimodal narrative.

### *3.2.1 Halliday’s systemic functional grammar and verbal transitivity*

Language, as Halliday (1994) asserts, “enables human beings to build a mental picture of reality to make sense of what goes on around them and inside them” (p. 106). As a fundamental concept in Halliday’s (Halliday & Matthiessen, 2014) systemic functional grammar (SFG), transitivity is a powerful tool in the analysis of experiential meaning. According to SFG, people's real-life experience, which they can express through clauses (i.e., the basic unit of grammar), consist of (a) a process unfolding through time, (b) participants being involved in the process, and (c) in optional addition, circumstances that are indirectly involved in the process such as time, space, and manner. The grammatical system that a clause employs to reflect and impose order on those various experiences and events is conceptualized as *transitivity* (Halliday & Matthiessen, 2014, p. 213). For a more straightforward explanation given by Thompson (2014), transitivity refers to "a system for describing the whole clause, rather than just the verb and its Object" (p. 94), with the functional components in the transitivity structure of a clause labeled as *processes*, *participants*, and *circumstances*. These three terms, as Thompson (2014) illustrates, reflect "our view of the world as consisting of 'goings-on' (verbs) involving things (nouns) that may have attributes (adjectives) and which go on against background details of place, time, manner, etc. (adverbials)" (p. 92).

In transitivity analysis, processes are usually expressed by the verbal group in a clause. Given that a clause is generally concerned with an action, state, or event, processes can be regarded as the essence of a clause. Halliday and Matthiessen (2014, p. 213-216) further classify six process types in language: Material (doing), Mental (sensing), Relational (being), Verbal (saying), Behavioral (behaving), and Existential (existing). Each type represents a specific kind of experience. For example, verbs like "eat" and "throw" are Material processes as they involve physical actions, and verbs like "think" and "want" are considered Mental processes for they pertain to the internal domain of someone's mind. Participants normally refer to the nominal group which are involved in the process of a clause. Participants can be further categorized into different terms, corresponding with the process types in which the participants are involved. For example, a participant is coined the Actor and Goal in material processes, Senser and Phenomenon in mental processes, Carrier and Attribute or Token and Value in relational processes, Behaver in behavioral processes, Sayer and Verbiage in verbal processes, and Existent in existential processes. Circumstances are usually concerned with the adverbial groups or prepositional phrases, which function as the background about how the process takes place in a clause. Based on their functions, circumstances can be further classified into nine common types: Extent, Location, Manner, Cause, Contingency, Accompaniment, Role, Matter, and Angle. The process-participant-circumstance model, which can be further elaborated into more delicate categories, provides a handy toolkit for examining the diverse facets of verbal representations of experience.

The transitivity model can also be applied to the analysis of translated texts (e.g., Calzada Pérez, 2007). In the study of subtitling, transitivity analysis has been employed to look into the representational features of subtitles and, more importantly, how subtitling may lead to transitivity shifts from the original spoken texts (e.g., change of process types in the target subtitles) and bring about certain contextual effects on the audiovisual narrative. In one of the early explorations, da Silva (1998) employed the transitivity system to compare how the ST and the interlingual target subtitles may differently construe the internal as well as external experience of a film character. It was observed that both the ST and the target subtitles managed to construct the character as a self-centered person unable to extend her actions, feelings, and sayings to others, given that the character tended not to address any other participants in the material, verbal or mental processes. The study by da Silva's (1998) has contributed to exhibiting the feasibility of the transitivity model as a tool for analyzing subtitles systematically through linguistic description. Mubenga (2010) and Noverino et al. (2020) investigated the translation shifts in interlingual film subtitling in terms of transitivity components. Mubenga (2010) used one example to demonstrate how the actor of the material process was reduced in the target subtitles, assuming that the omitted actor can be deduced from the character portrayed in the image. Noverino et al. (2020), based on a larger set of over 400 clausal instances, found that all the three transitivity components were sometimes reduced in the target subtitles. In other words, the original information concerning the participant(s) involved in an event, the event itself (i.e., processes), and the conditions accompanying the event (i.e., circumstances), could be lost in the subtitles, suggesting a different representation of the experiential meaning compared to the original verbal texts.

These previous studies have demonstrated how subtitling can affect characterization as well as the contextual representations in a film narrative. However, what remains unknown is why those translation shifts of transitivity components occur. As Noverino et al. (2020) admit, "there is no actual information of why the [transitivity] constituents are reduced by the subtitler" (p. 1030-1031). One possible explanation is that certain transitivity components are in line with the accompanying visual elements and the textual components are thus considered redundant and omitted. In this light, for conducting verbal transitivity analysis of subtitling, it appears necessary to pay additional attention to the visual transitivity components as well, which will be discussed in the following section.

### *3.2.2 Kress and Van Leeuwen’s visual grammar and visual transitivity*

Inspired by Halliday’s SFG, Kress and Van Leeuwen (2021) expanded this theory typically in the language domain to visual communication and developed the theoretical framework for visual analysis, that is, visual grammar (VG). The presupposition of VG is that visual communication can realize relatively the same fundamental social features of meaning as language can do. It is believed that the semiotic mode of visual communication has its own quite particular means of “expressing active relations between participants…processes…and circumstances” (Kress & Van Leeuwen, 2021, p. 45).

The visual transitivity system in VG consists of two main types of structures: Narrative structures and Conceptual structures. The Narrative structures present participants involved in “unfolding actions and events, processes of change and transitory spatial arrangements” (Kress & Van Leeuwen, 2021, p. 76). The processes in Narrative structures can be further classified into Action processes, Reactional processes, Speech processes, or Mental processes. The Action processes illustrate “doing” and “happening” that normally involve two participants, an Actor and a Goal. The Reactional processes are triggered by the eyelines and glance of participants, who are identified as a Reacter or a Phenomenon. The Speech processes normally depict the verbal actions with the tail of a “dialogue balloon” (like in comic strips) connecting two participants, a Sayer and an Utterance. The Mental processes, similarly, illustrate activities in the Senser’s mind, normally with the tail of a “thought bubble” which contains the Phenomenon. As for the circumstances allocated to the processes, three types are identified, that is, Setting (e.g., foreground and background), Means (e.g., manners of an action), and Accompaniment (e.g., a companion of a participant). The Conceptual structures, on the other hand, represent participants in terms of their more or less static and stable state of being. Conceptual processes are divided into three types, i.e., Classificational processes that indicate taxonomic relationships between participants, Analytical processes that construe a part-whole relationship between participants), and Symbolic processes that indicate the meaning or identity of the participants.

The visual transitivity system provides a set of elaborate guidelines for elucidating the representational meaning of images. In particular, it delineates the possible elementary units that researchers can sift out to scrutinize visual communication. Some applications of the system to audiovisual texts have been explored by Martinec and Salway (2005; see Section 3.1.1). The following section will discuss some critical issues on the combination of Halliday’s verbal transitivity model and Kress and Van Leeuwen's visual transitivity system for the analysis of text-image relations in audiovisual texts.

### *3.2.3 Combining verbal transitivity and visual transitivity*

From the above description of transitivity systems in SFG and VG, some connections between the two systems can be observed and summarized. For example, The Narrative structures in VG are well aligned with the Material processes in SFG, both representing the happening physical actions of an experience. The Conceptual structures in VG, on the other hand, are related to the Relational and Existential processes in SFG, as they all serve to depict the rather static status and features of the participants. Moreover, the three identified functions of circumstances in VG can find their counterparts in the nine common types classified in SFG.

While the visual transitivity model shares such similar coinages and theoretical grounds with the verbal transitivity system, the comparability between the two systems is highlighted by Kress and Van Leeuwen (2021):

> We have drawn attention to the fact that, while both visual structures and verbal structures can be used to express meanings drawn from a common cultural source, the two modes are not simply alternative means of representing 'the same thing'. It is easy to overemphasize either the similarity or the difference between the two modes. Only a detailed comparison can bring out how in some respects they realize similar meanings, albeit in different ways, while in other, perhaps most respects, they represent the world quite differently. (p. 73)

This statement is a reminder for researchers. As it emphasizes, each semiotic mode (like texts and images) has its own particular means of depicting a similar semantic experience. It should not be assumed that all the transitivity relations realized linguistically can also be represented visually, or vice versa. What is represented in the visual mode can seldom be exact equivalent to what is shown in the linguistic mode. The "sameness" or "difference" observed between the image and the text is a relative concept rather than an absolute value.

In this respect, when applying the two transitivity systems to the analysis of subtitling and looking into the text-image relations in audiovisual products, it is not to claim that the identified relations are absolute. Instead, they only suggest a converging or diverging representational meaning depicted in the verbal and visual processes, participants, or circumstances.

When applying the two transitivity systems to the analysis of audiovisual texts, it is important to note that the three basic transitivity components in a linguistic clause or in a visual image are not invariably obligatory. In the linguistic context, processes are the essential component in a clause whereas participants and circumstances can be optional. This issue is addressed by Halliday and Matthiessen (2014):

> Circumstantial elements are almost always optional augmentations of the clause rather than obligatory components. In contrast, participants are inherent in the process: every experiential type of clause has at least one participant and certain types have up to three participants – the only exception being, as just noted above, clauses of certain meteorological processes without any participants such as *it's raining*, *it's snowing*, *it's hailing* (but not all; for example, we say *the wind's blowing* rather than *it's winding*) (p. 221, emphasis in original).

In addition, on some occasions, a participant might only be implied in a clause and understood as an implicit part of the experiential meaning. An example of this kind is the imperative clause, where the Actor "you" is not mentioned but is understood implicitly as the participant of the process. Such issues are critical, especially when it comes to corpus-based analysis. The exclusion or inclusion of the implicit participants can have a direct impact on the final statistic results in term of frequency counts (Thompson, 2014).

Different from the optional nature of circumstances and participants in the linguistic transitivity system, all three components in the visual transitivity system appear to be consistently present. For example, when an object (participant) occurs in a film, it is always accompanied by its corresponding action or state of being (process), together with certain background information such as settings of time or place (circumstance). The major difference between their constant presence lies in their degree of *visual salience*, which is defined by Kress and Van Leeuwen (2021, p. 182) as the relative "visual weight" of the pictorial elements perceived by viewers. Salient objects may capture more attention from viewers due to factors such as size, color contrasts, sharpness of focus, and placement in the foreground or background in the image (Kress & Van Leeuwen, 2021). In this regard, while the components in the verbal transitivity system revolve around the status as either obligatory or optional, those in the visual system are the matter of being either salient or indistinctive.

In short, the sole focus on the verbal transitivity patterns of the subtitles is by its nature insufficient because, for example, participants in an audiovisual product are not only presented by what they say in language but what they do and how they do it in the visual mode. To acquire a fuller picture of the representational features of an audiovisual narrative, it is more reasonable to integrate both verbal and visual transitivity analysis.

### 3.3 A proposed framework of text-image relations for subtitling studies

So far, except for some recent attempts (e.g., Dicerto, 2018; Ramos Pinto & Mubaraki, 2020), few studies have applied the frameworks of text-image relations mentioned in Section 3.1 to the analysis of translation phenomena. To examine translation activities with these frameworks, adjustments are necessarily required. As Martinec and Salway (2005, p. 367) point out, "it may well be that the system will need to be modified as new image–text genres evolve or that at least the realizations of existing categories will change". When building a framework of text-image relations for the analysis of subtitling in AVT, five issues are worth considering, that is, whether the framework is characterized by: 1) adequate **categories** with clear boundaries, 2) unambiguous **terminologies**, 3) proper **analytical units**, 4) clear indication of the **direction** of intermodal comparison, and 5) applicability to dynamic audiovisual **media**. Simply put, we need a scheme that supports our analysis of subtitling without being overly detailed but with enough details to make the observations rigorous and adequate.

Taking these five issues into consideration, a new theoretical framework of text-image relations is proposed for the analysis of subtitling. The framework synthesizes previous insights on multimodal functional grammar and multimodal relations (Halliday & Matthiessen, 2014; Kress & Van Leeuwen, 2021; Martinec & Salway, 2005; Pastra, 2008; Unsworth, 2007). As shown in Figure 3.1, the framework consists of three major parts, i.e., text-image relations (4 Cs), and the verbal as well as visual transitivity components that intersect to form a certain relation for narrative representations. The term "text" in the framework broadly refers to any verbal elements in an audiovisual product, including the written representation/translation of spoken dialogues (i.e., subtitles), the verbal utterance (i.e., original dialogues), and the diegetic verbal contents (e.g., words on a cell phone which a film character would recognize). On the other hand, the term "image" pertains to any non-verbal information visually presented in an audiovisual product. The following subsections will describe the framework in terms of its categories, terminologies, analytical units, intermodal direction, and applicable media.

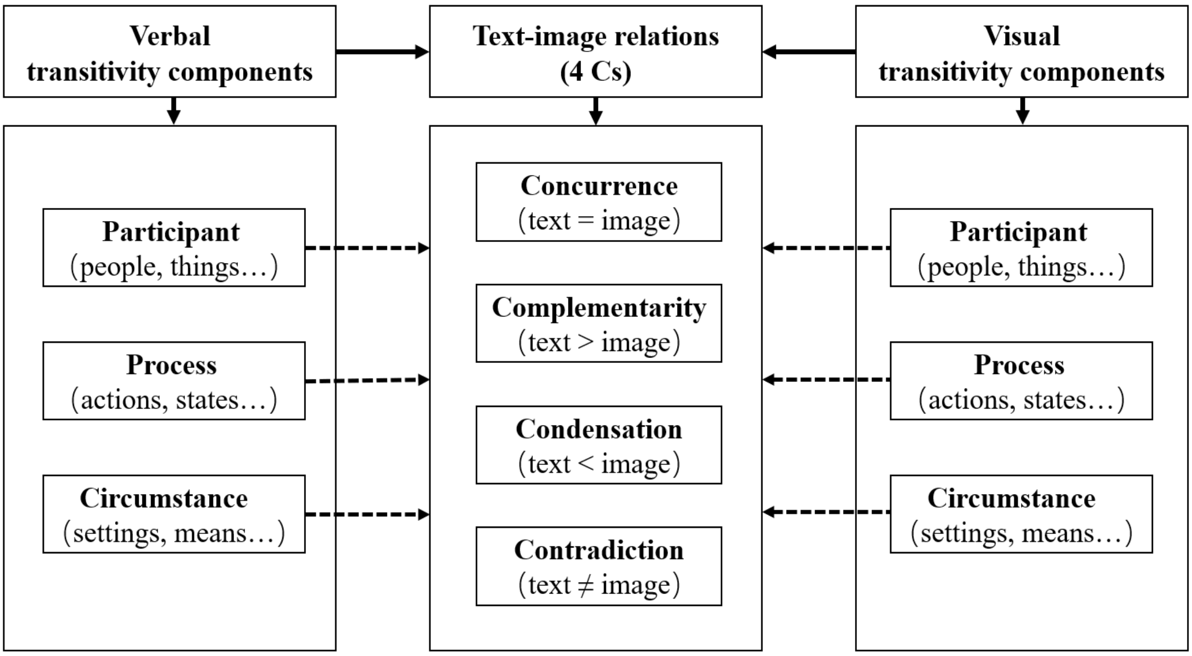

Verbal transitivity components
Text-image relations (4 Cs)
Visual transitivity components
Participant (people, things…)
Process (actions, states…)
Circumstance (settings, means…)
Concurrence (text = image)
Complementarity (text > image)
Condensation (text < image)
Contradiction (text ≠ image)
Participant (people, things…)
Process (actions, states…)
Circumstance (settings, means…)

**Figure 3.1** A proposed theoretical framework of text-image relations (4 Cs) for subtitling studies

*3.3.1 Categories of the framework*

Categorization, first of all, is a delicate task. It is seldom easy to systematically sort things into categories without subtle or overlapping boundaries. By definition, a category is a system where "two or more distinguishable objects or events are treated equivalently" (Mervis & Rosch, 1981, p. 89). The two dimensions of a category system, according to Rosch (1978, p. 30), are the "vertical" and the "horizontal" dimensions. The vertical dimension refers to "the level of inclusiveness of the category" (Rosch, 1978, p. 30), concerning how far and how many objects the category encompasses. The horizontal dimension, on the other hand, pertains to "the segmentation of categories at the same level of inclusiveness" (Rosch, 1978, p. 30), concerning how distinctive and flexible the category can be. In other words, a category system should be: (a) comprehensive enough to cover adequate objects or concepts and (b) distinctive enough to preclude blurred boundaries between the (sub)categories.

To guarantee both comprehensiveness and distinctiveness in categorizing text-image relations, one of the effective ways is to draw on the mathematics register. The "precise, robust and flexible" system of mathematical notations can describe mathematical relations "in the most economical way possible" and "without ambiguity" (O'Hallaron, 2015, p. 69). It is thus arguably practical to draw analogies to text-image relations from mathematical relations and categorize the former relations based on the latter. When the messages conveyed in texts and images are considered as numeric entities, the categories of text-image relations can be conceptualized through common mathematical notations (see also Section 3.1.4). The first common relation is equality, symbolized as “=”, which can be used for a relationship of relative equivalence between the text and the image. The relation of inequality, on the other hand, can be described as either “greater than” (symbolized as “>”) or “less than” (symbolized as “<”). It means that either the text or the image carries more information than the other. Another more general symbol for the unequal relationship is “≠”. It can be used to refer to occasions where the image and the text convey contradictory meanings. Another common relation, approximate equality (symbolized as “≈”), although commonly found in previous frameworks (see Table 2), is not included in the category, given its elusive nature and the difficulty in determining the level of approximation of multimodal meaning. Moreover, it overlaps with the relations of being “greater than” and “less than” and thus stands against the vertical dimension (or distinctiveness) of a category system. Juxtaposition (“:”) is also not included because it does not exactly capture the interactional relation between the text and the image. As Baumgarten (2008) asserts, in subtitled audiovisual products, when the text (subtitle) interacts with the image, “the two are explicitly connected” rather than implicitly juxtaposed (p. 11).

### *3.3.2 Terminologies of the framework (4 Cs)*

“Terms are the linguistic representation of concepts” (Valeontis & Mantzari, 2006, p. 4). After conceptualizing the four categories of text-image relations, the question that follows is how to form proper terminologies for these categories. As observed from the comparison of previous frameworks shown in Table 2 in Section 3.1.4, it is a common practice to use different terms to signify similar or even the same text-image relations. For example, regarding the relatively equivalent relation between the text and the image, Martinec and Salway (2005) choose the term *elaboration*, while Unsworth (2007) turns to the name of *concurrence* and Pastra (2008) *equivalence*. The choice of terms, to a great extent, is arbitrary and relies on the personal preference of the researchers. Although the arbitrariness of setting up terminologies is acceptable, Valeontis and Mantzari (2006) propose three possible systemic methods of term formation, that is, “creating new forms”, “using existing forms”, and “translingual borrowing” (p. 5). First of all, as previous researchers have brought forth a quite sufficient set of nomenclature to identify text-image relations, it is reasonable to use existing terms from the proceeding frameworks to signify the four proposed categories wherever appropriate.

Nevertheless, on the principle of transparency and consistency in linguistic representation of concepts (Valeontis & Mantzari, 2006), new words can also be created or borrowed to better represent the concepts in the research of text-image relations. From the aforementioned reasons, four types of text-image relations (4 Cs) are hereby designated:

— ***Concurrence*** (text = image), where the text or the image almost equally refers to what is presented in the other. In other words, both modes contain the same salient information in a narrative. This term is a direct use of the existing term by Unsworth (2007). The concept of *concurrence* is closely linked to textual-visual redundancy, a topic that has been examined in previous subtitling studies on viewer cognition and comprehension (e.g., Kruger & Liao, 2022; Liao et al., 2021; see also Section 2.1.2 for a review).

— ***Complementarity*** (text > image), in which the text further modifies or explicates what is shown in the image. In such a relation, the text provides more explanatory or illustrative information for the image (see Section 3.3.4 for detailed explanations of the importance of differentiating the direction of intermodal comparison). This term is derived from Unsworth's (2007) and Pastra's (2008) frameworks. It should be noted that Zabalbeascoa (2008) uses the term in a different way, referring to the notion that the text and the image complement each other. However, in this context, the term is defined as the subtitle extending the meaning of the image, with a "unidirectional" perspective (see Section 3.3.4).

— ***Condensation*** (text < image), where the text simplifies or understates what is shown in the image. In a *condensation* relation, the information encoded in the text is less specific than that in the image. This term is borrowed from Gottlieb (1992, p. 166) and Díaz Cintas and Remael (2021, p. 151). They used the term to identify a common subtitling strategy: reducing information from the oral source text due to (a) the spatial and temporal constraints inherent in subtitling and/or (b) "intersemiotic redundancy" (Gottlieb, 2001, p. 321). While the term "condensation" originally focuses exclusively on linguistic transfer, its meaning is extended in this framework to describe the reduced amount of information in the subtitles compared with that of the visual information. Moreover, the term does not refer to the process of thickening a liquid by removing some of its water content; instead, it denotes the act of making spoken, written, or depicted content more concise by omitting details or using fewer words to convey the information.

— ***Contradiction*** (text ≠ image), when the text or the image shows the divergent or opposite meaning of the other. In a *contradiction* relation, the text and the image are thematically correlated but they present contradictory meaning. The term is obtained from Pastra's (2008) scheme, and this relation is identified as *divergence* in Unsworth's (2007) system.

It is necessary to hereby clarify the meaning of the equal sign "=" in relation to *concurrence* within the framework. The symbol does not imply that a text can fully describe all aspects of the corresponding visual information. For example, when the subtitle uses a name to address a character in a film scene, it is impractical and uncommon for the text to include all visual details of the character such as body shape, skin color, clothing, etc. This kind of information is conceptualized by Baldry and Thibault (2006, p. 198) as "visual collocation", which pertains to the secondary items that specify either the character's role or the activity they are engaged in. Building on Baldry and Thibault's (2006) approach, when assessing the equality of meaning between the text and the image, the analysis of the visual elements primarily focuses on the salient items directly relevant to the ongoing narrative. Therefore, when a *concurrence* relation is identified, it signifies that the text equally represents the salient subject depicted in the image, rather than encompassing all the collocated visual elements. The "equivalence" or "sameness" identified between the image and the text is a relative concept rather than an absolute value.

To illustrate the four proposed concepts, some conceptual examples of text-image relations are presented in Figure 3.2. Examples of using this framework to code real-life subtitled content can be found in Section 4.2.2.

### *3.3.3 Analytical units of the framework*

To apply the four types of text-image relations into audiovisual multimodal analysis, it is crucial to determine a comparable unit. An improper standard for dividing units in the verbal and visual content can lead to inconsistency or even blunders in analysis. To determine proper analytical units, an analyst may encounter a series of questions. For example, which part of the text is related to which part of the image? Should the text be decomposed into as small as lexical elements, or as big as discursive units? How should the elements in an image be divided into comprehensible units? If it is pointless to analyze the smallest components in an image, i.e., pixels, then what is the reasonable size of visual components for analysis? If there is a person presented in an image, should the analytical units be the whole body of that person? If not, then what element of the person should be observed separately?

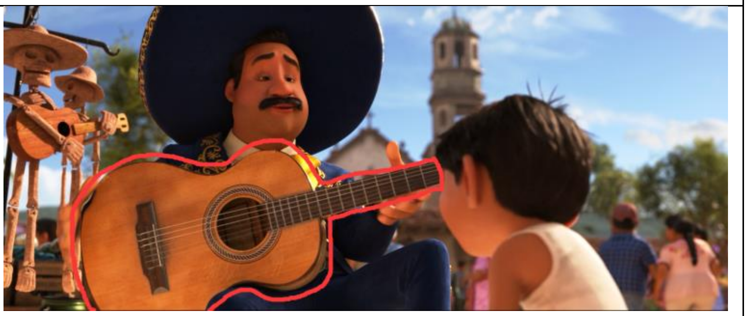
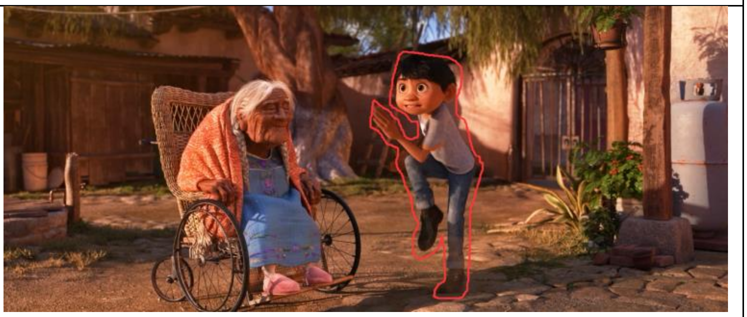
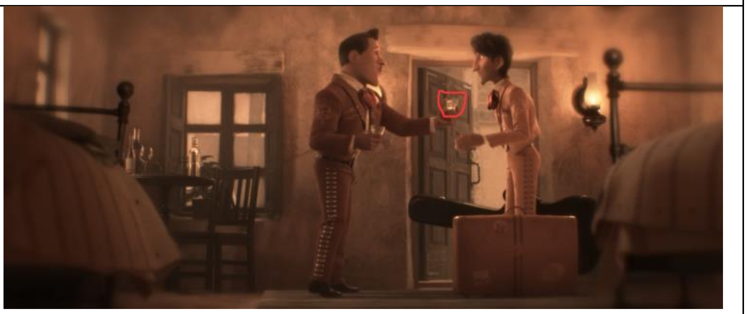
| **Text-image relations between different transitivity components** | **Examples** |
|---|---|
| Relations between *participants* |  |
| Concurrence (text = image) | He was just showing me **his guitar**. |
| Complementarity (text > image) | He was just showing me **his self-made guitar**. |
| Condensation (text < image) | He was just showing me **his instrument**. |
| Contradiction (text ≠ image) | He was just showing me **his violin**. |
| Relations between *processes* |  |
| Concurrence (text = image) | But now I **run** like this which is way faster. |
| Complementarity (text > image) | But now I **sprint** like this which is way faster. |
| Condensation (text < image) | But now I **do** like this which is way faster. |
| Contradiction (text ≠ image) | But now I **fly** like this which is way faster. |
| Relations between *circumstances* |  |
| Concurrence (text = image) | I'm sending you off **with a toast**. |
| Complementarity (text > image) | I'm sending you off **with a toast of Whisky**. |
| Condensation (text < image) | I'm sending you off **with this**. |
| Contradiction (text ≠ image) | I'm sending you off **with a flower**. |

**Figure 3.2** Conceptual examples of text-image relations adapted from dialogues in *Coco*

From the examples provided in the previous studies, it can be argued that the analytical units examined in previous frameworks have not been consistent. For instance, in the examples offered by Pastra (2008), the textual units can be words, phrases or sometimes the whole sentences. In general, the textual units in most of Pastra's examples are words. For instance, she takes the word "helmets" to illustrate its *type–token* relation with the image showing someone wearing a helmet (p. 309). But sometimes, a phrase, which is a larger unit than a word, is used, as in her example where the phrase "the President of Greece" and the image that shows Kostis Stephanopoulos are recognized as a *defining apposition* relation (p. 311). On the other hand, when Pastra demonstrates a *symbiosis* relation, a whole sentence, i.e., "so, what the women were doing then?", is taken as a unit to relate with the image (p. 313). Such inconsistent choices of textual elements indicate the necessity to further specify analytical units.

To determine the units of text in translation, Vinay and Darbelnet (1995) once defined the units of translation as the lexical elements that are combined to form a single element of thought, or in other words, "the smallest segment of the utterance whose signs are linked in such a way that they should not be translated individually" (p. 21). In this sense, during the process of subtitling, the translator is not dealing with individual words but thoughts and ideas in the source text. Thus, it is more suitable to examine the text on the semantic basis rather than a lexical unit. For example, in the sentence "you bet", the two words "you" and "bet" are combined to form one single meaning (or thought), i.e., "certainly". It would be erroneous to analyze "you" and "bet" separately during subtitling.

The task of specifying the units in images, on the other hand, is more complicated. As Pastra (2008) points out, images, by their very nature, "give much more descriptive information for real world entities than what is mentioned through speech/text" (p. 312), such as colors and shapes. It is impractical, and often meaningless, to focus on every detail of an image. A more practical way to determine the analytical element of an image, as Pastra suggests, is to look at the "complete body-movements", "the single objects, clusters of objects, foreground or background parts of images or whole images" (p. 317). Briefly put, it is rarely the case when a relation is formed by the whole text and the whole image (Bateman, 2014b). It is more often the specific components of the text (e.g., words) and the image (e.g., objects) that interweave to form a relation.

In this proposed framework, the basic unit for annotation of the text draws on the system of transitivity from Halliday's systemic functional grammar (Halliday & Matthiessen, 2014), and the analytical unit of the image is in respect with the transitivity system from visual grammar by Kress and Van Leeuwen (2021). The components of the verbal and visual transitivity analyses are threefold, involving *participants*, *processes*, and *circumstances* of both the image and the text. *Participants* refer to people and things, either concrete or abstract; *processes* represent actions or the state of being of the *participants*; *circumstances* concern the place where these actions or state of being occur or how they are performed (Kress & Van Leeuwen, 2021, p. 45; see also Section 3.2).

### *3.3.4 Intermodal direction of the framework*

To alleviate the complication of specifying a clear unit for analyzing text-image relations, an effective solution is to determine the analytical direction. So far, few researchers on text-image relations have addressed the issue of intermodal direction. Their frameworks tend to deal with the verbal and visual modalities in a two-way direction. For example, in the relation of *complementarity*, it can be either that the text complements the image or vice versa. This mutual analytical direction may double the researchers' analytical efforts because they have to look at the text and the image back and forth to identify possible text-image interplay. More importantly, the mutual analytical direction may bring up redundant identifications of relations. For example, when the text complements (>) the image, it can always be said in the other way around that the image is less specific than (<) the text. If we attempt to explore the text-image relations in a multimodal corpus, such mutual practice can be problematic by adding up unwanted frequencies. In the very context of subtitling, although it is technically possible to alter the visual elements on the screen, the linguistic content "continues to be the sole recipient of transposition" (Taylor, 2012, p. 26). A video can have various versions of subtitles, but the visual content is seldom changed. After all, subtitling always involves language, not exclusively so, but inevitably so.

In this sense, it is more efficient and assumedly adequate to adapt a unidirectional approach to the analysis of text-image relations. In other words, an analyst can focus on how the text interacts with the image, taking the text as the starting point of intermodal comparison (i.e., comparing the text with the image). Therefore, the proposed framework examines whether the text concurs with (=), complements (>), condenses (<), or contradicts (≠) the semiotic message in the image.

### *3.3.5 Applicable media of the framework*

The last but important factor for a framework of text-image relations is media. Subtitling is a work dealing with dynamic audiovisual media. In such a dynamic medium, the text-image relations are constantly in a state of flux. Unlike traditional static media such as textbooks and newspapers where the text and image are in a relatively fixed position and steady for analysis, audiovisual media features a dynamic flow of text and image, whose interplay is indefinite in terms of time and space. For example, a line of subtitle in a film may not refer to the objects in its synchronous frame but to the frame shown a few seconds earlier or even minutes later. Furthermore, a word may refer to an object that occurs on the screen for several times. It is of question whether this word should be related to only the first time the object is presented or to all the times. Such flowing spatial relativity and often asynchronous interaction between texts and images require a further specification of how to group the text and the relevant image in a definite (or relatively stable) set for observation. So far, not much attention in previous frameworks of text-image relations has been paid to this media-specific problem (see also Section 3.1.4). Researchers tend to tacitly assume the peripheral role of the problem. Even when they provide examples from dynamic texts, they analyze them in a way not much different from when they discuss a static text. One of the subtitling studies that address this issue is by Chen and Wang (2016), who acknowledge the subtitle's reference pointing back or forward to the visual elements in a whole film.

To address the problem of grouping texts and images in dynamic media, it is advisable to turn to theories of film art and look at the basic unit of a film narrative. The smallest unit of the visual narrative in films is a frame, which is a single still image on the film. One single second of content of a film can be composed of 24 frames. In a single frame, the text and the image are in a static position and thus their interaction is not much different from that in a static media. The sequence of 24 frames per second often eludes viewers' conscious perception. Instead, what viewers commonly apprehend as a cohesive element within the film's narrative structure is the larger unit known as a "shot". A shot in a finished film is "one uninterrupted image, whether or not there is mobile framing" (Bordwell et al., 2017, p. G-5). In other words, a shot is a collection of frames and is recorded from the time that a camera begins recording and then stops. In a film shot, the characters or objects are able to act or move and they start to form a more comprehensible narrative. However, most of the time, a complete narrative is formed by a larger unit, i.e., the scene. A scene is defined as "a segment in a narrative film that takes place in one time and space or that uses crosscutting to show two or more simultaneous actions" (Bordwell et al., 2017, p. G-5). That is to say, the film maker can use several shots to make a scene which can tell the viewer a coherent plot. In this sense, it seems appropriate for the proposed framework to group the text and the image in a scene and analyze their interaction in this narrative unit.

As the visual and textual elements are dynamic and constantly progressing within a scene, there may be a time lapse between the appearance of an element in the image and its corresponding text. This asynchrony makes it tricky and challenging to identify the text-image relation. To address this issue, the concept "**cumulative narrative**" is borrowed from Newcomb (2004). In this study, cumulative narrative refers to the situation when the new element in a scene relies on or makes references to the plots or characters that have occurred in previous narrative. In other words, the new narrative is continuously built upon previous narratives. This also holds true from the viewers' perspective. A previous study by Germeys and d'Ydewalle (2007) showed that film viewers tended to constantly perceive and make sense of the narrative continuity of a story. Their visual attentions were "almost entirely" directed towards "the most informative parts" of the narrative (Germeys & d'Ydewalle, 2007, p. 464) even when those core informative stimuli are moving or repositioned by shot transitions. It suggested that viewers' perception of the film's storyline is cumulative.

In this sense, a text-image relation can be identified based on the cumulative narrative even though the subtitle and the image are not presented simultaneously. For instance, in the case where a dog is saliently shown on the screen and the subtitle mentions "Dante" (the dog's name) a bit later, we can still argue that there exists a *concurrence* relation between the text and the image. This is because the text is directly connected to the ongoing cumulative narrative and the figure of the dog is, assumedly, already present in the viewers' mind.

### 3.4 A proposed framework of translation shifts in text-image relations for subtitling studies

The proposal of the present framework was inspired by the findings and challenges identified in the corpus study conducted in the project (see Chapter 4). Initially, the corpus annotation focused solely on text-image relations. To investigate shifts in these relations through subtitling, the author compared the total frequency of different text-image relations among the source and the target subtitles. However, this approach appeared to be problematic due to the presence of substantial noise in the comparison. For example, in some cases, a change in a text-image relation was not a result of the subtitler's consideration of the non-verbal elements during translation, but rather due to the difference in transitivity structures between Chinese and English clauses (Martin et al., 2023; see also Section 4.3.2 for more examples). To analyze translation shifts in a corpus, it is essential to take into account the systemic differences between languages (Cyrus, 2009).

To address these challenges and better understand the complex synergy of text-image interaction in AVT, this new framework is proposed. It takes a bottom-up approach and is based on actual cases of shifts observed in the film corpus. It incorporates five major types of shifts:

— ***Non-shift***: no change in text-image relations between the source and the target subtitles

— ***Obligatory shift***: inevitable change in text-image relations due to inherent differences between the source and the target language systems

— ***Preferential shift***: optional change in text-image relations caused by the stylistic preferences between the source and the target language systems

— ***Strengthening shift*** (***4 Es***: ***expansion***, ***explicitation***, ***enhancement***, and ***elaboration***): typical image-induced change in text-image relations by adding more specific or explicit information in the target subtitles to build a more relevant connection with the image

— ***Weakening shift*** (***4 Ds***: ***detachment***, ***diminution***, ***dilution***, and ***decrement***): typical image-induced change in text-image relations by making the information in the target subtitles more implicit or less specific and thus establishing a more distant connection with the visual information

The first three types of shifts are believed to be driven by the subtitler's linguistic consideration between the target and source languages, while the latter two are considered as more typical shifts resulting from the subtitler's conscious or unconscious attention to the non-verbal elements during translation.

The two image-induced shifts, *strengthening shift* and *weakening shift*, are further categorized into subtypes, as depicted in Figure 3.3. The *strengthening shift* comprises four subtypes (referred to as *4 Es*): *expansion*, *explicitation*, *enhancement*, and *elaboration*. These subtypes indicate the tendency of target subtitles to establish a closer connection with the visual information compared to source subtitles. The *weakening shift* also consists of four subtypes (labeled as *4 Ds*): *detachment*, *diminution*, *dilution*, and *decrement*. These subtypes describe the trend of target subtitles representing a looser interplay with the visual information compared to source dialogues.

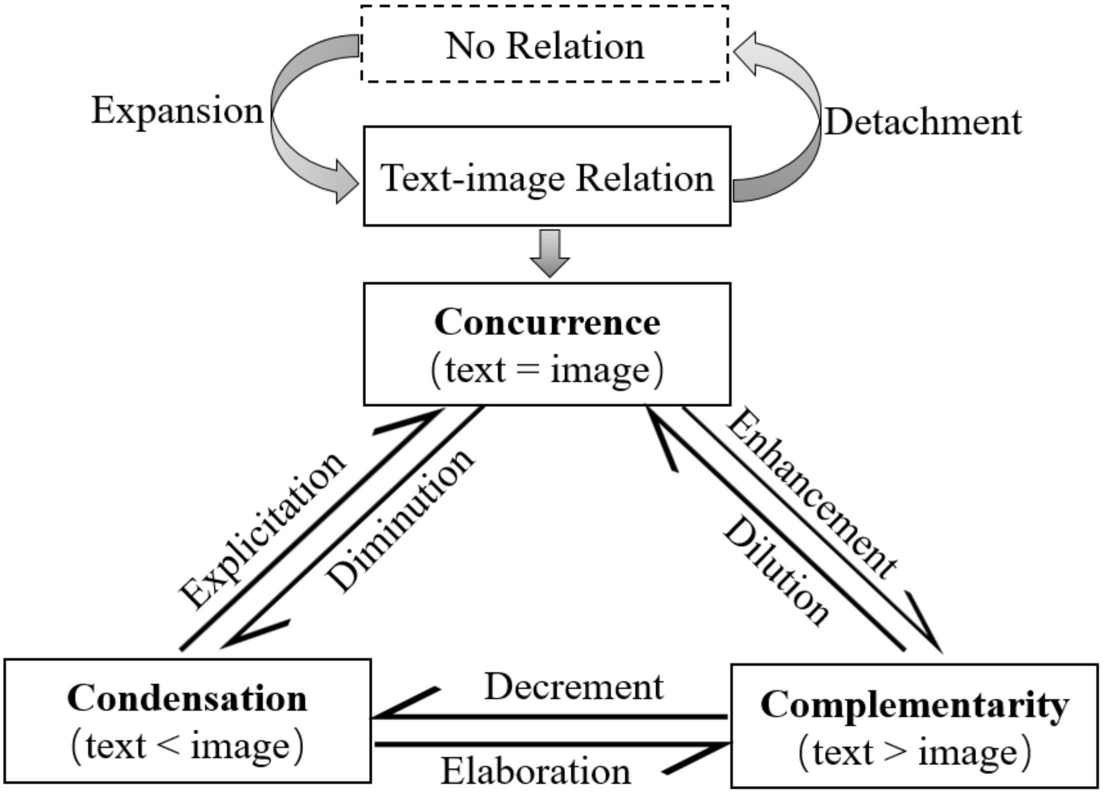


**Figure 3.3** A proposed framework of image-induced translation shifts (4 Ds & 4 Es) for subtitling studies

The identification of translation shifts within this framework is guided by the verbal and visual transitivity systems, which are also used to observe text-image relations in the study. In other words, the basic logic for the shift analysis is whether a target subtitle changes the original text-image relation within the *participants*, *processes*, or *circumstances* (for a more vivid example, see Section 3.4.1). Similar to Van Leuven-Zwart's (1990b, as cited in Cyrus, 2009) standpoint (see Section 2.1.4), the framework maintains a neutral attitude towards translation shifts. It does not prescribe what subtitlers should or should not do, but rather describe what shifts have been made. It specifically emphasizes the divergences between a source text-image relation and a target text-image relation. From this perspective, a translation shift in interlingual subtitling is not solely a linguistic shift, but also involves the subtitler's conscious or unconscious alteration of the multimodal layout of the audiovisual content. In the following subsections, each type of translation shift within the proposed framework will be explained.

### *3.4.1 Non-shifts*

*Non-shifts* refer to situations where the target subtitle maintains the original text-image relation by reproducing the same experiential meaning of the transitivity components conveyed in the source dialogues. Given that translation shifts can occur at various levels, ranging from small semantic elements to larger changes in language systems (Pekkanen, 2007), it is crucial to establish a clear **unit of comparison** for the text-image relations in the framework. In this study, the unit of comparison for non-shifts (as well as for other shifts in the framework) revolves around the experiential meaning of the subtitle, rather than the specific linguistic form of the languages involved. In other words, the comparison is made based on the verbal/visual *participants*, *processes,* and *circumstances*, rather than the linguistic structures such as part of speech and syntax.

For example, supposedly in a film scene there is a little girl holding a ticket in her hand and a boy standing beside her. The boy is threatening to take the girl's ticket with his arm outstretched and says, "Give me the ticket." As for the target Chinese subtitle, the translation goes as "票给我" [Ticket give me]. To analyze this example from a purely linguistic perspective, if we do not restrict the unit of comparison, at least two shifts can be identified. The first shift is the change of syntax. While the source text represents a normal "predicate (verb) + indirect object + direct object" structure, the target subtitle establishes a typical inverted structure of "direct object + predicate (verb) + indirect object". The other shift is the change in grammatical elements, as the definite article "the" in the source English is omitted in the target Chinese subtitle.

However, if this example is analyzed based on the unit of comparison in this framework, i.e., transitivity components, no shifts can be found. To determine if there are any shifts in the text-image relations, we first compare the verbal/visual *participants, process,* and *circumstances* between the source dialogue and the target subtitle. In the source subtitle, there are three explicit transitivity components, i.e., one *process* ("give") and two *participants* ("me" and "the ticket"). These three components establish three instances of *concurrence* relation to the visual transitivity components in the image, i.e., the boy's gesture of stretching out the hand to take something as the concurrent *process,* the boy himself as the concurrent *participant,* and the ticket as the other concurrent *participant*. Although the sentence structure of the target Chinese subtitle differs from that of the source English, it still conveys the same representational meaning of the original transitivity components, with "给" [give] as the *process*, and “票” [ticket] and “我” [me] as the two *participants*. Therefore, no shifts in text-image relations are identified in this case, although there are some micro-levels of shifts regarding the linguistic structure. If a target subtitle generally conveys the same representational meaning as the original dialogue, then it can be considered a case of *non-shift*.

Strictly speaking, *non-shifts* may not truly be considered *shift*, just as non-fiction is the antithesis of fiction. Nevertheless, some researchers on translation studies still regard *non-shifts* as a type of translation shifts (e.g., Calzada Pérez, 2007). Similarly, in this proposed framework, non-shifts are classified as a type of shifts. In most previous corpus-based shift analyses, little attention has been given to the cases of *non-shifts* (e.g., Qian & Feng, 2020), with the undertone that only shifts are the marker of translation problems and thus worthy of examination. Although *non-shifts* are not the primary focus of this corpus-based study, identifying *non-shifts* and regarding them as a type of translation shifts can provide valuable indirect evidence for the extent to which a subtitler considers non-verbal elements during the subtitling process. The analysis of *non-shifts* can also shed light on the strategies employed by subtitlers to tackle non-verbal information during translation.

### *3.4.2 Obligatory shifts*

An *obligatory shift* refers to the unavoidable alteration of the original text-image relation when translating subtitles from the source language to the target language, due to inherent structural differences between the two languages. This term has been conceptualized similarly by Blum-Kulka (1986/2000, p. 312), Calzada Pérez (2007, p. 150) and Liao (2011, p. 351).

*Obligatory shifts* are primarily triggered by the absence of certain linguistic elements in different language systems. In the case of English and Chinese, there are properties in English, such as verb tense and definite articles, that do not have direct equivalents in Chinese. As a result, these elements can only be omitted or transformed in the translation process.

For example, when translating the spoken English sentence “Where were you?” into Chinese, if the subtitler wants to convey the past tense of the verb, the translation could be “你刚刚去哪了?” [You just now go where?]. Here, the original past tense in the English verb is rendered as “刚刚” [just now], which functions as a temporal adverbial phrase. This addition of a new transitivity component, a *circumstance*, in the target subtitle creates a new text-image relation. Based on the unit of comparison used in this study, this newly added component warrants further analysis.

However, this shift in this example is induced by specific language-pair grammatical and structural differences between English and Chinese. With the focus on shifts induced by the presence of visual information in films, *obligatory shifts* should be left out from the focal analysis in this study, since they "do not involve choice on the part of translators" (Liao, 2011, p. 352). *Obligatory shifts* in interlingual subtitling are more passive outcomes of the language-pair disparity, rather than the subtitler's attention to the visual elements of the mise-en-scène. Therefore, when calculating shifts made by the subtitler's awareness of the co-occurring visual information during the subtitling process, the annotation of *obligatory shifts* in the corpus can minimize the noise of analysis. Nevertheless, the identification of *obligatory shifts* also serves as a reminder that "translation still involves language – not exclusively so, but undeniably so" (Cyrus, 2009, p. 88). It would be risky for researchers analyzing AVT from a multimodal perspective to overemphasize the role of non-verbal elements in the subtitling process.

### *3.4.3 Preferential shifts*

Following Liao's (2011) definition, *preferential shifts* are identified in this study when the target subtitle alters the original text-image relation due to the constraints imposed by the norms of the target language. *Preferential shifts* are different from *obligatory shifts*, as the former are norm-governed while the latter are structure-dependent. *Obligatory shifts* serve to make the target subtitles grammatically correct and comprehensible, whereas *preferential shifts* aim to make the target texts read stylistically conventional and acceptable to the target readers. In some instances, it is possible to construct a grammatically correct sentence in the target language that adheres to the rules of the source language. but, as suggested by Klaudy (2009), the overall text will appear "clumsy and unnatural" (p. 106).

For instance, when translating "Give me a smile" from English into Chinese, the most direct transfer would be "给我一个微笑" [Give me a smile]. While this translation is grammatically correct, it sounds unnatural in Chinese. A more conventional way to express the same idea would be "给我笑一个" [Smile once for me]. In this alternative version, there is a shift in the word class of "smile" from a noun to a verb, resulting in a more natural and stylistically appropriate target text in Chinese. Through a transitivity analysis, we can also observe that the *participant* ("smile" as an object) in the source text is transformed into a *process* ("smile" as an act) in the target text. This omission of a *participant* in the source text leads to a reduction in the textual relation with the visual information. Conversely, the addition of a *process* in the target text establishes a new relation with the visual mode. Both of these shifts in text-image relations are considered *preferential shifts* in this study.

Similar to *obligatory shifts*, *preferential shifts* are not deemed as indicator of the subtitler's consideration of non-verbal elements during the translation process. Hence, they are more considered as language-induced shifts in text-image relations. The purpose of identifying *preferential shifts* is to reduce the noise in the results of the focal analysis.

### *3.4.4 Strengthening shifts (4 Es)*

In previous research on interlingual subtitling, the term "strengthen" has been employed to describe the way in which target subtitles enhance a particular aspect of the original audiovisual content, such as the enhancement of the film narrative (Remael, 2003) and the intensification of source taboo language (Chen, 2022). In this framework, strengthening refers to the tendency of the target subtitle to have a stronger semantic relationship with the visual elements. *Strengthening shifts* occur when the target subtitle, compared with the source subtitle, establishes a closer or more relevant interaction with the image. This tendency can be seen as a typical sign that the subtitler consciously or unconsciously considers non-verbal elements during the translation process. Based on the cases observed in the multimodal corpus, *strengthening shifts* can be further categorized into four types: *expansion*, *explicitation*, *enhancement*, and *elaboration* (*4 Es*).

*Expansion* refers to the addition of a new text-image relation in the target subtitle that is not present in the source text. By introducing an additional transitivity component in the target subtitle, the *expansion* shift can create a new relation of *concurrence*, *complementarity*, or *condensation* with the corresponding visual transitivity component. While *expansion* shifts concern the creation of new text-image relations during subtitling, the other three types of *strengthening* shifts involve transforming an existing text-image relation into another form.

*Explicitation* occurs when the target subtitle changes the original *condensation* relation (text < image) into a relation of *concurrence* (text = image). As explained in Section 3.3.2, a *condensation* relation suggests that the information conveyed in the subtitle (or a certain part of it) is less specific than the information in the image (or a part of it), whereas a *concurrence* relation indicates that the subtitle and the image (or parts of both) contain approximately the same amount of representational meaning. Sometimes, the *condensation* relation in the source text may undergo an *explicitation* shift in the target text, resulting in a *concurrence* relation and providing more explicit information in conjunction with the visual content.

*Enhancement* refers to a type of change in which the target subtitle alters the original *concurrence* relation (text = image) and transforms it into a relation of *complementarity* (text > image). As discussed in Section 3.3.2, in a *complementarity* relation, the subtitle, or a portion of it, offers additional illustrative or explanatory information about the image. When an *enhancement* shift takes place, the target subtitle modifies the *concurrence* relation in the source text and establishes a *complementarity* relation by providing more specific information relevant to the image.

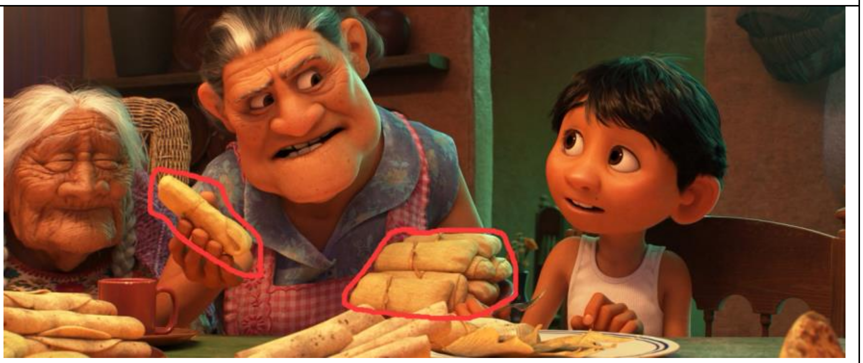
| **Strengthening shifts** | **Examples** | |
|---|---|---|
| |  | |
| | ST | TT |
| *Expansion*<br>(e.g., no relation → concurrence) | I asked if you would like to. | 我问你要不要一些玉米粽子<br>[I asked if you would like **some tamales**.] |
| *Explicitation*<br>(condensation → concurrence) | I asked if you would like **some**. | 我问你要不要一些玉米粽子<br>[I asked if you would like **some tamales**.] |
| *Enhancement*<br>(concurrence → complementarity) | I asked if you would like **some tamales**. | 我问你要不要**一些刚做好的玉米粽子**<br>[I asked if you would like **some freshly made tamales**.] |
| *Elaboration*<br>(condensation → complementarity) | I asked if you would like **some**. | 我问你要不要**一些刚做好的玉米粽子**<br>[I asked if you would like **some freshly made tamales**.] |

**Figure 3.4** Conceptual examples of *strengthening shifts* across *participants* in *Coco*

*Elaboration* is a phenomenon where the target subtitle modifies the original text-image relation of *condensation* (text < image) and instead establishes a relation of *complementarity* (text > image). In certain cases, while the meaning of the source dialogue is less specific than the visual content (*condensation*), the target subtitle provides additional information that goes beyond what is conveyed in the image, leading to an elaboration shift.

Overall, when *strengthening shifts* occur in interlingual subtitling, the target subtitle conveys more visually related information and represents more meaning than the source dialogues. The complex transition between different text-image relations is illustrated in Figure 3.3 at the beginning of Section 3.4. It is also important to note that this proposed framework of shifts does not include the relation of *contradiction*, as this type of relation is rarely observed in the corpus (see Section 4.2.1).

Some conceptual examples of different types of *strengthening* shifts are adapted from the translation of the original dialogues in *Coco*, as shown in Figure 3.4. Examples of using this framework to code real-life subtitled content can be found in Section 4.3.2.

### *3.4.5 Weakening shifts (4 Ds)*

*Weakening shifts*, positioned in the opposite direction of *strengthening shifts*, refer to the tendency of the target subtitles to create a looser or more distant connection with the visual information (illustrated in Figure 3.4). This is achieved by reducing certain representational meaning from the source text. Subtitlers employ *weakening shifts* on the assumption that viewers can decipher the original meaning from the visual cues (Gottlieb, 2001). Based on empirical instances observed in the multimodal corpus, *weakening shifts* can also be categorized into four types: *detachment*, *diminution*, *dilution*, and *decrement* (*4 Ds*).

*Detachment* refers to the omission of an existing text-image relation from the source text. This type of shift may occur when a particular transitivity component in the original dialogue is excluded in the target subtitles, resulting in the omission of the original corresponding text-image relation such as *condensation*, *concurrence*, and *complementarity*. *Detachment* is the opposite counterpart of *expansion* within this framework. While the *detachment* shift involves the complete exclusion of a specific text-image relation during subtitling, the other three types of *weakening shifts* are about altering an existing text-image relation into another.

*Diminution* refers to the change where the target subtitle alters the original *concurrence* relation (text = image) and transforms it into a *condensation* relation (text < image). This is achieved by replacing some originally explicit information in the source text with less specific content in the target subtitle. *Diminution* can be seen as the opposite counterpart of *explicitation* within this framework.

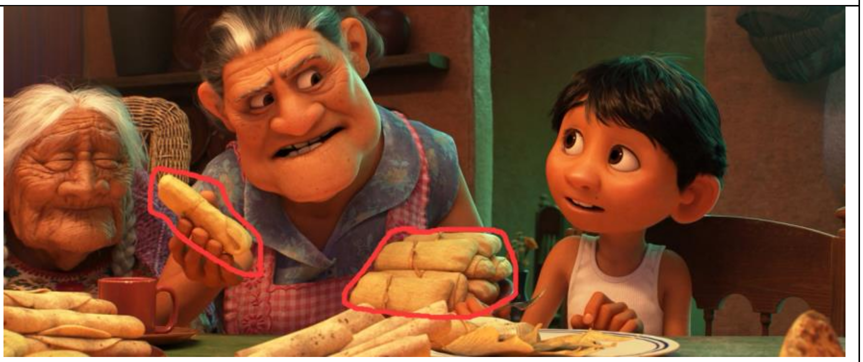
| **Weakening shifts** | **Examples** | |
|---|---|---|
| |  | |
| | ST | TT |
| *Detachment*<br>(e.g., concurrence → no relation) | I asked if you would like **some tamales**. | 我问你要不要<br>[I asked if you would like.] |
| *Diminution*<br>(concurrence → condensation) | I asked if you would like **some tamales**. | 我问你要不要**一些**<br>[I asked if you would like **some**.] |
| *Dilution*<br>(complementarity → concurrence) | I asked if you would like **some freshly made tamales**.] | 我问你要不要**一些玉米粽子**<br>[I asked if you would like **some tamales**.] |
| *Decrement*<br>(complementarity → condensation) | I asked if you would like **some freshly made tamales**.] | 我问你要不要**一些**<br>[I asked if you would like **some**.] |

**Figure 3.5** Conceptual examples of *weakening shifts* across *participants* in *Coco*

*Dilution* occurs when target subtitle changes the original *complementarity* relation (text > image) into a relation of *concurrence* (text = image). In this case, the target subtitle downgrades the specificity of the representational meaning in the source text, confining the textual specificity to the visual one. *Dilution* is the opposite counterpart of *enhancement* within this framework.

*Decrement* refers to the case where the target subtitle modifies the original text-image relation of *complementarity* (text > image) and instead establishes a relation of *condensation* (text < image). When a *decrement* shift takes place, the target subtitle extensively reduces the specificity of the meaning conveyed in the source text by providing less specific information related to the image. *Decrement* is the opposite counterpart of *elaboration* within this framework.

In general, in the case of *weakening shifts*, the target subtitle reduces the specificity of the representational meaning of the source dialogue and provides less visually relevant information. Similar to *strengthening shifts*, the categorization of *weakening shifts* does not consider the relation of *contradiction*, given the rare instances of this type of text-image relation in the corpus (see Section 4.2.1).

Figure 3.5 presents some examples of *weakening* shifts adapted from the translation of the original dialogues in *Coco*. Further examples of using this framework to code real-life subtitled content can be found in Section 4.3.2.

## 3.5 Summary

Two theoretical frameworks are proposed in this chapter to account for the dynamic interplay between subtitles and image in AVT. The first framework is used to analyze text-image relations in subtitled films. It is adapted from previous frameworks by Martinec and Salway (2005), Unsworth (2006, 2007), and Pastra (2008), with four major categories of text-image relations (4 Cs): *Concurrence* (text = image), *Complementarity* (text > image), *Condensation* (text < image), and *Contradiction* (text ≠ image). The other framework focuses on translation shifts in text-image relations specifically within the context of interlingual subtitling. This framework is formulated through a bottom-up approach, drawing on real-life subtitling cases. It comprises five major categories: *non-shifts*, *obligatory shifts*, *preferential shifts*, *strengthening shifts* (*4 Es*: *expansion*, *explicitation*, *enhancement*, and *elaboration*), and *weakening shifts* (*4 Ds*: *detachment*, *diminution*, *dilution*, and *decrement*). The first three types of shifts are believed to be language-induced, driven by the subtitler's linguistic consideration between the target and source languages. The latter two types are considered as image-induced shifts, arising from the subtitler's conscious or unconscious attention to the non-verbal elements during the translation process. The two frameworks serve as the theoretical foundation for the corpus study and the eye-tracking experiment in this project.

# CHAPTER 4 THE MULTIMODAL CORPUS-BASED STUDY

Chapter 4, addressing RQ1 and RQ2, provides an analysis of a self-constructed multimodal corpus of subtitled films. The chapter begins by introducing the methodological considerations (Section 4.1), such as the selection of corpus materials (English films with target Chinese subtitles), the size of the corpus (30 scenes sampled from ten popular and highly-regarded English films among Chinese viewers), and the annotation scheme employed for corpus analysis (based on the two theoretical frameworks proposed in this thesis). Subsequently, Section 4.2 presents the quantitative and qualitative outcomes related to the frequency distribution of text-image relations in the corpus. Section 4.3 then provides the results concerning translation shifts in text-image relations within the corpus. Building upon these findings, Section 4.4 engages in an overarching discussion on the role of subtitling as a multimodal representation in AVT.

## 4.1 Methodology

### *4.1.1 Corpus design and materials*

The advantages and feasibility of multimodal corpora of audiovisual content have been acknowledged and proved in previous research (Bonsignori, 2018; Desilla, 2014; Ramos Pinto & Mubaraki, 2020; Soffritti, 2019; see Section 2.2.1 for more details). Despite its notoriously time-consuming nature and the current dearth of unanimously accepted annotation schemes (Abuczki & Ghazaleh, 2013), a multimodal corpus of AVT can provide quantitative evidence on how the translated texts serve in the multimodal communication with other non-verbal meaning-making resources in the audiovisual product. Moreover, with many annotation tools currently available for the compilation of multimodal corpora such as ELAN (Wittenburg et al., 2006) and Anvil (Kipp, 2001), researchers can, in a more efficient way, pre-process, align, and annotate the textual and non-verbal content in the corpus.

In this light, the multimodal corpus in this study is compiled from some popular and well-received English films with both the source English subtitles and the target Chinese subtitles (see Appendix 1). The aim of the corpus is to investigate the text-image relations in films and to examine how those relations may be changed through subtitling. To systematically select films for the corpus, an inclusion criterion was adapted from Fox (2018), where films that had gained success in both box office and public reviews between 2000 and 2020 were selected from four top-250 lists of the most influential films:

- Top-grossing English films in North America (US and Canada) (on Box Office Mojo[1])
- Top-grossing English films in China (on Maoyan Entertainment[2])
- Top-rated English films among international viewers (on IMDb[3])
- Top-rated English films among Chinese viewers (on Douban[4])

The four lists were chosen to represent films that were successful in both the source-language market and the target language market. The 250 top-grossing English films in North America were based on the list provided by Box Office Mojo, an American website that systematically tracks box-office revenue of films. This list presented films that had achieved success in the source-text market. The 250 top-grossing English films in China were selected from the list offered by a Chinese online ticketing service provider Maoyan Entertainment. This list ranked blockbuster films in the target-text market. The 250 top-rated films among international viewers were chosen from the list of "IMDb Top 250 Movies", where films were ranked according to their ratings made by millions of users around the world. The 250 top-rated films among Chinese viewers are selected from the list of "Douban Top 250 Movies" offered by Douban, one of the most popular Chinese social networking websites where registered users can make comments on film, books, and music. A threshold of 250 was made for the inclusion criterion on the ground that it was the total number of films that were available in the two lists regarding viewers' ratings.

Based on Fox's (2018) criteria, each of the films in the four top-250 lists was given a score ranging from 1 to 250 according to its rank in the list (e.g., the film ranked first in the list was assigned a score of 250 and the one ranked second with a score of 249 and so forth). Then, a mean score was counted for each of the films. For example, the scores of the film *The Dark Knight* in the four lists were 239, 0, 247, and 224, so its mean score was 177.5 (derived from dividing the sum of the scores by four).

[1] https://www.boxofficemojo.com/chart/top_lifetime_gross/?ref_=bo_cso_ac [2021-12-31].

[2] https://piaofang.maoyan.com/rankings/year [2021-12-31].

[3] https://www.imdb.com/chart/top?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=470df400-70d9-4f35-bb05-8646a1195842&pf_rd_r=FQM2H4QBJFGEN203G47J&pf_rd_s=right-4&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_ql_3 [2021-12-31].

[4] https://movie.douban.com/top250 [2021-12-31].

Subsequently, the 100 English films released between the years 2000 and 2020 with the highest mean scores were selected as the data source of the corpus. There were three films that were ranked 100th, all of which had the same mean scores. These films were included in the corpus, resulting in a final selection of 102 films. It should be noted that the author in this study did not analyze all the 102 listed films but only the Top 10 films due to limitations in resources and time. However, the full list can serve as a valuable reference for future research in the fields of AVT or film studies.

The audiovisual content and subtitles chosen for the corpus analysis were the versions provided by Tencent Video, one of the most popular video streaming platforms among the Chinese audience. Although it would have been ideal to analyze the official target subtitles provided at cinemas, they were not accessible to the author of this study. Another viable option would have been the DVD versions officially released and circulated in the target market. However, over the past 20 years, DVDs have been increasingly overtaken by the rise of video streaming (Díaz Cintas & Remael, 2021). In recent years, there has been a growing interest among AVT researchers in interlingual subtitling on video streaming platforms like Netflix (e.g., Kuscu-Ozbudak, 2022; Pedersen, 2018), as streaming service providers are becoming “the most influential force driving subtitling norms” (Pedersen, 2018, p. 81). Therefore, Tencent Video is considered a suitable source of AVT materials for the present study.

### *4.1.2 Corpus size and compilation*

Balancing the resources available to the author and the research objectives, the corpus analysis in this study focused on sampled scenes from the Top 10 English films among Chinese audience based on the four rankings explained in Section 4.1.1. The sampling criteria were as follows: 1) the length of the scene should be about two minutes, given that in modern films “most scenes run between 1.5 and 3 minutes” (Bordwell, 2006, p. 57); 2) the selected scenes should primarily focus on human conversations rather than non-verbal representation such as singing, dancing, or fighting; 3) the majority of conversations in the scene should involve at least one main character in the film such as the protagonist or antagonist. Based on these criteria, a total of about six minutes of AVT content was sampled from each of the Top 10 films. The sampled content comprised three scenes (each lasting about two minutes) evenly distributed from the beginning, the middle, and the end of each film. Taking scenes from different places in the film can ensure the balance of samples in the corpus (Baker, 2018).

The resultant multimodal corpus was composed of 30 scenes sampled from the Top 10 films, including the original audiovisual content, the source English subtitles, and the target Chinese subtitles provided by the widely used streaming platform Tencent Video. The corpus was deemed to be reasonably representative in terms of timespan, film genre, and viewership, as presented in Table 4.1 (for a further discussion on potential limitations of the corpus, see Section 6.3). The frequency of text-image relations and the translation shifts observed in the corpus will constitute the basis for the design of the follow-up experiments, which are described in Chapter 5.

**Table 4.1** Annotated scenes from top 10 films in the corpus

| Rank | English & Chinese title | Year | Views on Tencent Video | Annotated duration | English words | Chinese words | Lines of subtitles |
|---|---|---|---|---|---|---|---|
| 1 | *The Dark Knight*<br>《蝙蝠侠：黑暗骑士》 | 2008 | 110 million | 6m 21s | 820 | 1391 | 113 |
| 2 | *The Lord of the Rings: The Return of the King*<br>《指环王 3：王者无敌》 | 2003 | 56 million | 6m 26s | 232 | 353 | 58 |
| 3 | *Avengers: Endgame*<br>《复仇者联盟 4：终局之战》 | 2019 | 420 million | 6m 32s | 573 | 886 | 119 |
| 4 | *Avengers: Infinity War*<br>《复仇者联盟 3：无限战争》 | 2018 | 260 million | 7m 22s | 901 | 1503 | 143 |
| 5 | *Inception*<br>《盗梦空间》 | 2010 | 71 million | 6m 35s | 531 | 820 | 82 |
| 6 | *Avatar*<br>《阿凡达》 | 2010 | 70 million | 6m 55s | 377 | 595 | 61 |
| 7 | *Coco*<br>《寻梦环游记》 | 2017 | 150 million | 6m 30s | 747 | 1151 | 124 |
| 8 | *Zootopia*<br>《疯狂动物城》 | 2016 | 300 million | 6m 20s | 809 | 1290 | 121 |
| 9 | *The Lord of the Rings: The Two Towers*<br>《指环王 2：双塔奇兵》 | 2002 | 24 million | 6m 42s | 431 | 746 | 76 |
| 10 | *Interstellar*<br>《星际穿越》 | 2014 | 200 million | 6m 29s | 508 | 745 | 91 |
| Total | | | | 67m 35s | 5929 | 9480 | 988 |

### *4.1.3 Corpus annotation*

The annotation scheme for this multimodal corpus is based on the theoretical frameworks proposed in Section 3.3 and Section 3.4. It synthesizes previous insights on multimodal functional grammar, multimodal relations and descriptive translation studies. As shown in Figure 4.1, the annotation scheme consists of three dimensions.

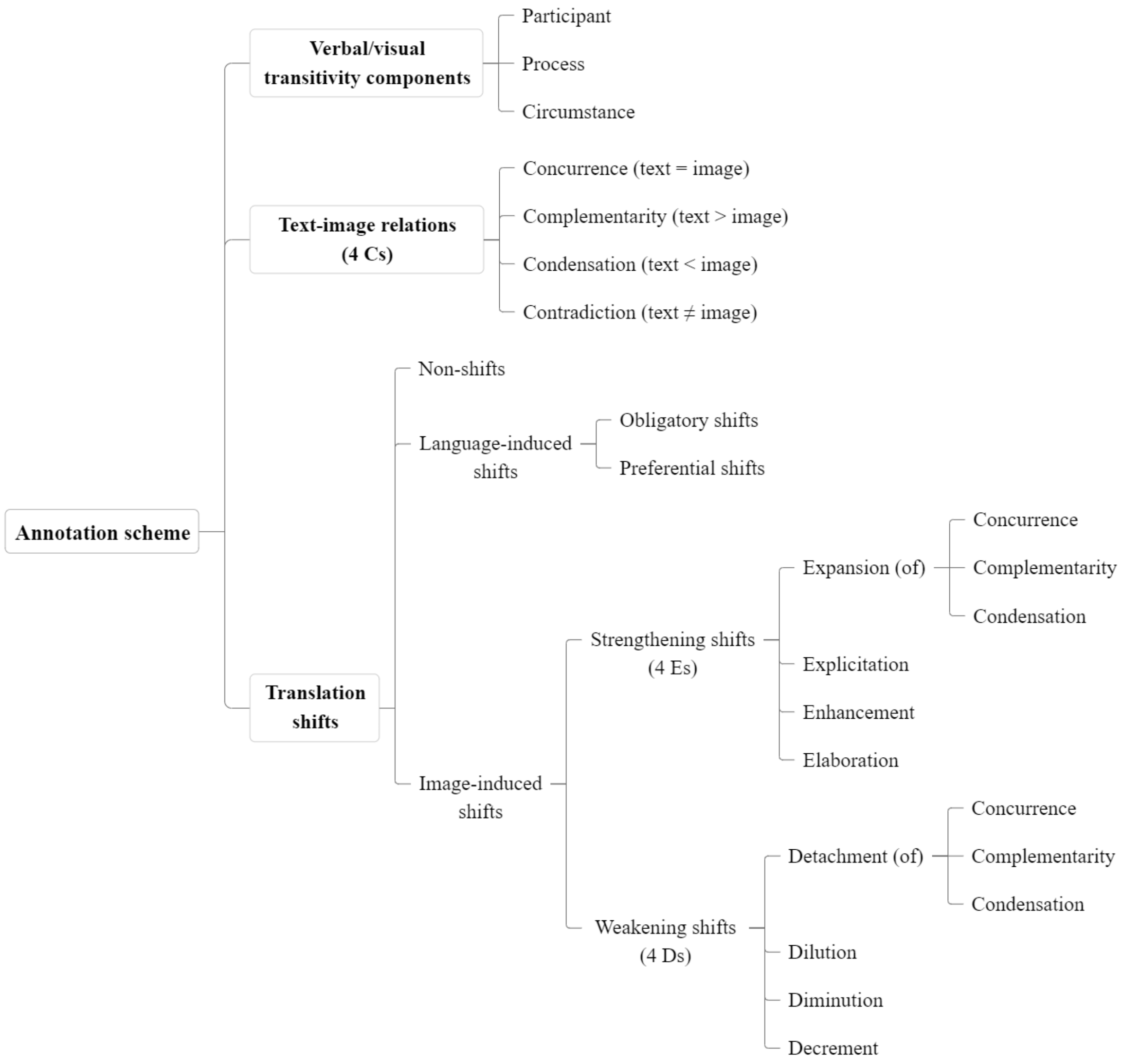
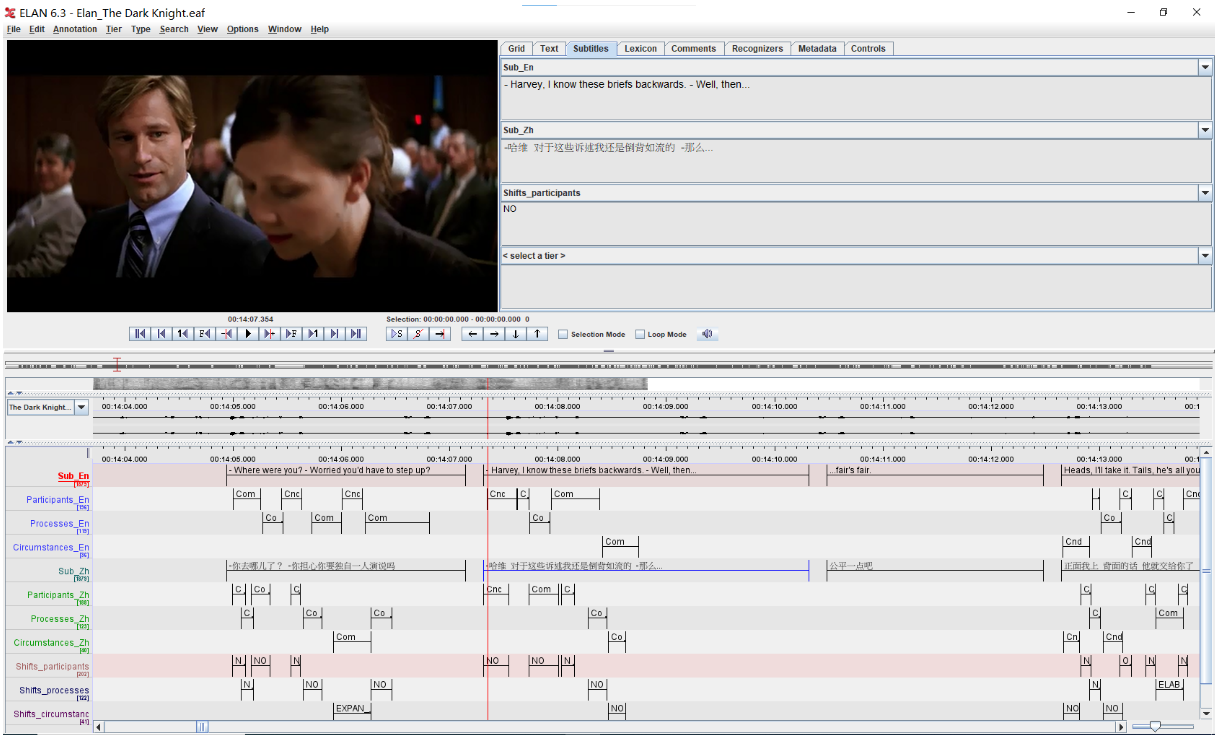


**Figure 4.1** Annotation scheme for text-image relations in the corpus of subtitled films

The first dimension concerns the analytical units of the text and the visual content, which draws on the system of transitivity in Halliday's systemic functional grammar (Halliday & Matthiessen, 2014) and the transitivity system in visual grammar by Kress and Van Leeuwen (2021). The components of the verbal and visual transitivity analyses are threefold, involving *participants*, *processes*, and *circumstances* of both the image and the text. *Participants* refer to people, places, and things, either concrete or abstract; *processes* represent actions or the state of being of the *participants*; *circumstances* concern the place where these actions or state of being occur or how they are performed (see Section 3.2 and 3.3.3). The second dimension specifies the text-image relations proposed in Section 3.3, i.e., *concurrence*, *complementarity*, *condensation*, and *contradiction*. The third dimension indicates the translation shifts in text-image relations through subtitling, as proposed in Section 3.4. Coding examples of the corpus entries can be found in Section 4.2.2 and 4.3.2.

All the tags in the annotation scheme were manually coded using the video annotation software ELAN (Wittenburg et al., 2006). The coding system created in the software was multi-leveled, comprising eleven tiers, as shown in Figure 4.2 (the name of each tier is displayed on the left).



ELAN 6.3 - Elan_The Dark Knight.eaf

Sub_En: - Harvey, I know these briefs backwards. - Well, then...

Sub_Zh: -哈维 对于这些诉述我还是倒背如流的 -那么...

Shifts_participants: NO

Selection: 00:00:00.000 - 00:00:00.000 0

Sub_En: - Where were you? - Worried you'd have to step up? | - Harvey, I know these briefs backwards. - Well, then... | ...fair's fair. | Heads, I'll take it. Tails, he's all you

Sub_Zh: -你去哪儿了？ -你担心你要独自一人演说吗 | -哈维 对于这些诉述我还是倒背如流的 -那么... | 公平一点吧 | 正面我上 背面的话 他就交给你了

**Figure 4.2** A screenshot from an annotation session in ELAN

To begin the annotation process, all the sampled scenes from the same film were imported into an ELAN project as the visual content for annotation. The English and Chinese subtitles of the scenes were then imported as the first two tiers in the corpus. These two tiers contained all the textual content of the films for analysis. Afterward, under each subtitle tier were added three tiers regarding the three major transitivity components as the analytical units. Different tags of text-image relations were then manually coded into the tiers accordingly. Below these eight tiers, three additional tiers were added to analyze the translation shifts through subtitling. To prevent spelling mistakes in the manual annotation, a set of controlled vocabularies was created in the software. This allowed the annotator to select codes from the vocabulary list instead of typing them manually.

It is worth noting here that the annotation scheme only includes the three most general transitivity components and does not go into more elaborate transitivity categories (e.g., material and relational processes). Ideally, it would be reasonable to analyze the transitivity features by detailed categories. However, the present study does not attempt to seek how the transitivity features reveal the discursive characteristics of the depicted characters, which would require a granular analysis of transitivity categories. Instead, the research focus of the study is the convergence or divergence between verbal and visual representations. It is thus assumed that the general model is sufficient and more manageable. Nevertheless, this does not imply that a more refined categorization is devoid of interest (for a comprehensive linguistic application of the transitivity system to interlingual subtitling, see Noverino et al., 2020).

To ensure coding reliability, two coders were involved in the annotation process. The first coder, the author of this thesis, annotated all the selected subtitled scenes. The other coder, the author's chief academic supervisor, was introduced to the coding scheme and then independently annotated the three scenes from the film *Interstellar*, which accounted for approximately 10% of the coded data. The perfect inter-coder agreement reached 80.97%, indicating the reliability of the manual coding and the validity of the coding scheme. Coding discrepancies mainly arose from multiple possible interpretations of text-image interplay in the film. For example, in a shot where water flooded into a spacecraft, a subtitle line went as "The engines are flooded!". The first coder interpreted this case as establishing an instance of *complementarity* relation between the verbal and visual *participants* ("the engine"), as the engine was not distinctly displayed on the image and thus the text provided additional information for the image. However, the second coder regarded this case as *concurrence*, believing that the text directly corresponded to the engine that was vaguely shown in the background together with other equipment in the spacecraft.

Upon completion of the annotation, a tab-delimited text file was exported from ELAN for the calculation of frequency distributions of each type of text-image relations and translation shifts identified by the first coder. The results can show which text-image relations are more commonly found in the subtitled scenes and the extent to which the text-image relations are preserved, added, omitted, or altered through subtitling.

## 4.2 Analysis of text-image relations in subtitled films

The results derived from the annotated data in the corpus were twofold, regarding (a) the frequency distribution of different text-image relations in the subtitled films (RQ1), and (b) the translation shifts in text-image relations through interlingual subtitling (RQ2) (see Section 1.2).

### *4.2.1 Quantitative results of text-image relations*

Table 4.2 shows the frequency distribution of different verbal-visual relations from the annotation of more than 60 minutes of audiovisual content in the corpus. The most frequently occurring relation was *complementarity* (text > image), with a total of 1,347 instances in the source subtitles. The second most commonly observed relation was *concurrence* (text = image), with a slightly lower count of 1,272 instances in the source subtitles. The *condensation* relation (text < image) was not very common, totaling 184 cases in the source subtitles. The relation of *contradiction* (text ≠ image), on the other hand, were rarely found in the corpus, with only 3 cases in the source subtitles. The text-image relations among the target subtitles also showed the same distributions.

**Table 4.2** Frequency of different text-image relations in the corpus

| **Text-image relations (4 Cs)** | **Source subtitle** | **Target subtitle** | **Total** |
|---|---|---|---|
| Complementarity (text > image) | 1347 | 1365 | 2712 |
| Concurrence (text = image) | 1272 | 1240 | 2512 |
| Condensation (text < image) | 184 | 148 | 332 |
| Contradiction (text ≠ image) | 3 | 3 | 6 |

The results suggest at least two patterns of text-image interaction in subtitled films. Firstly, the linguistic content (i.e., dialogues) in the films appeared to be more semantically specific than that of the visual content (i.e., images), as the frequency of *complementarity* was more than seven times that of *condensation*. Moreover, in a representational sense, the linguistic and visual content in the films tended to be closely comparable, as evidenced by the substantial number of *concurrence* relations identified. These patterns of text-image interaction shed light on the potentially complex dynamics between language and visuals in AVT.

### *4.2.2 Qualitative analysis of text-image relations*

This section provides a range of qualitative examples that illustrate the quantitative results presented in the last section. The examples of text-image relations are selected from *Zootopia* in the corpus, as presented in Figure 4.3.

A *concurrence* relation refers to the case where the text (or part of the text) equally corresponds to the salient element depicted in the image. For instance, the subtitle in Example 4.3-1 contains the *participant* "a fox", and the fox is simultaneously presented as the visual *participant* in the image. In Example 4.3-2, the subtitle contains the verbal *process* "return" and the bunny in the image stretches out her hand to indicate a substantially equivalent *process*. In Example 4.3-3, the arctic shrew utters the locative *circumstance* "at my wedding"; meanwhile, her wedding dress shown in the image also suggests the same semantic information about the location.

A *complementarity* relation means the text further explicates or modifies what is shown in the image. In Example 4.3-4, the bunny explains that the bullets in the sheep's gun are blueberries from her family farm, which provides additional descriptive information for the *participant* (i.e., the bullets in the gun) in the image. In Example 4.3-5, the tiger cub expresses the verbal *process* "can hunt for" to provide additional details about his action of standing on a stage and acting as an actuary dressed in a suit. Here, the verbal *process* is more illustrative than the visual *process*. In Example 4.3-6, the fox mentions a locative *circumstance* "in your dumb little stage play", which is not presented in the given scene, so the subtitle here complements the image in terms of *circumstance*.

A *condensation* relation, the opposite counterpart of *complementarity*, refers to the case where the text is semantically less specific than the image. In Example 4.3-7, the *participant* in the text is verbally signified by the exophoric reference "that", the meaning of which is revealed by the injury on the bunny's face. In other words, the verbal *participant* in this case is not as vivid as the visual one. In Example 4.3-8, the bunny notices her colleague, the cheetah, is putting his stuff into a box (the visual *process*). Then, in confusion, she tries to confirm what she sees by asking about what the cheetah is "doing" (the verbal *process*). It can be thus argued that the verbal *process* in this subtitle is less specific than the visual action in the image. In Example 4.3-9, the fox mentions the temporal *circumstance* "now", whereas the image reveals more precisely that this is a moment of bullying, as the fox pushes the bunny to the ground for humiliation.

A *contradiction* relation is established when there is a disparity between the meaning conveyed by the text and the image. Example 4.3-10 shows a case concerning the *participants*. In this example, the arctic shrew is humorously referred to as “Mr. Big” despite its visibly small size. This bizarre incongruity subverts the viewer’s expectation and generates a comedic effect. Thus far, no examples of *contradiction* relations have been found in relation to *processes* or *circumstances*. This aligns with the infrequent occurrence of this type of relation, as evidenced by the results presented in Table 4.2. Some conceptual examples, though, can be found in Figure 3.2 in Section 3.3.2.

These illustrative examples vividly show the intricate interplay between subtitles and images in films. To further explore the potential impact of interlingual subtitling on the interaction between text and image, it is essential to undertake a closer examination of translation shifts, which will be presented in the subsequent section.

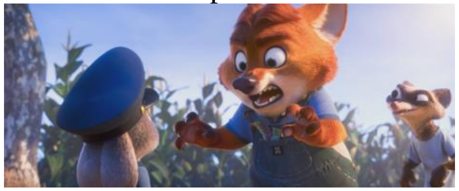
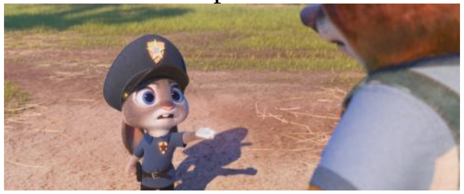
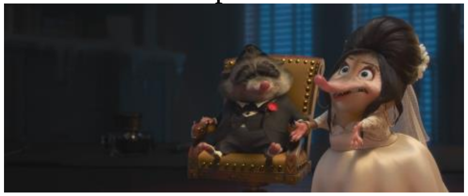
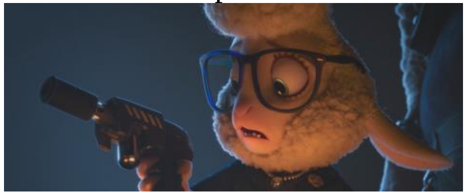
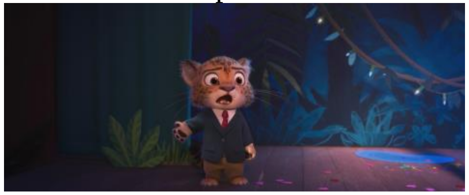
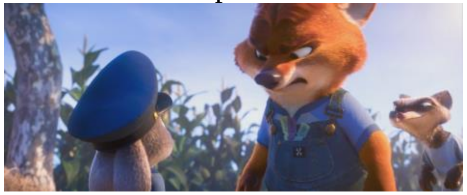
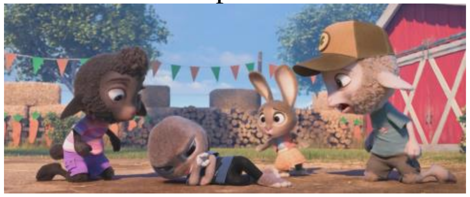
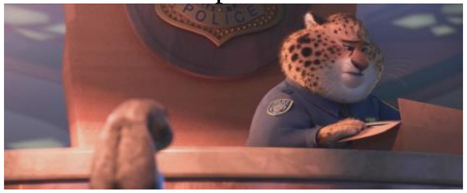
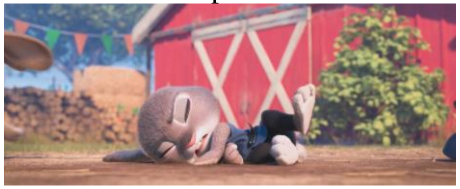
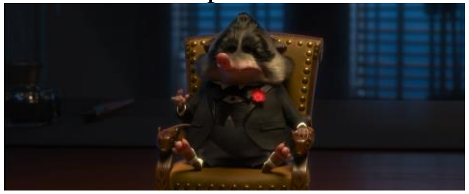
| **Text-image relations (4 Cs)** | **Verbal and visual Transitivity components** | | |
|---|---|---|---|
| | Participant | Process | Circumstance |
| Concurrence<br>(text = image) | Example 4.3-1<br><br>…because I'm **a fox**, | Example 4.3-2<br><br>Kindly **return** my friend's tickets. | Example 4.3-3<br><br>No icing anyone **at my wedding**! |
| Complementarity<br>(text > image) | Example 4.3-4<br><br>Those are **blueberries, from my family's farm**. | Example 4.3-5<br><br>Today I **can hunt for** tax exemptions. | Example 4.3-6<br><br>…and like you said **in your dumb little stage play**, |
| Condensation<br>(text < image) | Example 4.3-7<br><br>**That** looks bad. | Example 4.3-8<br><br>What **are** you **doing**? | Example 4.3-9<br><br>Scared **now**? |
| Contradiction<br>(text ≠ image) | Example 4.3-10<br><br>**Mr. Big,** sir,… | (Not found in the corpus, but see Figure 3.2 for a conceptual example) | (Not found in the corpus, but see Figure 3.2 for a conceptual example) |

**Figure 4.3** Examples of text-image relations in *Zootopia*

## 4.3 Analysis of translation shifts in text-image relations through subtitling

### *4.3.1 Quantitative results of translation shifts in text-image relations*

Table 4.3 presents how the text-image relations were shifted through interlingual subtitling across different transitivity components. In general, the frequency of the *concurrence* relation slightly increased after subtitling (from 1,347 to 1,365), while the *complementarity* and *condensation* relations experienced a mild decrease (from 1,272 to 1,240; from 184 to 148). However, as mentioned in Section 3.4, this broad comparison does not fully capture the potential impact of non-verbal elements on the verbal presentation during subtitling. For example, while there was an overall increase in instances of *concurrence*, the number of this relation between *participants* actually decreased through subtitling (from 541 to 500). This is why the framework of translation shifts in text-image relations is proposed in Section 3.4, as it allows for a more precise examination of translation shifts from a multimodal perspective.

**Table 4.3** Comparison of text-image relations between source and target subtitles across different transitivity components

| **Text-image relations (4 Cs)** | **Participants** | | **Processes** | | **Circumstances** | | **Total** | |
|---|---|---|---|---|---|---|---|---|
| | **ST** | **TT** | **ST** | **TT** | **ST** | **TT** | **ST** | **TT** |
| Concurrence | 541 | 500 | 626 | 654 | 180 | 211 | 1347 | 1365 |
| Complementarity | 943 | 886 | 261 | 284 | 68 | 70 | 1272 | 1240 |
| Condensation | 90 | 64 | 59 | 49 | 35 | 35 | 184 | 148 |
| Contradiction | 3 | 3 | 0 | 0 | 0 | 0 | 3 | 3 |

Table 4.4 presents the frequency of translation shifts across different transitivity units observed in the corpus. As explained in Section 3.4, translation shifts in the proposed framework are identified as non-shifts, language-induced shifts, or image-induced shifts. Image-induced shifts include *strengthening shifts* and *weakening shifts*, which suggest the subtitler's conscious or unconscious consideration of the visual elements during translation. Language-induced shifts, on the other hand, encompass *obligatory shifts* and *preferential shifts*, as they are more driven by the linguistic disparity or different norms between the target and source language systems. Overall, *non-shifts* were the most frequently observed type of shift, with 2,497 cases in the corpus. This suggests that the target subtitles were largely semantically similar to the source subtitles. *Obligatory shifts* and *preferential shifts* were relatively infrequent in the corpus.

Regarding the image-induced shifts, *strengthening shifts* and *weakening shifts* occurred with similar frequencies (149 and 166, respectively). Thus, it would be premature to draw a definitive conclusion about whether subtitling strengthens or weakens the connection between text and image.

**Table 4.4** Frequency of translation shifts in text-image relations across different transitivity components

| **Translation shifts** | | **Transitivity components** | | | **Total** |
|---|---|---|---|---|---|
| | | **Participant** | **Process** | **Circumstance** | |
| Non-shifts | | 1364 | 875 | 258 | 2497 |
| Language-induced | Obligatory | 46 | 42 | 8 | 96 |
| | Preferential | 17 | 23 | 10 | 50 |
| Image-induced | Strengthening (4 Es) | 53 | 57 | 39 | 149 |
| | Weakening (4 Ds) | 140 | 14 | 12 | 166 |

**Table 4.5** Frequency of subcategories of image-induced translation shifts

| **Translation shifts** | **Transitivity components** | | | **Total** |
|---|---|---|---|---|
| | **Participant** | **Process** | **Circumstance** | |
| Strengthening shifts | 53 | 57 | 39 | 149 |
| Expansion of concurrence | 36 | 28 | 6 | 70 |
| Expansion of complementarity | 3 | 11 | 30 | 44 |
| Expansion of condensation | 3 | 3 | 0 | 6 |
| Explicitation | 6 | 10 | 2 | 18 |
| Enhancement | 4 | 3 | 0 | 7 |
| Elaboration | 1 | 2 | 1 | 4 |
| Weakening shifts | 140 | 14 | 12 | 166 |
| Detachment of concurrence | 100 | 4 | 5 | 109 |
| Detachment of complementarity | 22 | 8 | 5 | 35 |
| Detachment of condensation | 14 | 0 | 1 | 15 |
| Diminution | 1 | 1 | 1 | 3 |
| Dilution | 3 | 1 | 0 | 4 |
| Decrement | 0 | 0 | 0 | 0 |

To further scrutinize the pattern of image-induced shifts in text-image relations, the frequency of the subcategories of *strengthening* and *weakening shifts* are compared, as presented in Table 4.5.

As shown in the first column of the transitivity components, among the image-induced shifts identified between participants, *detachment of concurrence* had the highest frequency (n = 100). Moving on to the second column, between processes, *expansion of concurrence* was the most common shift (n = 28). As for the cases between circumstances, *expansion of complementarity* had the highest frequency (n = 30). It is also worth noting that no instances of *decrement* were observed in the corpus, indicating its infrequent use in subtitling. Other rarely observed shifts include *expansion of condensation* (n = 6), *enhancement* (n = 7), *elaboration* (n = 4), *diminution* (n = 3), and *dilution* (n = 4).

To find out whether there was a statistically significant difference in the frequency of each subtype of shift with respect to each transitivity component, the chi-square test was employed. It should be noted that the shift of *decrement* was not included in the test due to all of its values being zero, which would further violate the assumption of the test that expected value in each cell should be greater than 5.

Table 4.6 presents the inferential statistical results of the chi-square test conducted on the 11 types of image-induced shifts. Among the 33 cells analyzed, 19 cells (57.58%) had an expected frequency of less than 5, violating the assumption of the chi-square test. Consequently, the alternative Fisher's exact test was utilized. The results of the test showed a highly significant association between the type of translation shifts and the type of transitivity components ($p$ < .001). However, the overall significant result did not specify which cells contributed the most to this effect. To determine the significance of each cell, standardized residuals were inspected. A cell value is considered significant if its standardized residual falls outside the range of ±1.96 (Field et al, 2012, p. 826). For example, the standardized residual of the *expansion of complementarity* shift between *participants* is -4.614, indicating that the observed frequency of this shift is significantly less than its expected frequency. Since the objective of this study is to find out shifts that frequently take place, it is reasonable to focus only on the cells with a positive standard residual rather than the negative ones.

**Table 4.6** Chi-square results of 11 types of image-induced translation shifts

| Translation shifts | Statistical items | Participant | Process | Circumstance | Total |
|---|---|---|---|---|---|
| Expansion of concurrence | Count | 36 | 28 | 6 | 70 |
| | Expected values | 42.889 | 15.778 | 11.333 | |
| | Std residual | -1.052 | **3.077** | -1.584 | |
| Expansion of complementarity | Count | 3 | 11 | 30 | 44 |
| | Expected values | 26.959 | 9.917 | 7.124 | |
| | Std residual | -4.614 | 0.344 | **8.571** | |
| Expansion of condensation | Count | 3 | 3 | 0 | 6 |
| | Expected values | 3.676 | 1.352 | 0.971 | |
| | Std residual | -0.353 | 1.417 | -0.986 | |
| Explicitation | Count | 6 | 10 | 2 | 18 |
| | Expected values | 11.029 | 4.057 | 2.914 | |
| | Std residual | -1.514 | **2.950** | -0.536 | |
| Enhancement | Count | 4 | 3 | 0 | 7 |
| | Expected values | 4.289 | 1.578 | 1.133 | |
| | Std residual | -0.139 | 1.132 | -1.065 | |
| Elaboration | Count | 1 | 2 | 1 | 4 |
| | Expected values | 2.451 | 0.902 | 0.648 | |
| | Std residual | -0.927 | 1.157 | 0.438 | |
| Detachment of concurrence | Count | 100 | 4 | 5 | 109 |
| | Expected values | 66.784 | 24.568 | 17.648 | |
| | Std residual | **4.065** | -4.150 | -3.011 | |
| Detachment of complementarity | Count | 22 | 8 | 5 | 35 |
| | Expected values | 21.444 | 7.889 | 5.667 | |
| | Std residual | 0.120 | 0.040 | -0.280 | |
| Detachment of condensation | Count | 14 | 0 | 1 | 15 |
| | Expected values | 9.190 | 3.381 | 2.429 | |
| | Std residual | 1.586 | -1.839 | -0.917 | |
| Diminution | Count | 1 | 1 | 1 | 3 |
| | Expected values | 1.838 | 0.676 | 0.486 | |
| | Std residual | -0.618 | 0.394 | 0.738 | |
| Dilution | Count | 3 | 1 | 0 | 4 |
| | Expected values | 2.451 | 0.902 | 0.648 | |
| | Std residual | 0.351 | 0.104 | -0.805 | |

Note: The bold number indicates a standardized residual that is positive and exceeds the threshold significance level of 1.96.

As Table 4.6 shows, the most significantly frequent shift that occurred between participants was *detachment of concurrence*, with a standardized residual of 4.065. This suggests that the target subtitles tended to avoid mentioning the relevant visual objects and leave the signified participant implicit in the target text. In terms of shifts occurring between processes, *expansion of concurrence* was predominantly employed, with a standardized residual of 3.077. This indicates a tendency in the target subtitles to provide additional specification for the co-occurring visual actions. When it comes to cases involving shifts between circumstances, *expansion of complementarity* was most frequently observed. This suggests that the target subtitles tended to include extra linguistic modifications to describe the visual processes.

### *4.3.2 Qualitative analysis of translation shifts in text-image relations*

In addition to the quantitative analysis, this section includes examples annotated in the corpus to facilitate the reader's understanding of the various translation shifts. The examples will first cover *non-shifts*, *obligatory shifts*, and *preferential shifts*. Subsequently, examples will be provided for four subtypes of image-induced shifts that were frequently observed in the corpus, indicated by a significantly high standardized residual shown in Table 4.6. These four shifts with a significantly higher frequency than expected encompass the following: *expansion of concurrence* between verbal and visual *processes*, *expansion of complementarity* between *circumstances*, *explicitation* between *processes*, and *detachment of concurrence* between *participants*.

Examples of *non-shifts* are shown in Figure 4.4. A n*on-shift* refers to the case where the target subtitle maintains the original text-image relation by reproducing the same experiential meaning of the transitivity components conveyed in the source dialogues. The frequent use of *non-shifts* in the corpus suggests that the subtitlers are largely faithful in conveying the original text-image relations. In the examples, cases of *non-shifts* can be observed between different transitivity components. Firstly, between the *participants*, the source and target subtitles both convey the verbal participant ("your ticket") that forms a *concurrence* relation with ticket visually present in the image. Moreover, between the *processes*, both the source and target subtitles include a more specific *process* ("give") to modify the visual *process* in the image, leading to a text-image relation of *complementarity*. Furthermore, regarding the *circumstances,* both the source and target subtitles contains an adverb phrase as the textual *circumstance* ("right now" and "immediately"), which provides more illustrative information for the action taking place in the image and forms a *complementarity* relation between the verbal and visual circumstance. As the transitivity components and the meaning of the target subtitles are basically equivalent to those of the source dialogues, no particular changes in text-image relations are observed, and they are identified as *non-shifts*.

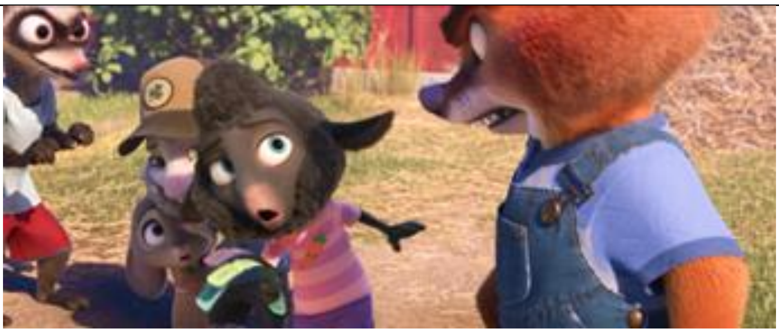
| **Background information** | In a scene from *Zootopia*, during a county fair event, Gideon, the fox, and his ferret friend are bullying two sheep and a bunny child. He tries to scare them and take their fair tickets. |
|---|---|
| **Visual frames** |  |
| **Source subtitles** | Gideon: ***Give me your tickets right now***,<br>or I'm gonna kick your meek little sheep butt. |
| **Target subtitles** | 吉丁: ***快把你的票给我***<br>[***Immediately give your ticket to me***,]<br>否则我把你的羊屁股踢烂！<br>[or I'll kick off your sheep butt!] |

**Figure 4.4** Example of *non-shifts* between the verbal and visual *participants, processes and circumstances*

An example of *obligatory shifts* is shown in Figure 4.5. An *obligatory shift* is the unavoidable alteration of the original text-image relation during subtitling due to inherent structural disparities between two languages. In this example, the first line of the source subtitles consists of one participant ("Mr. Frodo") and one process ("haven't had any sleep"), and they both correspond to the relevant visual elements on the image, where Frodo doesn't fall asleep to protect the ring throughout the scene. However, in the target subtitle, in addition to the similar *participant* and *process*, a new circumstance ("all this time") is added to represent the present perfect tense in the original English. Since there is no such verb tense in Chinese, the target subtitle has to employ an adverbial phrase to represent the same meaning. In this way, the text-image relation is changed, but it is primarily due to the linguistic differences between the two languages rather than the impact of the co-occurring visual elements. Therefore, it is identified as a case of *obligatory shift*.

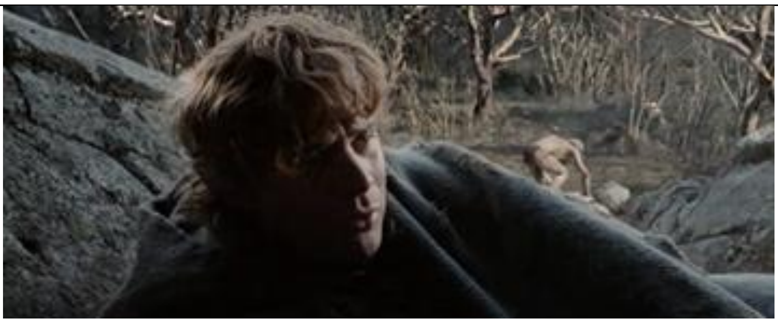
| **Background information** | In a scene from *The Lord of the Rings: The Return of the King*, Frodo, the protagonist, and Sam, his friend, are taking a rest in a cave. Sam wakes up and is worried whether it is too late to continue their adventurous journey. |
|---|---|
| **Visual frames** |  |
| **Source subtitles** | Sam: Haven't you had any sleep, Mr. Frodo?<br>I've gone and had too much. |
| **Target subtitles** | 山姆: 佛罗多 你***一直***都没睡吗<br>[Frodo, you didn't sleep ***all this time***?]<br>我睡得太久了<br>[I slept too long.] |

**Figure 4.5** Example of *obligatory shifts* between the verbal and visual *participants*

An example of *preferential shifts* is presented in Figure 4.6. A p*referential shift* occurs when the target subtitle adheres to the norms of the target language and thus alters the original text-image relation, making the target texts more acceptable to the target readers. In this example, Thanos' original line, "I'm a survivor", is rendered in the target subtitle as "我只是侥幸活下来了而已" ("I just luckily survived"). This translation transforms the original *participant* ("a survivor") into a *process* ("survived"), which causes an omission of the original text-image relation between the *participants* and leads to an addition of a new relation between the *processes*. Such shifts in text-image relations, however, are arguably not the result of the subtitler's multimodal awareness of the visual elements during translation, but rather the impact of conventional usage of the Chinese language. While it is grammatically correct to translate it word-for-word as "我是一个幸存者" ("I'm a survivor"), it sounds unnatural in Chinese and the sentence structure appears Europeanized. In this sense, the two shifts in this example are considered as language-induced shifts, and more specifically, *preferential shifts*.

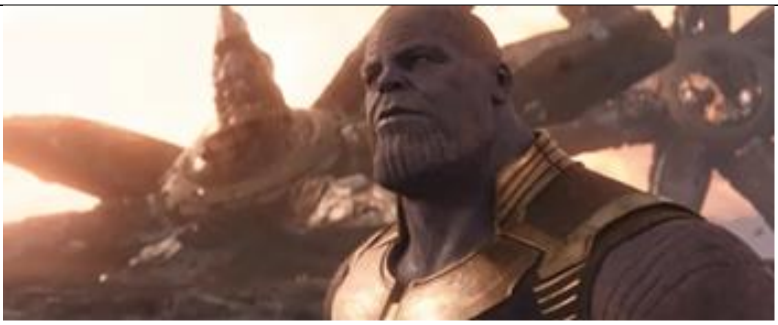
| | |
|---|---|
| **Background information** | In a scene from *Avengers: Infinity*, Dr. Strange and Thanos have a tense conversation before they have a fight. Thanos explains why and how he committed a genocide on his home planet, to which Dr. Strange responds in sarcasm and disgust. |
| **Visual frames** |  |
| **Source subtitles** | Dr. Strange: Congratulations, you're a prophet.<br>Thanos: I'm ***a survivor***. |
| **Target subtitles** | 奇异博士: 恭喜了 真是伟大的预言家<br>[Congratulations, really great prophet.]<br>灭霸: 我只是侥幸***活下来了***而已<br>[I just luckily ***survived***.] |

**Figure 4.6** Example of *preferential shifts* between the verbal and visual *participants* and *processes*

*Expansion* shifts refer to the addition of a new text-image relation in the target subtitle. The added text-image relation can be *condensation* (text < image), *concurrence* (text = image), or *complementarity* (text > image). The corpus comprises 28 cases of *expansion* of *concurrence* between the verbal and visual *processes* in subtitling. In the example presented in Figure 4.7, the target subtitles verbalize the action shown in the image. In the source dialogue, Lila simply says "No" as a response that her view is blocked by her father's hand. However, the target subtitle is rendered as "挡住啦" (blocked), which conveys more explicitly the co-occurring visual action in the image and establishes a *concurrence* relation regarding the *processes* conveyed in the text and the image.

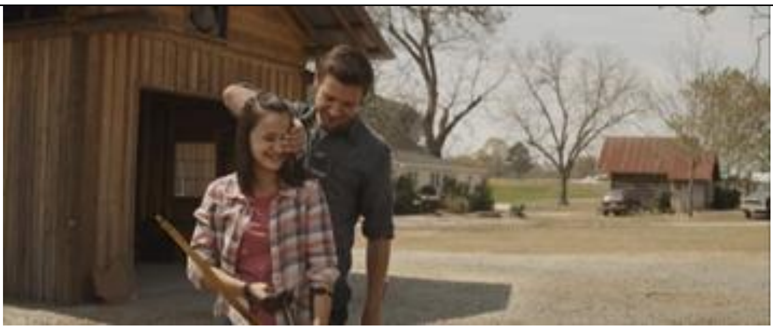
| | |
|---|---|
| **Background information** | In the beginning scene from *Avengers: Endgame,* Clint (also known as Hawkeye) is teaching archery to his daughter, Lila. At this particular frame, Clint is teasing Lila by blocking her view while she is trying to shoot an arrow. |
| **Visual frames** |  |
| **Source subtitles** | Clint: Can you see?<br>Lila: Yeah.<br>Clint: Are you sure?<br>Lila: Mm-hmm.<br>Clint: How about now? Can you see now?<br>Lila: ***No***. |
| **Target subtitles** | 鹰眼: 看得见吗 [Can see?]<br>莱拉: 可以 [Yes.]<br>鹰眼: 确定吗 [Sure?]<br>莱拉: (不译) [(No translation)]<br>鹰眼: 现在呢 还看得见吗 [How about now? Still can see?]<br>莱拉: ***挡住*** 啦 [***Blocked***.] |

**Figure 4.7** Example of *expansion of concurrence* between verbal and visual *processes*

Another common shift observed in the corpus is *expansion of complementarity*, between verbal and visual *circumstances*, with a frequency of 30 cases. An example of this type of shift is shown in Figure 4.8. In this example, the target subtitle includes the additional modifier “suddenly” to provide further clarification on the character’s hesitation, as expressed in the verbal process “don’t want to go”. This addition serves to elucidate the unexpected circumstance of the visual situation, establishing a new *complementarity* connection between the verbal and visual *circumstances*.

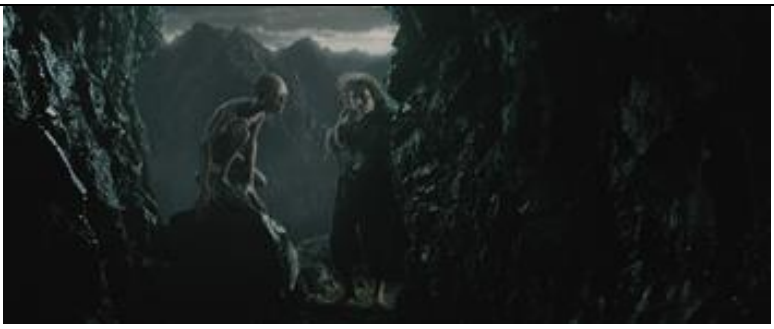
| | |
|---|---|
| **Background information** | In a scene from *The Lord of the Rings: The Return of the King*, Frodo is deceived by Gollum into entering a treacherous tunnel. The tunnel, ominously dark, instills a sense of unease and apprehension within Frodo. |
| **Visual frames** |  |
| **Source subtitles** | Gollum: Master must go inside the tunnel.<br>Frodo: Now that I’m here, I don’t think I want to. |
| **Target subtitles** | 咕噜: 主人得进到隧道里去<br>[Master must go inside the tunnel.]<br>佛罗多: 走到这儿 我***突然***不想往里走了<br>[Getting here, I ***suddenly*** don’t want to go inside.] |

**Figure 4.8** Example of *expansion* of *complementarity* between verbal and visual *circumstances*

*Explicitation* refers to the change in which the target subtitle alters the original *condensation* relation (text < image) into a relation of *concurrence* (text = image). In the corpus, 10 cases of *explicitation* were observed between the verbal and visual *processes*. As shown in the example in Figure 4.9, the target subtitle provides more explicit details in conjunction with the visual action. In the source subtitle, Joker only mentions a general action “start with”. The viewer has to infer from the visual information that Batman hits Joker’s head. Thus, the source subtitle is not as specific as what is shown in the image and they form a *condensation* relation. On the other hand, the target subtitle explicitly indicates Batman’s action of hitting, building a *concurrence* relation with the visual information. Compared with the source subtitle, the target subtitle in this case conveys more explicit information that aligns with the image, resulting in an *explicitation* shift.

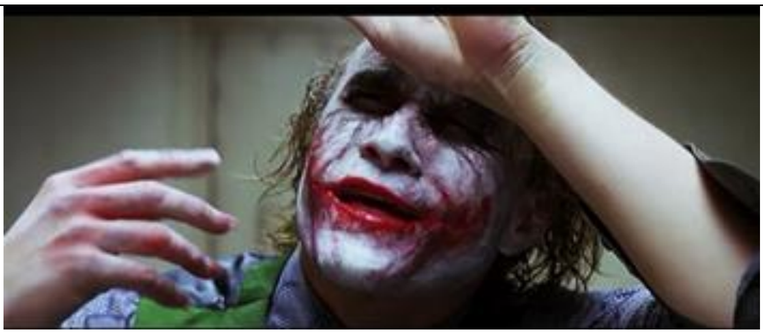
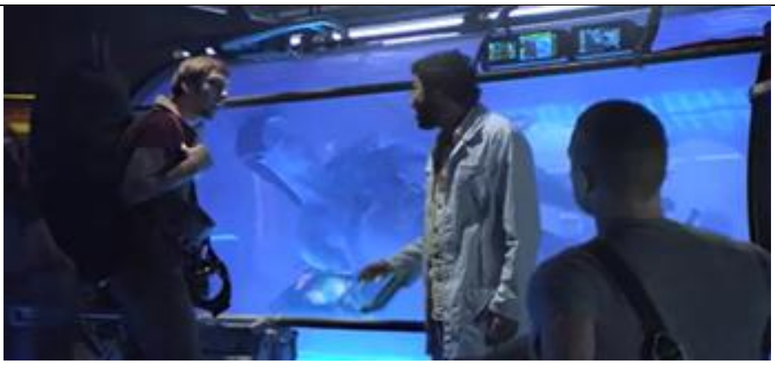
| **Background information** | In a scene from *The Dark Knight*, Batman suddenly shows up in the interrogation room, standing right behind Joker. Without Joker being aware, Batman swiftly strikes Joker's head against the table. |
|---|---|
| **Visual frames** |  |
| **Source subtitles** | Joker: Never ***start with*** the head.<br>The victim gets all fuzzy. |
| **Target subtitles** | 小丑: 别***打***头啊<br>[Don't ***hit*** the head.]<br>受害者会晕的<br>[The victim gets fuzzy.] |

**Figure 4.9** Example of *explicitation* between verbal and visual *processes*

*Detachment* shifts, frequently occurring in the corpus, refer to the omission of an existing text-image relation from the source subtitle. The omitted text-image relation can be *condensation* (text < image), *concurrence* (text = image), or *complementarity* (text > image). Noticeably, there were 100 cases of *detachment of complementarity* between the verbal and visual *participants* annotated in the corpus. Figure 4.10 presents a few examples of them. In these examples, the target subtitles omit certain *participants* present in the source subtitles, including "they", "us", "them", "you guys", and "them". All of these participants have their corresponding subjects in the image, forming the original relations of *concurrence*. However, these *participants* are absent and only implied in the target subtitles, leading to the translation shifts of *detachment of concurrence* between the *participants*.

| | |
|---|---|
| **Background information** | In a scene from *Avatar*, Norm, an avatar driver, introduces Jake, a newcomer, to the avatar lab. They are observing an avatar inside a container, accompanied by a doctor at the lab. |
| **Visual frames** |  |
| **Source subtitles** | Jake: Damn! They got big.<br>Norm: Yeah, they fully mature on the flight out.<br>So the proprioceptive sims seem to work really well.<br>Doctor: Yeah, ***they***’ve got great muscle tone.<br>It’ll take ***us*** a few hours to get ***them*** decanted,<br>but ***you guys*** can take ***them*** out tomorrow. |
| **Target subtitles** | 杰克: 见鬼，他们长大了<br>[Damn, they grew big.]<br>诺曼: 是的，他们已经完全成熟<br>[Yeah, they already fully mature.]<br>本体感知模拟器发育良好<br>[The proprioceptive sims develop well.]<br>博士: 是的，而且 Ø 肌肉强健有力<br>[Yeah, and Ø muscle strong and powerful.]<br>再有几个小时 Ø Ø 就可以出箱<br>[In a few more hours, Ø Ø can be decanted.]<br>不过明天 Ø Ø 便可出去<br>[But tomorrow Ø Ø can be out] |

*Note: the null sign (Ø) indicates the missing participant in the target subtitle.*

**Figure 4.10** Example of *detachment of concurrence* between verbal and visual *participants*

## 4.4 Discussion

### *4.4.1 Summary of corpus findings*

As shown in Table 4.4, Table 4.5 and Table 4.6, in interlingual subtitling, text-image relations (particularly *concurrence*, *complementarity*, and *condensation*) could undergo some degree of alteration across different transitivity components, either preserved (*non-shifts*), strengthened (*strengthening shifts*), or weakened (*weakening shifts*). *Non-shifts* were found to be the most frequently observed type of shift in the corpus. This finding tallies with pervious observations that a translator, often by default, translates the source text literally, without employing any translation shifts unless the literal approach becomes unavailable or problematic (Ivir, 1981; Tirkkonen-Condit, 2005).

A translation shift, in terms of the original meaning or the relationship between the original text and images, reflects the conscious or unconscious decision-making of the translator/subtitler in considering the concurrent visual content within the audiovisual product. Through an analysis of standardized residuals, three significantly frequent translation shifts in text-image relations between each transitivity component (*process*, *participant*, and *circumstance*) can be identified as follows:

1. The shift of *expansion of concurrence* between the verbal and visual *processes* was frequently observed in the target films, suggesting that target subtitles tend to explicate the co-occurring visual actions.
2. The shift of *detachment of concurrence* between the verbal and visual *participants* was frequently found in the target films, indicating the tendency to avoid mentioning the corresponding visual objects and leave the signified *participant* implicit in the target subtitles.
3. The shift of *expansion of complementarity* between the verbal and visual *circumstances* was commonly observed in the target films, suggesting that target subtitles are apt to modify or provide more descriptive information for the visual kinesics.

The *expansion of concurrence* and *complementarity* relations in the target subtitled content is in line with Baumgarten's (2008, p. 20) observation on AVT that "translations display a greater redundancy between the verbally and the visually given meanings" by creating a stronger and more explicit link between the image and the verbal text. Nevertheless, subtitling can also make the original text-image links more implicit. As indicated by the frequent cases of *detachment of concurrence* between *participants*, the target subtitles sometimes omit the original verbal *participant* in the source dialogues and leave the intended meaning implicit and revealed by the image. Such a subtitling method echoes what Taylor (2012, p. 27) calls "the minimalist approach", by which the subtitler intends to shorten sentences and let the viewers derive the meaning from other semiotic resources at their disposal. In general, the translation shifts observed in the corpus suggest closer and more intensive text-image interactions in the translated audiovisual content.

### *4.4.2 Subtitling as multimodal representation*

The major findings from the multimodal corpus analysis have unraveled an important role of subtitling as multimodal representation. It is evident that interlingual subtitling involves divergence not only from the original text but from the original text-image interactions.

In the corpus, the observed patterns of translation shifts were more or less inconsistent. While *weakening shifts* were frequently identified (e.g., *detachment of concurrence* between *participants*), *strengthening shifts* in the opposite direction were also commonly found (e.g., *expansion of complementarity* between *circumstances*). This seemingly paradoxical pattern aligns with Tortoriello's (2011) observation that the subtitling method for the verbal and non-verbal interaction is often not identical or consistent. Furthermore, this pattern complicates the previous observation that subtitling is a linguistically reductive process (e.g., Díaz Cintas & Remael, 2021; Han & Wang, 2014), as some new linguistic elements can be added in the target subtitles to interact with the visual information. Such inconsistent patterns, indicative of the subtitler's complex decision-making process when transferring non-verbal information into the target subtitles, can be explained by the possibility that different verbal or visual elements may require different approaches, as they have varying roles in representing experiential meanings (Chen, 2019). These experiential meanings, from the standpoint of SFG (Halliday & Matthiessen, 2014) and VG (Kress & Van Leeuwen, 2021), are represented through different transitivity components, namely *processes*, *participants*, and *circumstances*.

In terms of *processes*, the first major finding from the corpus is the tendency of target subtitles to explicate the corresponding visual kinesics depicted in the image, i.e., *expansion of concurrence* between verbal and visual *processes*. This tendency can be explained by the subtitler's intension to enhance the multimodal cohesion in the audiovisual text, one of the key aspects of subtitling emphasized in previous AVT research (e.g., Lautenbacher, 2015; Remael & Reviers, 2019; Taylor, 2016), which has also been termed as semiotic cohesion (e.g., Tortoriello, 2011; Vitucci, 2017) with similar meaning. On the one hand, Chaume (2004) remarks that subtitlers should consider the interwoven information conveyed in the interaction between different meaning-making modes (e.g., verbal, visual) rather than transmit the meaning within each mode separately. On the other hand, Tortoriello (2011) reminds the subtitler that a subtitle which explicates the message originally conveyed in the non-nonverbal channel may also have a negative effect of "semiotic tautology" (p. 74). In light of these divergent views, the observed tendency in this corpus (represented in Table 4.6) suggests that, in practice, or at least in the context of English-Chinese subtitling, the subtitlers prioritize semiotic cohesion over avoiding semiotic tautology. They reiterate the visual meaning in the verbal information and thereby creating redundancy between the text and image in the audiovisual text for the target viewers. Unlike the predominantly qualitative analyses of multimodal cohesion in previous studies, the tendency observed in this study provides quantitative evidence to support the endorsement of amplifying multimodal cohesion, at least for the case of verbal and visual *processes*.

The second major finding of the corpus analysis relates to the transitivity component of *participants*. Through frequent *detachment of concurrence* between verbal and visual *participants*, the target subtitles tended to omit the deictic information in the original dialogues and thus leave the signifying entities, such as pronouns, character names, and non-human objects, implicit in the target text. These losses in textual and visual concurrence regarding the *participants* contrast with the patterns for *processes* mentioned earlier. However, it is consistent with Wang et al.'s (2017) observation that nearly one thirds of the English pronouns are missing in the Chinese target subtitles for films and TV programs. One possible explanation is that Chinese, unlike English, is a pro-drop language (Huang, 1984; Wang et al., 2017), where certain pronouns can be omitted while still remaining grammatically or pragmatically comprehensible. Similar tendencies, with a multimodal consideration of the non-verbal elements, are identified by Bogucki (2020), who argues that characters' names are more frequently to be substituted by pronouns or omitted when the referred persons are visible in the scene. This "paring down of the verbal component", as asserted by Taylor (2004, p. 161), is justifiable as long as the meaning of the film text is still encoded by other non-verbal modes (e.g., visual clues, facial expressions, etc.). The absence of the *participant* in the target subtitles seems to be a natural resort to the spatial and temporal constraints in subtitling. A multimodal approach to subtitling has encouraged this method of omission, given "the availability of a number of semiotic modes which are capable of compensating for missing textual messages" (Bączkowska, 2011, p. 61).

As for the transitivity component of *circumstances*, another major finding derived from the corpus is the tendency of target subtitles to additionally modify or provide more descriptive information for the visual kinesics, i.e., *expansion of complementarity* between the verbal and visual *circumstances*. This subtitling method echoes Barthes's (1977, cited in Martinec & Salway, 2005) remark that dialogue in films "functions not simply as elucidation but really does advance the action by setting out (…) meanings that are not found in the image itself" (p. 341). The expanded complementary elements observed in the corpus typically took the form of adverbial words or phrases related to the emotional state of the character as visually depicted in the scene, for instance, highlighting the character's hesitation (see Figure 4.8). It can be inferred that the subtitler, based on the visual context, attempts to "reinforce the characterization" of the portrayed character (Messerli, 2019, p. 538). By providing more linguistic details that correspond to the visual entities, the visual and verbal elements of the film narrative "do not simply co-exist" but "are internally related to each other" and fused together as one (Baumgarten, 2008, p. 11).

The aforementioned corpus findings have underlined the role of subtitling as multimodal representation. However, these findings also complicate this multifaceted role by demonstrating that subtitling can **reiterate** (*expansion of concurrence* between *processes*), **remove** (*detachment of concurrence* between *participants*), or **reinforce** (*expansion of complementarity* between *circumstances*) the interactions between text and image in subtitled films. From this multimodal standpoint, it is difficult to determine whether subtitling is a reductive or expansive procedure. It appears that subtitling methods do not adhere to a strict dichotomy but rather coexist dynamically. Different transitivity components demonstrate varying tendencies of shifts through subtitling. While the corpus-based study sheds important light on the patterned shifts of subtitle-image relations, a question that still remains is whether such shifts will affect target viewers' reception of the translated content. To delve into this aspect, a reception experiment will be conducted, the details of which will be outlined in the forthcoming chapter.

# CHAPTER 5 THE EYE-TRACKING EXPERIMENT

As observed in the last chapter of the corpus study, in interlingual subtitling, text-image relations (particularly *concurrence*, *complementarity*, and *condensation*) could undergo alteration across different transitivity components. Based on these corpus findings, Chapter 5 investigates the potential impact of certain types of translation shifts in text-image relations on viewer reception (RQ3). The chapter commences with a brief overview of the design and insights obtained from the pilot experiment. Then, the design of the main experiment is outlined, covering the hypotheses, stimulus materials, participants, treatment conditions, variables, instruments, data processing, and procedure. The chapter concludes with a comprehensive analysis, both quantitative and qualitative, followed by an overall discussion.

## 5.1 Pilot experiment

The pilot experiment was designed to examine the feasibility of the eye-tracking approach to be used in the main experiment. It did not serve to provide a meaningful effect size estimate for planning the main experiment due to small samples. Nevertheless, it contributed to identifying and avoiding potential problems for the main experiment. The pilot experiment, as part of the PhD project, was approved by the Ethics Committee of The Hong Kong Polytechnic University (ref: HSEARS20220228002).

### *5.1.1 Participants, materials, and procedures*

Before the pilot experiment, a few printed posters for participant recruitment were put up on the university campus. Twenty-one participants who were native Chinese speakers and habitual viewers of simplified Chinese subtitles voluntarily signed up for the pilot experiment. They were randomly divided into a Control Group (n=11) and an Experimental Group (n=10). Both groups watched the same video clips but with different subtitles (i.e., treatment conditions), as shown in Table 5.1. The Control Group (CG) watched video clips with the subtitles adapted from official translation, while the Experimental Group (EXG) watched video clips with the subtitles containing more translation shifts in text-image relations based on the pilot results of the corpus-based analysis.

**Table 5.1** Material design and treatment conditions

| Clip | Source film | Duration | Control Group | Experimental Group |
|---|---|---|---|---|
| 0 | *Gifted* (2017) | 1m 3s | Clip P0 | Clip P0 |
| 1 | *Avengers: Endgame* (2019) | 1m 47s | Clip P1c<br>(0 translation shifts in *concurrence* relations between verbal and visual *processes*) | Clip P1e<br>(10 translation shifts in *concurrence* relations between verbal and visual *processes*) |
| 2 | *Avengers: Infinity War* (2018) | 2m 16s | Clip P2c<br>(0 translation shifts in *condensation* relations between verbal and visual *participants*) | Clip P2e<br>(10 translation shifts in *condensation* relations between verbal and visual *participants*) |
| 3 | *Zootopia* (2016) | 1m 34s | Clip P3c<br>(0 translation shifts in *complementarity* relations between verbal and visual *circumstances*) | Clip P3e<br>(10 translation shifts in *complementarity* relations between verbal and visual *circumstances*) |

Four film excerpts were selected as the audiovisual stimuli for the pilot experiment. The original language of the films was all English, but the dubbed versions in Spanish was used instead so as to control the confounding factor of the participants' varied English proficiency level. The excerpts were labeled as Clip P0 (with P referring to pilot), Clip P1, Clip P2, and Clip P3. Clip P0 was exactly the same for the CG and EXG participants, serving as a benchmark material to ensure that both groups had comparable levels of (a) comprehension performance, (b) cognitive abilities, and (c) perception of subtitle quality. The other three clips were presented differently to the two groups in an attempt to provide tentative information for the proposed hypotheses, as Table 5.1 shows.

The independent variable in the experiment was the version of the subtitles (no translation shifts in text-image relations vs. translation shifts in text-image relations). The dependent variables in the experiment were threefold, i.e., visual attention, comprehension performance, and perception of subtitle quality. These variables were tested by eye-tracking data, questionnaires, and interviews. The participant's visual attention during watching the videos was gauged with eye-tracking data. Open-ended questions were used to test the participants' comprehension performance. Closed-ended questions, in the form of six-point Likert scale guided by Künzli's (2021) CIA model of subtitle quality, were applied to examining the participant's perception of subtitle quality. Besides these quantitative methods, an interview was conducted with each participant to attain comments about their viewing experience, expectation of subtitles, and their feelings during the pilot experiment. Their thoughts about the experiment could provide insights for optimizing the design of the main experiment. To control confounding variables and ensure comparability, a demographic questionnaire was used to collect the participant's information concerning their age, gender, language proficiency in Spanish (the dubbing language in the videos) and viewing habits.

The participants were tested individually in an eye-tracking lab. For the CG participants, they were first briefed on the general goal of the experiment before filling out the demographic questionnaire regarding their general personal information. Afterward, they were guided to sit in front of a screen equipped with an eye-tracking device and they were presented the four video clips (i.e., Clip P0, Clip P1c, Clip P2c, Clip P3c) one after another. At the end of each clip, they answered two sets of questionnaires concerning their comprehension of the audiovisual content (e.g., plots, characters) and their perception of the subtitle quality. After they finished all the clips and questionnaires, the author had a short interview with them. The EXG participants went through a similar process, except that they watched different clips with manipulated subtitles (i.e., Clip 0, Clip P1e, Clip P2e, Clip P3e). At the end of the experiment, the individual participant was informed of the specific purpose of the experiment.

### *5.1.2 Insights gained from the pilot study*

The analysis of the data collected in the pilot study suggested that the proposed instruments were suitable for investigating the viewers' visual attention, comprehension, and perception of subtitle quality. The open-ended questions were feasible for evaluating viewers' comprehension of the audiovisual content. The six-point Likert scale used to assess viewers' perception was proved reliable by the reliability test of Cronbach's alpha ($\alpha = 0.79$ in Clip P0; $\alpha = 0.84$ in Clip P1; $\alpha = 0.89$ in Clip P2; $\alpha = 0.86$ in Clip P3). The one-hour duration of the experiment was acceptable to the participants and did not lead to any noticeable attention loss or cognitive fatigue during the eye-tracking session.

However, three major issues were observed during the pilot experiment. The first problem was the scattered narratives of the audiovisual stimuli throughout the experiment. Since the video clips were selected from different films, some participants commented that they felt like they were opening a blind box every time they were exposed to a new video. As a result, at the beginning of each viewing, they had to make extra cognitive efforts to get used to the new characters and make sense of the new storyline. To address this issue, excerpts from one single film were used as audiovisual stimuli in the main experiment.

The second issue was the participants' familiarity with the audiovisual content. The video clips selected for the experiment were excepts from popular English films among Chinese viewers, as indicated by the film list in the self-built corpus. Many participants mentioned that they had already watched the film before the experiment, which somewhat assisted their comprehension performance in the follow-up questionnaires. To manage this issue, an older and less popular film was selected for the main experiment.

The last major issue was the viewers' poor eye-tracking data quality when watching Clip P3. Out of the 21 participants, 10 viewers had a gaze sampling ratio below 70%, a high level of data loss compared with that for the viewing of other clips. This issue was later attributed to the relatively low brightness of the video, as a darker screen may cause a vaguer corneal reflection and wider pupil dilation in the viewer's eyes, making it difficult for the eye tracker to accurately track the point of gaze. To prevent similar technical issues, film scenes with insufficient lighting were not considered in the main experiment.

## 5.2 Main experiment

### *5.2.1 Hypotheses*

The main experiment was designed following a similar structure to that of the pilot experiment, but with some adjustments based on the insights gained from the pilot study. The hypotheses for the main experiment were formulated by taking into account the results of the corpus-based study, preliminary findings from the pilot experiment, and previous AVT reception studies (Chen, 2020; Lautenbacher, 2015):

H1: Viewers who watch films with more *expansion of concurrence* relations in the target subtitles regarding visual *processes* allocate shorter gaze time to the image areas, have higher comprehension scores, and perceive better quality of the subtitles.

H2: Viewers who watch films with more *detachment of concurrence* relations in the target subtitles regarding visual *participants* allocate longer visual attention to the visual elements but do not decrease in comprehension scores or perception of subtitle quality.

H3: Viewers who watch films with more *expansion of complementarity* relations in the target subtitles regarding visual *circumstances* allocate less visual attention to the image areas, have higher comprehension scores, and perceive better quality of the subtitles.

### *5.2.2 Materials*

The audiovisual stimuli selected for the main experiment consisted of four excerpts from the American comedy film *Meet the Parents*, released in 2000. The film revolves around a young male nurse Greg and his first visit to his girlfriend's house, where he meets his prospective parents-in-law. Despite Greg's good-heartedness, the visit turns into a nightmarish experience due to a series of unfortunate incidents. The film was chosen as an ideal source of audiovisual stimuli for this experiment for mainly two reasons: (a) it was released more than two decades ago and was not commonly known among the Chinese audience; (b) it contains short and self-contained scenes that can be selected and combined to form a comprehensive narrative, effectively addressing the issue of scattered storylines identified in the pilot experiment. While the original language of the films was English, the film excerpts presented to the participants were dubbed in Thai, a foreign language unknown to the participants in order to control the confounding factor of their varied English proficiency level. The dubbed excerpts in Thai were from the DVD officially released by Universal Studios & DreamWorks LLC in 2000. The scenes selected as the stimulus video clips are described in Table 5.2.

Two sets of subtitles were designed for the experiment based on the target subtitles in simplified Chinese originally provided on the DVD. These two sets were assigned to the Control Group (CG) and the Experimental Group (EXG) respectively. The main difference between the two sets of subtitles lay in their semantic interplay with the visual information in the videos (for further comparative description, see Section 5.2.4). While the textual contents were different between the subtitles, their level of readability was controlled. Additionally, the length of the subtitle lines, measured by the number of Chinese characters in each line of the subtitles, was kept consistent between the two sets. The representative features of the subtitles, such as font size, color, position, and presentation speed, were also kept as the same, as they are potential confounding variables in this study (Silva et al., 2022). The key linguistic and technical characters of the designed subtitles are presented in Table 5.3.

**Table 5.2** Design and descriptive information of the audiovisual stimuli

| **Clip** | **Duration** | **No. of original English words** | **Scene description** |
|---|---|---|---|
| 0 | 1m 8s | 153 | **Location**: airport<br>**Plot**: Greg arrives at the airport with a bag that is too big to carry on. He is informed by the airline staff that he needs to check the bag. Unfortunately, the airline loses his bag. |
| 1 | 2m 24s | 321 | **Location**: Pam’s house<br>**Plot**: Greg and his girlfriend Pam arrive at her house, where Greg is introduced to Pam’s parents and their pet cat for the first time. They engage in initial conversations and get to know each other. |
| 2 | 1m 47s | 201 | **Location**: Pam’s house<br>**Plot**: Greg and Pam’s family gather in the living room, where Pam’s mother opens the gift that Greg brings to the family and they engage in a conversation. |
| 3 | 1m 43s | 346 | **Location**: airport<br>**Plot**: Greg boards the airplane after his nightmarish visit to Pam’s house. He tries to fit his oversized bag into the overhead luggage rack and refuses the flight attendant’s request to check the bag, leading to a heated argument. Greg becomes upset and causes a scene and is carried out of the plane by the security guards. |

The subtitles were created and designed by the author with a free and open-source software for video and subtitle editing, titled 人人译世界 (1sj.tv). After the subtitles were finalized, they were hardcoded within the software into the selected film excerpts. The final output of the audiovisual stimuli for the experiment comprised two sets of videos. Both sets contained the same audiovisual content, which was dubbed in Thai. However, they differed in terms of the target subtitles in Chinese. The videos had a resolution of 1280 × 720 and were encoded at a frame rate of 30 fps.

**Table 5.3** Linguistic and technical characteristics of the designed subtitles

| Subtitle characteristics | Clip 0 | | Clip 1 | | Clip 2 | | Clip 3 | |
|---|---|---|---|---|---|---|---|---|
| | CG | EXG | CG | EXG | CG | EXG | CG | EXG |
| Tier one frequently used Chinese characters (%)[1] | 95.78 | | 96.88 | 96.62 | 98.03 | 98.03 | 98.99 | 98.99 |
| No. of Chinese characters | 162 | | 382 | | 254 | | 297 | |
| No. of lines | 17 | | 55 | | 30 | | 36 | |
| Avg. Chinese characters per line | 9.53 | | 6.95 | | 8.47 | | 8.25 | |
| Avg. subtitle speed (cps)[2] | 8.15 | | 8.50 | | 8.40 | | 7.70 | |

### *5.2.3 Participants*

Before the main experiment, some printed posters for participant recruitment were put up on the university campus, with the prior approval of the university's ethics committee (ref: HSEARS20220228002). To enroll in the experiment, interested individuals were required to complete an online questionnaire concerning their demographic information, including their name, age, foreign language proficiency, academic background, and preferred time for participation in the experiment. During the experiment, the participants were also asked to

[1] "Tier one frequently used Chinese characters" refers to the 3,500 most frequently used characters in the Chinese language (MOE, 2013). A higher proportion of tier-one characters within a Chinese text implies a higher level of readability. The figures in the Table indicate that the Chinese subtitles in the video clips were easily comprehensible, as all of them had a percentage exceeding 95%.

[2] The speed of a subtitle line in this study was determined by dividing the duration of the subtitle line by the total number of characters it contains. The average subtitle speed for a given clip was calculated by taking the mean value of the speeds of each subtitle line within the clip. This calculation method has been commonly employed in experimental AVT research (Fresno & Sepielak, 2022). The subtitle speed in the clips for this experiment was approximately 8 characters per second (cps). This speed was deemed suitable, as 95.7% (314 out of 328) of the responses from the 82 viewers in the main experiment indicated that they *somewhat agreed*, *agreed*, or *strongly agreed* to the appropriateness of subtitle speed in each of the four stimulus clips (i.e., assigning ratings of 4 or above on a 6-point scale). Hence, the viewers had sufficient time to comprehend the textual information presented. It should also be noted that when calculating the speed of Chinese subtitles, each Chinese character takes up two spaces, equivalent to two English letters since Chinese characters are known as full-width characters that usually occupy two spaces instead of one space. For example, if a subtitle line has 6 Chinese characters and appears for 4 seconds, the cps of this subtitle line is (6*2) / 4 = 3.

complete another demographic questionnaire about their gender, viewing habit and previous exposure to the audiovisual stimuli used in the experiment. The contained demographic information served to control the potential confounding variables (age, gender, source language proficiency, viewing experience) and ensure inter-group comparability.

More than 120 participants voluntarily signed up for the experiment. After preliminary screening, 88 participants were recruited, who were native Chinese speakers and habitual viewers of simplified Chinese subtitles. Prior to their individual invitation to the eye tracking lab, they were divided into a Control Group or an Experimental Group based on their age so as to ensure age control among them. Six participants were unable to complete the entire experiment as they did not pass the eye tracking calibration session. Consequently, the final number of participants for the experiment was 82. The demographic information of the two groups is shown in Table 5.4.

**Table 5.4** Demographic information of the participants in the main experiment

| | | **Control Group (CG)** | **Experimental Group (EXG)** |
|---|---|---|---|
| Number of Participants | | 40 | 42 |
| Age | Range | 18-30 | 18-30 |
| | Mean | 25.35 | 25.26 |
| | SD | 2.69 | 2.91 |
| Gender | Male | 13 | 13 |
| | Female | 27 | 29 |
| Academic Background | | Not related to linguistics or language studies | |
| Language Proficiency | | No or limited proficiency of Thai (the dubbed language) | |
| Viewing Habits | | Regular viewers of subtitled audiovisual content | |
| Previous Exposure | | Not familiar with the film excerpts | |

Age was controlled in the experiment. According to data from Maoyan Movie, China's largest online ticket retailer, 48% of the Chinese cinema goers in 2021 were at the age between from 20 to 29, with similar proportions marked in 2019 and 2020 (Maoyan, 2022). Participants in this experiment were thus selected from a similar age range of 18-30, with the main focus on the largest group of film consumers in China to ensure more representative results of the experiment. The average age of the CG participants was 25.35 (SD = 2.69), similar to 25.26 as among the EXG viewers (SD = 2.91).

Gender was also considered in the experiment. As reported in Burczynska (2017), gender difference was not observed in a subtitled film reception study. Nevertheless, in this study, male and female participants are balanced in both groups. There were 13 male and 27 female viewers in the CG, and similarly, there were 13 male and 28 female subjects in the EXG.

The viewers' language background was also controlled. Given the growing English proficiency among Chinese young people in recent years, the film excerpts presented to the participants were the dubbed versions in Thai. In this way, the viewers had the same level of proficiency in the source language of the films and they had to rely on the target subtitles to comprehend the verbal content and assist their understanding of the audiovisual content. It is also worth noting that none of the participants had a professional background in language or linguistics. Thus, it can be assumed that their reception of the subtitled content was natural and unaffected by theories from linguistics or translation studies.

The participant's viewing habit and previous exposure to the audiovisual stimuli may affect their ultimate reception and were thus controlled for the experiment. The majority of the viewers had never seen the audiovisual stimuli before the experiment. Only one viewer reported having watched the film before, who watched a few excerpts of the film on the Internet one year ago and had only fuzzy memories of the film, so the data retrieved from this participant was not excluded. Based on the results obtained from a six-point Likert scale question (1 = never, 6 = always) regarding the frequency of reading target subtitles when watching foreign films, both groups reported reading subtitles often or more than often. The results from the one-sample t-test showed that both groups had a mean score significantly over 4 (CG: mean = 4.60, $t$ = 3.509, $p$ = 0.001; EXG: mean = 4.73, $t$ = 4.787, $p$ < 0.001). This suggests that the viewers in the experiment were accustomed to reading Chinese subtitles for audiovisual content in a foreign language.

Given the aforementioned controlled factors, it can be inferred that any significantly different viewing patterns observed later between the groups can be attributed to the treatment conditions in the experiment, as described in the upcoming section.

### *5.2.4 Treatment conditions*

Both groups watched the same audiovisual stimuli but with different subtitles (treatment conditions), as shown in Table 5.5. The CG viewers watched video clips with the subtitles adapted from official DVD translation, while the EXG viewers watched clips with the subtitles containing more translation shifts in text-image relations based on the results of the corpus-based analysis (for the full design of the manipulated subtitle versions, see Appendix 2).

**Table 5.5** Treatment conditions

| Clips | Control Group | Experimental Group |
|---|---|---|
| 0 | Clip 0 | Clip 0 |
| 1 | Clip 1c<br>(0 cases of *expansion* of *concurrence* relations between verbal and visual *processes*) | Clip 1e<br>(10 cases of *expansion* of *concurrence* relations between verbal and visual *processes*) |
| 2 | Clip 2c<br>(0 cases of *detachment* of *concurrence* relations between verbal and visual *participants*) | Clip 2e<br>(10 cases of *detachment* of *concurrence* relations between verbal and visual *participants*) |
| 3 | Clip 3c<br>(0 cases of *expansion* of *complementarity* relations between verbal and visual *circumstances*) | Clip 3e<br>(10 cases of *expansion* of *complementarity* relations between verbal and visual *circumstances*) |

Clip 0 was presented to both the CG and EXG participants with the same target Chinese subtitles. This clip served as a baseline material to ensure that both groups had comparable levels of: (a) comprehension performance, (b) cognitive abilities, and (c) perception of subtitle quality. The other three clips, namely Clip 1, Clip 2, and Clip 3, were presented with different Chinese subtitles to the two groups. This was designed to test the proposed hypotheses of the experiment (for more details about the video clips, see section 5.2.1).

Clip 1 was designed to testify Hypothesis 1, concerning the impact of expanded *concurrence* relations between the verbal and visual *processes* on viewers' reception. In Clip 1c, the target subtitles were generally direct transfer of the original dialogues. In contrast, the subtitles in Clip 1e were more explicit about the co-occurring visual actions in the scene, incorporating ten additional cases of *concurrence* relations. An example of the different treatment is presented in Figure 5.1.

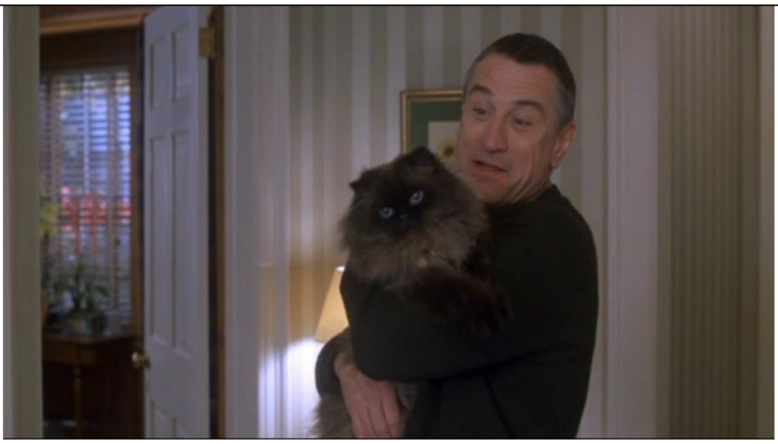
| | |
|---|---|
| **Background information** | Pam’s father, Jack, is introducing his beloved pet cat to Greg. |
| **Visual frames** |  |
| **Source subtitles** | Jack: Attaboy. That took me another week. |
| **CG subtitles** | 杰克: 我又花了一个星期教它这个<br>[I spent another week teaching it this.] |
| **EXG subtitles** | 杰克: 我又教了一星期 它就***会挥手***<br>[I taught for another week and it ***can wave***.] |

**Figure 5.1** Example of *expansion of concurrence* between *processes* in Clip 1

Clip 2 served to verify Hypothesis 2, regarding the potential impact of detached *concurrence* relations between the verbal and visual *participants*. The target subtitles in Clip 2c generally maintained the *participants* that were originally present in the source dialogues and semantically correspondent to the characters in the mise-en-scène. However, in Clip 2e, ten cases of the verbal participants in the source dialogues were omitted in the target subtitles, leaving the referential information implicit for the target viewers. Figure 5.2 shows an example of such a case.

| | |
|---|---|
| **Background information** | Greg is complaining to Pam that she shouldn’t have told her parents that Greg doesn’t like cats, given that her father is a cat lover. |
| **Visual frames** | |
| **Source subtitles** | Greg: You know, I wish you hadn’t told your parents I hate cats.<br>Pam: But you do hate cats.<br>Greg: You didn’t have to tell them right when we met.<br>Pam I know. I’m sorry. It just kinda slipped out. |
| **CG subtitles** | 阿基: 真希望你没说我讨厌猫<br>[Really wish you hadn’t said I hate cat.]<br>白梅: 你真的讨厌猫啊<br>[You do hate cats.]<br>阿基: 你不用一见面就告诉他们<br>[You didn’t have to tell them at the first meeting.]<br>白梅: 对不起 我只是顺口<br>[Sorry, I just slipped out.] |
| **EXG subtitles** | 阿基: 真希望你没说我讨厌猫<br>[Really wish you hadn’t said I hate cat.]<br>白梅: 你真的讨厌猫啊<br>[You do hate cats.]<br>阿基: Ø 也不用一见面就说出来的 Ø<br>[Ø Didn’t have to say Ø at the first meeting.]<br>白梅: 对不起 Ø 就只是顺口<br>[Sorry, Ø just slipped out.] |

*Note: the null sign (Ø) indicates the missing original participant in the target subtitle.*

**Figure 5.2** Example of *detachment of concurrence* between *participants* in Clip 2

Clip 3 was used to substantiate Hypothesis 3, which examined how the viewer’s viewing experience could be affected by the expanded *complementarity* relations between verbal and visual *circumstances*. In comparison to Clip 3c, the target subtitles in Clip 3e offered more illustrative information for the actions depicted in the image. By adding ten extra circumstantial elements, the subtitles in Clip 3e heightened the main character’s emotion and intensified the tension in the scene. Figure 5.3 presents an example of the designed treatment.

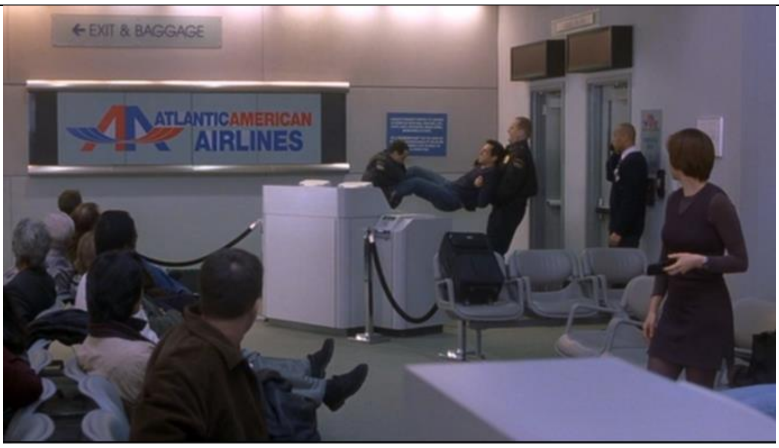
| | |
|---|---|
| **Background information** | Greg has a quarrel with the flight attendant on the airplane, resulting in being carried out of the plane by the security guards. |
| **Visual frames** |  |
| **Source subtitles** | Greg: Get off of me! Get off of me! |
| **CG subtitles** | 阿基: 放开我啊 放开我啊<br>[Get off of me. Get off of me.] |
| **EXG subtitles** | 阿基: ***快***放开我 ***快***放开我<br>[***Immediately*** get off of me. ***Immediately*** get off of me.] |

**Figure 5.3** Example of *expansion of complementarity* between *circumstances* in Clip 3

### *5.2.5 Variables*

The independent variable in this study is the version of the subtitles (no translation shifts in text-image relations vs. 10 cases of translation shifts of text-image relations). The dependent variables examined in this experiment are the participants’ visual attention, comprehension performance, and perception of subtitle quality.

The participant’s visual attention during watching the videos is gauged with eye-tracking data. As Smith (2013) suggested in his review of recent eye-tracking research in film studies, “eye tracking has the potential to provide a real-time insight into viewer cognition” (p.

185). Analyzing participants' visual attention through eye-tracking data can offer evidence of their cognitive processing while watching subtitled films. To this end, it is necessary to first define the area of interest (AOI), which is the spatial area "for which the eye-movement data will be extracted from the eye-movement recording for further analysis" (Godfroid, 2020, p. 159). Based on the research questions of this study, three groups of AOIs were designed as follows:

— ***Subtitle AOI***: the area where the subtitle texts are presented.

— ***Primary Image AOI***: the area where the subtitle-related visual objects are presented. It should be noted that arguably all visual objects can be related to subtitles to a greater or lesser extent. However, for the purpose of this experiment, the focus is put on the salient visual objects that are directly linked to the 10 manipulated subtitle texts in each clip, hence the term ***Primary Image AO****I*.

— ***Global AOI***: the area encompassing the entire visual frame. The Global AOI overlaps with the Subtitle AOI and the Primary Image AOI. The purpose of marking this area is to calculate the ***Secondary Image AOI***, namely subtracting the Subtitle AOI and the Primary Image AOI from the Global AOI. The Secondary Image AOI will allow us to know whether and how the viewers would explore the visual objects that are not directly related to the manipulated subtitle texts.

As discussed in Section 5.2.4, there are 10 cases of differences in the subtitled clips for the two participant groups. In each case, the AOIs were defined and drawn independently based on the specific subtitle line manipulated for between-group comparison. Figure 5.4 illustrates the drawing of AOIs for the subtitle line presented in Clip 1e (see also Figure 5.1 for the contextual details).

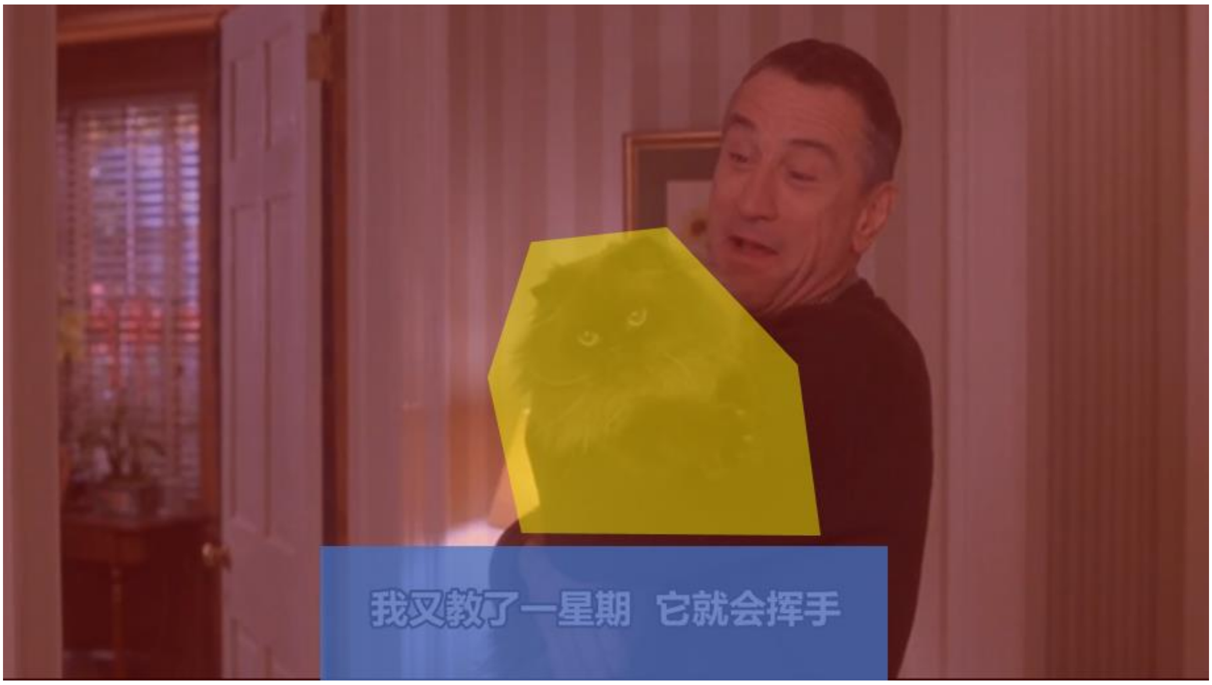


**Figure 5.4** Illustration of a subtitle AOI (in blue), a primary image AOI (in yellow) and a secondary image AOI (in red)

In this example, the subtitle AOI is represented by the blue square at the bottom of the screen. It covers a slightly larger area than the subtitle text to minimize the impact of some vertical drift in the recorded gaze data.

The primary image AOI, shown in yellow, covers the salient subject related to the subtitle content. In this particular example, the subtitle highlights the cat that greets people by waving its paw, making the cat the visually salient item signified by the subtitle. Sometimes, in the dynamic and multimodal film narrative, a salient object might be in motion rather than fixed. In such cases, a dynamic AOI is created for the subject, meaning that the image AOI moves along the same path as the object.

The global AOI, depicted in red, covers the entire visual frame. The secondary image AOI is the area outside the blue subtitle AOI and the yellow primary image AOI, i.e., the remaining area in red on the screen. For example, the face of Jack who is holding the cat, is covered by the secondary image AOI.

In addition to considering the spatial features of the defined AOIs, it is also essential to clarify the temporal design of the AOIs. In this study, the AOIs are only activated when the subtitles of interest are present on the scene (i.e., the 10 subtitle lines with translation shifts for EXG viewers and the corresponding subtitles lines without translation shifts for CG viewers).

In other words, eye-movement data are extracted for analysis from the moment a subtitle of interest appears on the screen until the moment the subtitle disappears[1].

To examine the participants' eye movements within the defined AOIs, four eye-tracking measures are chosen as follows:

— ***Dwell time (DT)***: the total duration of all fixation and saccades within an AOI, starting with the first fixation in an AOI to the end of the last fixation in the same AOI (Doherty & Kruger, 2018). A higher dwell time is an indication of higher processing difficulties and cognitive effort (Holmqvist et al., 2011). DT is reported in milliseconds (ms).

— ***Dwell time percentage of global dwell time (DT%)***: the proportion of DT on the subtitle or image AOI to DT of the global AOI. This measure, adapted from Kruger et al. (2014), is calculated by dividing DT on the subtitle or image AOI by DT on the global AOI and multiplying the result by 100 to obtain a percentage. While DT refers to the *absolute* visual duration among viewers, DT% indicates the viewers' *relative* visual duration by factoring in individual differences.

— ***Mean fixation duration (MFD)***: the average duration of each individual fixation within an AOI. To calculate this measure, the total fixation duration is divided by the total number of fixations within the AOI. Generally, there are "functional links between what is fixated and cognitive processing of that item – the longer the fixation, the 'deeper' the processing" (Holmqvist et al. 2011, p. 382). MFD is reported in ms.

— ***First fixation latency to the primary image AOI after subtitle reading* (*FFLIM*)**: the time elapsed before the first fixation on the primary image AOI after reading the subtitle. This measure is adapted from the more commonly used *first fixation latency* in AVT research (e.g., Black, 2022; Zheng & Xie, 2018). FFLIM captures the temporal span between the beginning time the viewer first fixates on the subtitle AOI to the time the viewer first fixates

[1] In some cases, the salient item on the image may linger slightly longer after the subtitle disappears. One possible approach to this issue is to extend the activated time for the image AOI in order to examine the potential prolonged impact of subtitle reading on image reading. However, the primary focus of this study is more about the effect of on-going interactions between the subtitle and the image on viewers' visual attention. Extending the activated time for the image AOI would introduce more noise in the potential "split-attention effect" on participants (Lin et al., 2016, p. 48), where the verbal and non-verbal information compete simultaneously for viewers' cognitive resources. Hence, this possible approach is not applied in this study.

on the primary image AOI[1]. A shorter FFLIM suggests that viewers, influenced by the subtitle, allocate their gaze more quickly to the focal visual information. The measure of FFLIM is reported in ms.

The second dependent variable in the experiment was the participants' comprehension performance. Comprehension has been a widely studied factor in previous reception studies of AVT (e.g., Desilla, 2014; Hu et al., 2020; Kruger et al., 2022). Moreover, from the perspective of target viewers, comprehension has been identified as their top concern when consuming subtitled audiovisual content (Wu & Chen, 2022). Both viewers and AVT researchers have shown extensive interest in video comprehension facilitated by subtitles. It is thus important to consider the potential impact of image-induced translation shifts on target viewers' understanding of the subtitled audiovisual content.

The final dependent variable in the experiment was the participants' perceived quality of the target subtitles. Quantifying and defining the quality of subtitling is challenging due to the varying industrial guidelines and academic standards (Díaz Cintas & Remael, 2021). Despite the absence of clear and systematic rules for target viewers, they can still form a general impression of subtitle quality. As Kuscu-Ozbudak (2022) observed, target viewers may associate a perceived decrease in subtitle quality with a decrease in content quality, leading to a loss of interest in the audiovisual content (e.g., unsubscribing from the audiovisual program). Hence, it is crucial to investigate this fundamental aspect of subtitling and examine how translation shifts in the text-image relations may impact viewers' perception of the target subtitle's quality.

The participant's preference for different subtitling methods was also investigated in the experiment. However, this variable was not included in the research hypotheses on the ground that, strictly speaking, it was not a dependent variable. It was more assumed to be *independent* of the audiovisual stimuli in the experiment. The participants' preferences are shaped by their personal knowledge and previous viewing experiences. Hence, it is more appropriate to regard their preferences as a possible *viewer-sociological variable* (see Section 2.2.2), which is not directly related to the variables investigated in this experiment but is worth exploring. As Božović (2022) states, "considering the needs, preferences, and attitudes of end-

[1] Similar to FFLIM, another possible measure is the viewer's first fixation time to the subtitle AOI after looking at the image. However, since this study focuses on how different subtitle versions may affect viewers' visual attention to the image, this measure is not considered in the experiment.

users is crucial in the decision and policy-making process and, thus, important for the industry" of AVT (p. 17). The analysis of viewers' preference for subtitling methods serves as a supplement to both the experimental results and the corpus findings in this study.

### *5.2.6 Instrument and data processing: eye tracking*

The participant's visual attention allocated to the subtitle and image areas was monitored by a remote eye tracker, Tobii TX300, set at the maximum sampling rate of 300 Hz for data collection. Tobii TX300 comprises an eye tracker unit and a removable 23", 1920x1080 widescreen monitor. The monitor acts as the presentation screen to display the audiovisual stimuli during the experiment. Importantly, in the experiment, the videos were presented in the central part of the monitor but did not occupy the full screen, as shown in Figure 5.5. Although it is technically feasible to make the videos fit the full screen, this deliberate adjustment was used to ensure less data loss and better data quality. As Holmqvist et al. (2011) pointed out, when the viewer's gaze moves towards the edge of the visual field of recording, "data will be gradually poorer" (p. 98). To this end, the videos were encoded with a resolution of 1280 × 720 (as mention in Section 5.3.1). Since the resolution of the videos was lower than that of the monitor, the videos were automatically resized and thus appeared smaller than the full screen.

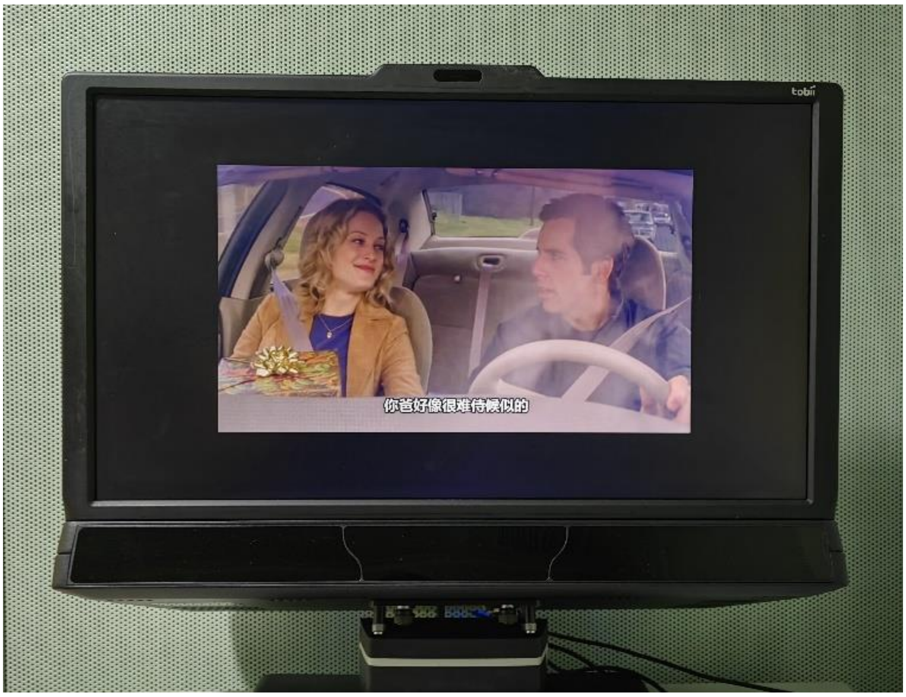


**Figure 5.5** Display of stimuli on the eye-tracking monitor

The eye-tracking recording and data analysis were conducted using the software package Tobii Studio 3.4.5. In cases where the data for a specific eye-tracking measure could not be directly retrieved from Tobii Studio (e.g., FFLIM), raw gaze data were exported and processed in R (version 4.3.0).

To detect and calculate fixation data, the I-VT Fixation Filter was used, a default fixation filter in Tobii Studio that classifies eye movements according to the velocity of the directional shifts of the eye. Based on the default setting of this filter, the minimum fixation duration was at 60 ms, with the eye-movement velocity threshold at 30 degrees/second (Muñoz, 2017; Liu et al., 2018). No upper limit on fixations was set in order to account for individual variability (Cui et al., 2023). To address fixations that were incorrectly split into multiple fixations (e.g., due to eye blinks), adjacent fixations were also merged, with the maximum time between separate fixations at 75 ms and the maximum visual angle of eyes between separate fixations at 0.5 degrees.

To ensure data quality for analysis, recordings from participants with a tracking ratio below 70% (assessed by Tobii Studio) were excluded (Hu et al., 2020; Fievez et al., 2023). Out of the total of 328 recordings obtained from 82 participants (each participant watching 4 clips), 58 recordings from 21 participants were discarded due to poor sample quality. To further denoise the data, recordings with noticeable data drift were removed. For example, some participants' gaze data consistently fell below the subtitle AOIs, indicating low accuracy in gaze detection during the recording. From a manual review of all the recordings, 40 recordings with severe drift data were excluded. Consequently, the final number of recordings included in the analysis was 230[1].

After eye-tracking data with poor quality were removed, 25 to 29 participants from each group were included for the final inter-group comparisons of their visual attention to the three clips (Clips 1-3). Although the eye-movement data were excluded for analysis from some participants, their data for comprehension, perception, and preference were still included in the analysis.

The statistical data analyses of the eye movements, as well as comprehension performance, perception rating, and preference comparison, were conducted using the statistic

[1] Overall, the proportion of sufficient data inclusion in this experiment was 70% (230/328), slightly below the general rate of over 80% observed in pervious eye-tracking research on subtitling (e.g., Szarkowska & Gerber-Morón, 2018; Liao et al., 2020). However, a lower data inclusion rate is not uncommon (e.g., 63.9% in Hu et al., 2020). This could be explained by Blignaut and Wium's (2014) observation that "the narrow eyes of Asian participants cause the eyetracker to lose the glint, and therefore, trackability is worse for Asians than for the other two ethnicities [Caucasian and African]. Asian participants also performed worse than the other two ethnicities with regard to accuracy and precision." (p. 75).

tools available in the packages in R (version 4.3.0), including *pastecs* (version 1.3.21, for descriptive statistics) and *stats* (version 4.3.0, for statistical tests such as the Mann-Whitney U test and the Fisher's exact test). To ensure the validity of statistical tests that require the assumption of data normality (e.g., t-tests), the Shapiro-Wilk test was performed to assess the normality of the data for each investigated variable within each participant group. In cases where the data for a variable were not normally distributed ($p < 0.05$) in both groups across all stimulus clips, the non-parametric Mann-Whitney U test was used instead.

### *5.2.7 Instrument and data processing: comprehension test*

To measure the independent variables of the participants' comprehension of the audiovisual content, comprehension tests were employed. Each video clip for the experiment was accompanied by a comprehension test with three open-ended questions, written in Chinese, the participants' mother tongue (for the complete comprehension tests and their translated English versions, see Appendix 3). The questions were designed to evaluate the participants' understanding of certain narrative details about the characters or the plot in the audiovisual stimuli. Open-ended questions were used here to prompt "richer, fuller and perhaps more genuine responses" from the participants about their comprehension (Coolican, 2019, p. 184).

Before further evaluation, the answers from the open-ended questions needed to be quantified. To convert these data into continuous values, the participants' comprehension performance was measured by the number of correct information points provided in their responses to the open-ended questions. Each information point was evaluated using the three-point rating scale adapted from Desilla (2014), as presented in Table 5.6. For instance, the comprehension questions for Clip 0 have a possible maximum of three correct information points, so the highest possible comprehension score for Clip 0 was 3*2 = 6.

**Table 5.6** Rating scale for comprehension

| **Score** | **Description** |
|---|---|
| 2 | The participants provided correct, clear answers. |
| 1 | The participants provided obscure, inconclusive answers. |
| 0 | The participants provided faulty answers or no answer. |

Two coders, the author (the first coder) and his chief supervisor (the second coder), were involved in rating the participants' comprehension performance. Initially, the first coder made a preliminary observation of the raw responses and developed a set of specific rating criteria. These criteria included examples of correct, partially correct, and incorrect answers for all the open-ended questions. Afterwards, based on the criteria, both coders independently rated all the comprehension responses. The inter-coder agreement was found to be 93.42%, indicating a high level of agreement. Any rating discrepancies between the coders were subsequently discussed and resolved.

### *5.2.8 Instrument and data processing: perception questionnaires*

To measure the independent variables of the participants' perception of subtitle quality, questionnaires were utilized. The perception questionnaire contained six items in the form of a six-point Likert scale ranging from 1 (not agree at all) to 6 (totally agree). With an even number of choices, the participants were not provided a neutral option and thus they had to choose a level of agreement or disagreement (Mellinger & Hanson, 2017). The questions were designed to gauge the participants' overall perception of the quality of the target subtitles based on Künzli's (2021) CIA model of subtitle quality. The model analyzes subtitle quality from three dimensions: *correspondence*, *intelligibility*, and *authenticity*. *Correspondence* refers to the degree of similarity (as specified by the client) between the source and target subtitles. *Intelligibility* pertains to the physical presentation and formulation of the subtitles, which may potentially affect viewers' comprehension. *Authenticity* concerns the natural use of the target language in the subtitles, independent of the co-presence of the original soundtrack. The relationship between the dimensions in the CIA model and the designed question items is outlined in Table 5.7. The complete perception scale originally written in Chinese and its translated English version can be found in Appendix 4.

In order to assess the reliability of the closed-ended questions, Cronbach's alpha coefficients were employed to measure the internal consistency across the CG and EXG participants. The obtained results, displayed in Table 5.8, reveal high Cronbach's alpha coefficients in both groups ($\alpha > 0.8$), suggesting that the six items utilized to evaluate perception of subtitle quality were highly reliable.

**Table 5.7** Questionnaire items in the six-point Likert scale

| Items | Measured variables | Dimension in the CIA model |
|---|---|---|
| 1 | Perceived naturalness of target subtitles | Authenticity: orality/idiomaticity |
| 2 | Perceived clarity of target subtitles | Intelligibility: readability-simplicity |
| 3 | Perceived conciseness of target subtitles | Intelligibility: readability-conciseness |
| 4 | Perceived speed of target subtitles | Intelligibility: perceptibility-speed |
| 5 | Perceived target subtitle-visual congruence | Correspondence: denotational |
| 6 | Perceived relevance of target subtitles to the traits and emotions of a film character | Correspondence: connotational |

**Table 5.8** Cronbach's Alpha Coefficients for perception of subtitle quality

| | Clip 0 | Clip 1 | Clip 2 | Clip 3 |
|---|---|---|---|---|
| **Cronbach's α** | 0.88 | 0.88 | 0.89 | 0.85 |

### *5.2.9 Instrument and data processing: interviews*

An interview was conducted at the end of each individual experiment. The interview followed a "semi-structured" format (Saldanha & O'Brien, 2014, p. 172), with a set of prepared questions but also allowing for flexibility to introduce new questions based on the participant's responses. The interview served as "supplementary sources of data" to explain participants' specific visual attention when watching subtitled films (Edley & Litosseliti, 2018, p. 210). In addition, the interview also examined the viewers' attitudes towards the investigated subtitling methods.

During the interview, participants were initially encouraged to freely discuss their thoughts about the experiment. They were asked about their feelings while watching the clips and their thoughts while completing the comprehension tests and the perception questionnaires. This free recalling session was followed by the semi-structured session. Both the CG version and the EXG version of subtitles for Clip 1, Clip 2 and Clip3 were presented to the participants. They were asked to closely examine the versions and answer the following questions:

1) Which version do you think is the one you just watched in the clip?

2) What is the major difference between the two versions?

3) Which subtitle version do you prefer, and why?

The participants' responses provided direct and genuine evidence of the general attitude towards the subtitling method frequently found in the corpus. Additional ad hoc questions were also raised during the interview if deemed necessary. For example, if the researcher/interviewer found in the replay of an eye-tracking recording that the participant focused on the image and skipped a subtitle line, they were asked to explain the observed behavior. If a participant displayed hesitation when filling out the questionnaires, they were asked for the possible reason behind it. Each individual interview lasted approximately twenty minutes.

All the interviews were audio-recorded with the participants' consent and then transcribed for subsequent analysis.

### *5.2.10 Procedures*

The participants were tested individually in an eye-tracking lab. For the CG participants, they were first briefed on the general objective of the experiment. Then, they were guided to sit before the monitor of the Tobii TX300 eye tracker, positioning their eyes at approximately 64 cm from the screen. To ensure accurate and precise tracking, participants were instructed to keep their head still during the experiment although their head was free to move.

The CG participants were presented with the four video clips (Clip 0, Clip 1c, Clip 2c, Clip 3c) sequentially. Prior to each clip, they underwent a nine-point eye-tracking calibration process. After watching each clip, they completed a questionnaire assessing their comprehension and perception of subtitle quality. To prevent cognitive fatigue, participants were given a few minutes of break after finishing each questionnaire. Upon completing the viewing of the four video clips, participants were asked to fill out a demographic questionnaire (see also Appendix 5), providing general personal information and indicating their proficiency in the dubbing language (Thai) used in the clips.

The EXG participants followed a similar process, with the exception that they watched the clips with different, manipulated subtitles (Clip 0, Clip 1e, Clip 2e, Clip 3e). At the end of the experiment, each participant was informed about the specific purpose of the study, followed by a semi-structured interview. The overall procedure of the experiment is illustrated in Figure 5.6.

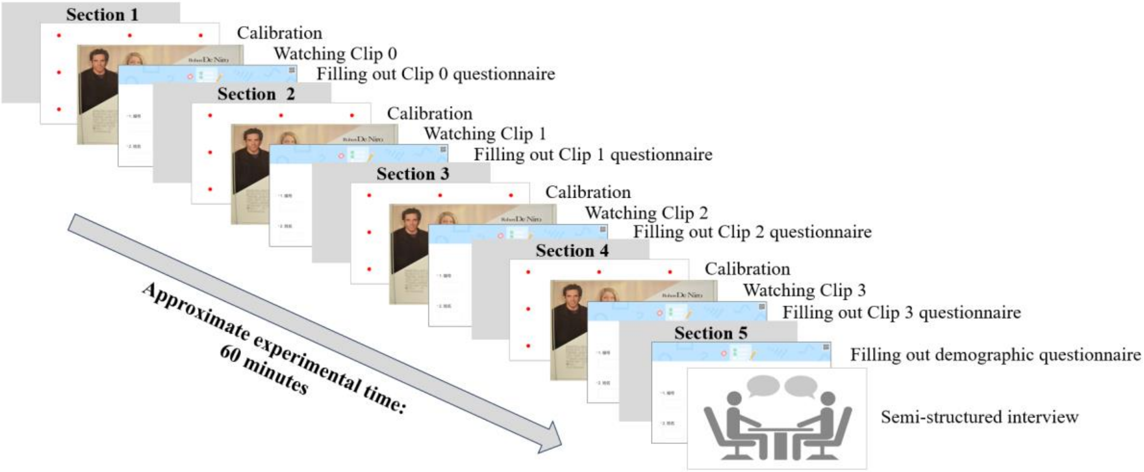

Section 1
Calibration
Watching Clip 0
Filling out Clip 0 questionnaire
Section 2
Calibration
Watching Clip 1
Filling out Clip 1 questionnaire
Section 3
Calibration
Watching Clip 2
Filling out Clip 2 questionnaire
Section 4
Calibration
Watching Clip 3
Filling out Clip 3 questionnaire
Section 5
Filling out demographic questionnaire
Semi-structured interview
Approximate experimental time:
60 minutes

**Figure 5.6** Experimental procedures

## 5.3 Quantitative results of the main experiment

### *5.3.1 Comparison of visual attention between groups*

As mentioned in Section 5.3.4, to investigate the participants' visual attention in the experiment, four eye-tracking measures are examined, including *dwell time* (DT), *dwell time percentage of global dwell time* (DT%), *mean fixation duration* (MFD), and *first fixation latency to the primary image AOI after subtitle reading* (FFLIM). To determine if there were statistical differences between the two groups, the non-parametric Mann-Whitney U test was employed given that the data deviated from a normal distribution and violated the assumption of normality required for t-tests (see also Section 5.3.5).

Table 5.9 presents the descriptive and statistical results for DT in the three groups of AOIs in each stimulus clip. Regarding visual attention in Clip 1, a significant difference was observed in DT in the primary image AOI between the two groups ($z$ = -2.269, $p$ = 0.023). Following conventional benchmarks for effect size (Cohen, 2013; Mellinger & Hanson, 2017), which classify effect sizes as small ($r$ = 0.1), medium ($r$ = 0.3), or large ($r$ = 0.5), the observed difference was of medium effect size ($r$ = 0.308). This suggests that EXG viewers, who watched subtitles with more *expansion of concurrence* shifts related to visual *processes*, allocated less visual attention to the corresponding visual information. On the other hand, no significant difference was found in DT in the subtitle AOIs ($z$ = -0.034, $p$ = 0.973), suggesting that the different subtitle versions elicited similar amount of visual attention from both groups. As the difference in DT in the global AOIs was not significant ($z$ = -0.963, $p$ = 0.336), it can be said that there was not a substantial individual difference in the total time the participants spent on the screen. In other words, the observed difference in DT in primary image AOIs was more likely attributable to the textual differences in the subtitles rather than inherent individual differences among the participants. In the case of Clip 2 and Clip 3, no significant difference in visual attention was observed. It appeared that the treatment conditions applied to the two groups did not result in noticeable variations in DT among the viewers.

**Table 5.9** DT in different AOIs and the Mann-Whitney test results

| **Clips** | **AOIs** | **N** | | **Median (ms)** | | ***z*** | ***p*** | ***r*** |
|---|---|---|---|---|---|---|---|---|
| | | **CG** | **EXG** | **CG** | **EXG** | | | |
| Clip 1 | Subtitle | 29 | 25 | 3385 | 3196 | -0.034 | 0.973 | -0.005 |
| | Primary image | 29 | 25 | 4796 | 4298 | -2.269 | **0.023** | -0.308 |
| | Global | 29 | 25 | 11796 | 11883 | -0.963 | 0.336 | -0.131 |
| Clip 2 | Subtitle | 29 | 29 | 9589 | 7987 | -1.396 | 0.163 | -0.183 |
| | Primary image | 29 | 29 | 6622 | 6725 | -0.278 | 0.781 | -0.036 |
| | Global | 29 | 29 | 23877 | 23801 | -1.145 | 0.252 | -0.150 |
| Clip 3 | Subtitle | 29 | 28 | 10947 | 10224 | -1.088 | 0.277 | -0.144 |
| | Primary image | 29 | 28 | 9771 | 11003 | -1.152 | 0.249 | -0.153 |
| | Global | 29 | 28 | 24793 | 24813 | -0.774 | 0.439 | -0.103 |

Table 5.10 displays the results for DT% in the AOIs. Similar to the results for DT in Table 5.9, a statistically significant difference with a medium effect size was found between the two groups' DT% in the primary image AOI for Clip 1 ($z$ = -2.316, $p$ = 0.020, $r$ = -0.315). This pattern further supports the observation in the DT results that EXG viewers tended to allocate less visual attention to the focal visual object signified by the target subtitles. In contrast, the insignificant difference in DT in the subtitle AOIs ($z$ = -0.173, $p$ = 0.862) indicates a similar distribution of visual attention in the subtitle area among the groups. Furthermore, another significant difference was observed in the secondary image AOI ($z$ = -2.325, $p$ = 0.020, $r$ = -0.316). As EXG viewers allocated more visual attention in the secondary image AOI than CG viewers, it suggests that EXG viewers processed the primary image AOI much faster and thus spared more time to explore the rest of the visual information in the scene. As for Clip 2 and Clip 3, no significant difference in DT% in the AOIs was observed, suggesting a similar impact of the different subtitle versions applied to the clips.

**Table 5.10** DT% in different AOIs and the Mann-Whitney test results

| **Clips** | **AOIs** | **N** | | **Median (%)** | | ***z*** | ***p*** | ***r*** |
|---|---|---|---|---|---|---|---|---|
| | | **CG** | **EXG** | **CG** | **EXG** | | | |
| Clip 1 | Subtitle | 29 | 25 | 29.6 | 27.3 | -0.173 | 0.862 | -0.024 |
| | Primary image | 29 | 25 | 41.4 | 36.0 | -2.316 | **0.020** | -0.315 |
| | Secondary image | 29 | 25 | 27.8 | 33.0 | -2.325 | **0.020** | -0.316 |
| Clip 2 | Subtitle | 29 | 29 | 40.4 | 37.4 | -1.291 | 0.197 | -0.169 |
| | Primary image | 29 | 29 | 28.3 | 28.8 | -0.583 | 0.560 | -0.077 |
| | Secondary image | 29 | 29 | 29.7 | 32.6 | -1.392 | 0.164 | -0.183 |
| Clip 3 | Subtitle | 29 | 28 | 44.2 | 41.5 | -1.253 | 0.210 | -0.166 |
| | Primary image | 29 | 28 | 39.5 | 44.3 | -1.006 | 0.315 | -0.133 |
| | Secondary image | 29 | 28 | 14.5 | 15.0 | -0.599 | 0.549 | -0.079 |

Table 5.11 shows the results for MFD in the AOIs. The analysis revealed no significant difference between the groups across the AOIs in all stimulus clips. It seems that, on average, EXG viewers did not engage in deeper cognitive processing of the textual or visual information compared to CG participants while watching the clips.

**Table 5.11** MFD in different AOIs and the Mann-Whitney test results

| **Clips** | **AOIs** | **N** | | **Median (ms)** | | ***z*** | ***p*** | ***r*** |
|---|---|---|---|---|---|---|---|---|
| | | **CG** | **EXG** | **CG** | **EXG** | | | |
| Clip 1 | Subtitle | 29 | 25 | 189 | 185 | -0.243 | 0.808 | -0.033 |
| | Primary image | 29 | 25 | 283 | 282 | -0.260 | 0.795 | -0.035 |
| | Global | 29 | 25 | 231 | 234 | -0.442 | 0.658 | -0.060 |
| Clip 2 | Subtitle | 29 | 29 | 225 | 199 | -1.346 | 0.178 | -0.177 |
| | Primary image | 29 | 29 | 329 | 342 | -0.381 | 0.703 | -0.050 |
| | Global | 29 | 29 | 270 | 248 | -0.677 | 0.499 | -0.089 |
| Clip 3 | Subtitle | 29 | 28 | 233 | 213 | -0.950 | 0.342 | -0.126 |
| | Primary image | 29 | 28 | 362 | 358 | -0.176 | 0.861 | -0.023 |
| | Global | 29 | 28 | 276 | 265 | -0.846 | 0.397 | -0.112 |

Table 5.12 presents the results for FFLIM in AOIs. No significant between-group differences were observed across the AOIs in all stimulus clips. Both EXG and CG viewers exhibited

similar response times in locating the key visual information in the film scenes. It suggests that the change in semantic relevance of the target subtitles to the image did not accelerate or defer viewers' cognitive attention to the visual information.

**Table 5.12** FFLIM in each stimulus clip and the independent samples t-test results

| Clips | AOIs | N | | Mean (SD) (ms) | | t | df | *p* |
|---|---|---|---|---|---|---|---|---|
| | | **CG** | **EXG** | **CG** | **EXG** | | | |
| Clip 1 | Primary image | 29 | 25 | 2409 (787) | 2206 (655) | 1.035 | 52 | 0.306 |
| Clip 2 | Primary image | 29 | 29 | 6388 (1535) | 6211 (1275) | 0.475 | 54 | 0.636 |
| Clip 3 | Primary image | 29 | 28 | 6335 (1657) | 6564 (1657) | -0.507 | 55 | 0.614 |

### *5.3.2 Comparison of comprehension between groups*

The participants' comprehension of the audiovisual stimuli was measured by open-ended questions and quantified by two coders using a rating scale that was developed for this study (see Section 5.3.6). Table 5.13 presents the results of the comprehension performance of the two participant groups.

**Table 5.13** Sum of comprehension scores and the Mann-Whitney test results

| Clips | Questions | N | | Median | | *z* | *p* | *r* |
|---|---|---|---|---|---|---|---|---|
| | | **CG** | **EXG** | **CG** | **EXG** | | | |
| Clip 0 | $Q_{sum1-3}$ | 40 | 42 | 6.0 | 6.0 | -0.097 | 0.923 | -0.011 |
| Clip 1 | $Q_{sum1-3}$ | 40 | 42 | 11.0 | 12.0 | -2.200 | **0.027** | -0.242 |
| Clip 2 | $Q_{sum1-3}$ | 40 | 42 | 15.0 | 15.0 | -0.172 | 0.863 | -0.019 |
| Clip 3 | $Q_{sum1-3}$ | 40 | 42 | 11.5 | 11.0 | -0.862 | 0.388 | -0.095 |

The results indicate that both groups had a similar level of comprehension performance for Clip 0 ($z$ = -0.097, $p$ = 0.923). This suggests that the cognitive abilities of the two groups were comparable, ensuring that any differences observed in the other clips were due to the textual variations intentionally included in the clips. Regarding Clip 1, the EXG participants demonstrated significantly better comprehension compared to the CG participants ($z$ = -2.200, $p$ = 0.027), and the observed difference was close to a medium effect size ($r$ = 0.242) (Cohen, 2013). In other words, viewers who watched films with more *expansion of concurrence* relations in the target subtitles regarding visual *processes* tended to have a better understanding of the actions in the mise-en-scène. As for Clip 2 and Clip 3, there were no significant differences between the groups in terms of their comprehension of the audiovisual content.

### *5.3.3 Comparison of perception of subtitle quality between groups*

The participants' perception of subtitle quality was evaluated using a set of six-point Likert scales. The scale consisted of six items and measured various aspects of subtitle characteristics: 1) naturalness, 2) clarity, 3) conciseness, 4) speed, 5) subtitle-visual congruence, and 6) relevance to the traits and emotions of film characters (see Section 5.3.6). Participants rated their agreement level on a scale of 0 (not agreeing at all) to 6 (totally agreeing) to indicate their perception of each aspect of the subtitle characteristics.

The test results from the Mann-Whitney test revealed a significant difference in the sum of perception scores between the groups for Clip 0, the baseline video ($z$ = -2.680, $p$ = 0.007, $r$ = -0.296). In this case, it is not advisable to directly compare the (non)differences between the groups in other clips, as these may be influenced by the initial baseline disparities.

To address the issue of baseline differences, the *gain score* approach was employed. Gain analysis is "a method of analysing pretest–posttest data when there is a significant difference in the pretest data" (Karim & Nassaji, 2020, p. 528). Despite its potential bias such as floor and ceiling effects, the approach still maintains "good Type I error control across all conditions" in the presence of moderate pre-test group differences (Jennings & Cribbiet, 2016, p. 218). A gain score is a participant's posttest score minus their pretest score (Lindstromberg, 2016). In this study, the gain score of perception was obtained from a participant's perception score of each of the three Clips (Clip 1-3) minus their perception rating of Clip 0. For instance, to assess the differential impact of the subtitling methods used in Clip 1, the *gain score* was obtained by subtracting the perception score of Clip 0 from the perception score of Clip 1 for each participant. Subsequently, the Mann-Whitney test was applied to compare the gain perception scores between the groups. The results are presented in Table 5.14.

According to the results shown in Table 5.14, there was a marginally significant difference in gain scores between the groups for Clip 2 ($z$ = -1.960, $p$ = 0.0499, $r$ = -0.216). The median gain score of perception for EXG was smaller than that of CG, hinting that EXG participants perceived worse subtitle quality than CG viewers. However, this result is only suggestive and inconclusive, given the small effect size and the marginal significance level.

**Table 5.14** Gain scores of Clips 1-3 and the Mann-Whitney test results

| **Clips** | **Items** | **N** | | **Median** | | ***z*** | ***p*** | ***r*** |
|---|---|---|---|---|---|---|---|---|
| | | **CG** | **EXG** | **CG** | **EXG** | | | |
| Clip 1 | Item$_{\text{gain1-6}}$ | 40 | 42 | 0.0 | 0.0 | -1.175 | 0.240 | -0.130 |
| Clip 2 | Item$_{\text{gain1-6}}$ | 40 | 42 | 1.0 | 0.0 | -1.960 | **0.0499** | -0.216 |
| Clip 3 | Item$_{\text{gain1-6}}$ | 40 | 42 | 1.0 | 0.5 | -0.192 | 0.848 | -0.021 |

### *5.3.4 Comparison of preferred subtitling methods between groups*

During the interview in the experiment, the participants were invited to compare the CG subtitles with the EXG subtitles in the three video clips. They were aware (through self-realization) or made aware by the author that each EXG subtitles contained ten additional instances of translation shifts in text-image relations. By indicating whether they favored either the CG or EXG subtitle version, participants expressed their personal preference for the subtitling methods.

Table 5.15 displays the distribution of the participants' preferences for the subtitle versions in each of the three EXG video clips. In other words, the Disfavored choice means a viewer prefers the CG subtitle version instead. Overall, the EXG participants showed a larger proportion of preference for the EXG subtitle versions compared to the CG participants, with 73.8% against 52.5% in Clip 1, 45.2% against 22.5% in Clip 2, and 92.9% against 75.0% in Clip 3. This difference in preferences could be attributed to the first impression bias (Asch, 1946; Lim et al., 2000), which refers to the tendency of human decision making to be influenced and overshadowed by initial information, leading to prejudiced evaluations of subsequent information. To assess the potential impact of the first impression bias, Fisher's exact tests were conducted to determine whether there was a significant relationship between the participant groups and their preferences. The results indicated a significant association between the participant groups and their preferences for subtitle versions in all three clips ($p$ = 0.012 in Clip 1, $p$ = 0.015 in Clip 2, $p$ = 0.024 in Clip 3). Therefore, it is important to avoid combining the two groups of frequency when interpreting the results. Instead, it is more appropriate to analyze the participants' preferences for subtitling methods by group.

**Table 5.15** Distribution of preferences for each investigated subtitling method

| Clips | Preference for subtitling method | N | | Percentage (%) | |
|---|---|---|---|---|---|
| | | CG | EXG | CG | EXG |
| Clip 1e | Favored | 21 | 31 | 52.5 | 73.8 |
| | Disfavored | 18 | 7 | 45.0 | 16.7 |
| | Ambiguous | 1 | 4 | 2.5 | 9.5 |
| Clip 2e | Favored | 9 | 19 | 22.5 | 45.2 |
| | Disfavored | 28 | 16 | 70.0 | 38.1 |
| | Ambiguous | 3 | 7 | 7.5 | 16.7 |
| Clip 3e | Favored | 30 | 39 | 75.0 | 92.9 |
| | Disfavored | 8 | 1 | 20.0 | 2.4 |
| | Ambiguous | 2 | 2 | 5.0 | 4.8 |

Table 5.16 presents the results of the 2 × 1 goodness of fit $\chi 2$ test (Wallis, 2021), which compared the frequency of the Favored choice and the Disfavored choice within each group across the three stimulus clips. In the case of Clip 1e, as indicated by the significance value ($\chi 2 = 7.000$, $p = 0.008$), EXG viewers significantly favored the subtitling method of *expansion of concurrence* between verbal and visual *processes*. However, the preference for this method among CG viewers was not significantly high. As for Clip 2e, CG viewers showed a significant preference for the subtitling method of *detachment of concurrence* between verbal and visual *participants* ($\chi 2 = 9.756$, $p = 0.002$), whereas EXG viewers did not exhibit a noticeable preference for it. In the case of Clip 3e, both groups of viewers unanimously and significantly voted for the method of *expansion of complementarity* between verbal and visual *circumstances* ($\chi 2 = 12.737$, $p < 0.001$ for CG; $\chi 2 = 28.125$, $p < 0.001$ for EXG). Overall, the preferences for subtitling methods in Clip 1e and Clip 2e appeared to vary between the groups, while the subtitling methods used in Clip 3e were appreciated by both groups.

**Table 5.16** Results of the 2 × 1 goodness of fit χ2 test for comparing Favored and Disfavored choices

| **Clips** | **Comparing pairs** | **CG** | | | **EXG** | | |
|---|---|---|---|---|---|---|---|
| | | N | χ2 | *p* | N | χ2 | *p* |
| Clip 1e | Favored | 21 | 0.231 | 0.631 | 21 | 7.000 | **0.008** |
| | Disfavored | 18 | | | 7 | | |
| Clip 2e | Favored | 9 | 9.756 | **0.002** | 19 | 0.257 | 0.612 |
| | Disfavored | 28 | | | 16 | | |
| Clip 3e | Favored | 30 | 12.737 | < **0.001** | 31 | 28.125 | < **0.001** |
| | Disfavored | 8 | | | 1 | | |

### *5.3.5 Summary of quantitative results*

The quantitative results from the previous sections are summarized in Table 5.17. In the case of Clip 1, where EXG viewers were exposed to more instances of *expansion of concurrence* between verbal and visual *processes* in the target subtitles, they demonstrated significantly better comprehension of the film narrative compared to CG viewers. However, it was observed that EXG viewers' improved comprehension performance was not necessarily attributed to their longer visual attention to the subtitle text or the salient visual entities. In fact, they allocated similar cognitive effort to subtitle reading but noticeably spent even less time processing the focal visual information compared to CG viewers. This pattern can be explained by further qualitative data, which will be discussed in Section 5.4.2. Regarding the viewers' preferences, EXG viewers significantly favored the subtitling method of *expansion of concurrence* between verbal and visual *processes,* whereas CG viewers didn't exhibit a noticeable preference for this method.

As for Clip 2, where EXG viewers watched target subtitles with more instances of *detachment of concurrence* in relation with visual *participants*, it was found that the perception of subtitle quality did not improve as much for EXG viewers compared to CG viewers. This significant difference in perception scores, albeit marginal, indirectly suggests that the EXG subtitle version was perceived as having lower quality. In terms of their personal preferences, CG viewers showed a significant preference for the subtitling method of *detachment of concurrence* between verbal and visual *participants*, but the preference for this method among EXG viewers was not significantly high.

**Table 5.17** Summary of the quantitative findings in the eye-tracking experiment

| **Clips (independent variables)** | **Comprehension of audiovisual content** | **Visual attention** | | | | **Perception of subtitle quality** | **Preference for subtitling methods** |
|---|---|---|---|---|---|---|---|
| | | DT | DT% | MFD | FFLIM | | |
| Clip 1 (*Expansion of concurrence* between *processes*) | EXG > CG | EXG < CG (PI AOI) | EXG < CG (PI AOI);<br>EXG > CG (SI AOI) | n.s. | n.s. | n.s. | EXG (favored > disfavored) |
| Clip 2 (*Detachment of concurrence* between *participants*) | n.s. | n.s. | n.s. | n.s. | n.s. | EXG < CG | CG (favored < disfavored) |
| Clip 3 (*Expansion of complementarity* between *circumstances*) | n.s. | n.s. | n.s. | n.s. | n.s. | n.s. | EXG & CG (favored > disfavored) |

Notes: The abbreviation “n.s.” means no statistically significant difference was found. The greater-than sign (>) and the less-than sign (<) indicate that a statistically significant result is observed in the inter-group or intra-group comparison. “PI AOI” refers to primary image AOI. “SI AOI” refers to secondary image AOI.

When it comes to Clip 3, where EXG viewers were exposed to more instances of *expansion of complementarity* between verbal and visual *circumstances*, no significant difference was observed in the dependent variables. The experimental conditions did not produce noticeable variations in viewers' comprehension, visual attention, or perception of subtitle quality. However, viewers from both groups, regardless of the subtitle versions they watched, demonstrated a strong preference for the subtitling method of *expansion of complementarity*. The inclusion of additional descriptive information in the target subtitles to modify the visual elements was well received by the viewers.

After presenting the quantitative results, a more comprehensive understanding of the viewers' perspectives is sought through a qualitative examination, which is presented in the next section.

## 5.4 Qualitative findings of the main experiment

### *5.4.1 H1: impact of expansion of concurrence between verbal and visual processes*

The subtitles designed for Clip 1 attempted to verify Hypothesis 1. It was assumed that incorporating additional explicit information from the concurrent visual actions into the target subtitles would result in reduced gaze time on the primary image areas, improved comprehension scores, and higher ratings for subtitle quality. Based on the statistical results, this hypothesis was partially confirmed.

As the statistical results indicate, EXG viewers, who watched target subtitles with greater textual reiteration of the visual actions, significantly allocated shorter gaze durations to the primary image AOIs and longer gaze time to the secondary image AOIs. An example of this is presented in Figure 5.7, which shows two heat maps generated from the gaze data of the two groups during the time frame when the manipulated subtitle was displayed on the screen (for more contextual details, see Figure 5.1). As observed from the red areas that signify longer absolute fixation duration, compared to CG viewers, EXG viewers exhibited higher concentration of fixations on the cat (the primary visual information) and on Jack's face (the secondary visual information). Since the CG subtitle did not explicitly specify the cat's behavior (waving its paw to greet people), CG viewers invested more cognitive attention in processing the primary visual information in order to understand the narrative. EXG viewers, on the other hand, seemed to grasp the key information in the narrative more quickly, allowing them more time to explore other relevant visual content, such as the facial expressions of the speaking character.

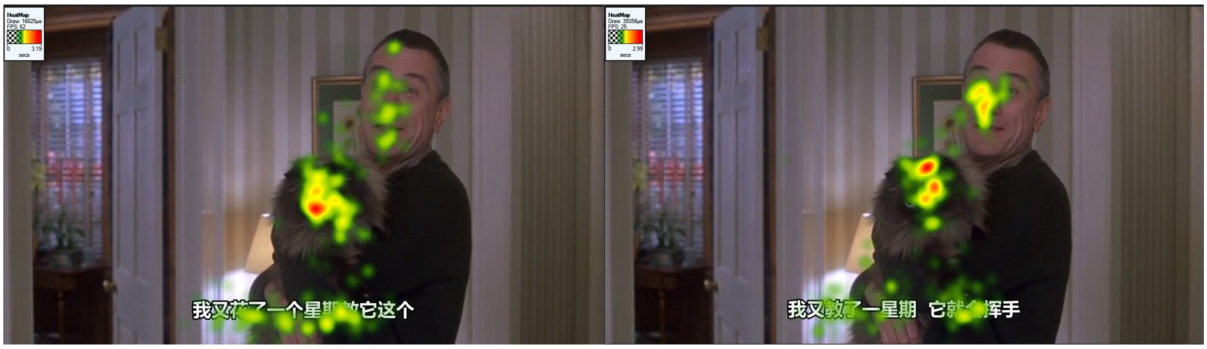


**Figure 5.7** Heat maps of the absolute fixation duration among CG (left) and EXG (right) on one subtitle line in Clip 1

This gaze pattern was also explained by some participants. As CGP138[1] commented:

> …because when he [Jack] said he taught it [the cat] this, the screen didn't seem to simultaneously show what was taught, so I kept seeking…I first located where the cat was, and then looked for which part [of it] was moving, and then I saw its paw, so the speed of locating it was a little slow…I think [to understand] this is still quite difficult...because the range of the paw's waving movement was actually not that large and it passed quickly.

In this sense, when the subtitle is not explicit enough, a viewer may be intensely preoccupied with the task of locating the essential information and as a result, they cannot afford additional attention to supplementary pictorial details in the film.

The increased difficulty in processing the narrative also accounted for the noticeably lower comprehension scores among the CG viewers. Many of the viewers who preferred the more explicit subtitle version commented that reiteration of the visual content in the text aided their comprehension and enhanced their perception of certain narrative details. For example, as expressed by EXGP238 and EXGP241, approximately 60-70% of their impression on a character's particular actions relied on the explicit verbal elements, while 30%-40% was attributed to visual cues. The multimodal reiteration seemed to improve their interpretation of

[1] The label for each participant consists of their group and participant number. For instance, CGP138 means that Participant 138 is from the Control Group. The interview extracts were translated from Chinese to English by the author of this thesis.

the narrative, resulting in a higher likelihood for the EXG viewers to recall the narrative details during subsequent comprehension tests.

However, prior to being informed about the subtitling method used in the video, the EXG viewers did not demonstrate a significantly higher gain score for the perception of subtitle quality. Although EXG viewers were not particularly aware of or sensitive about this issue when watching a subtitled film, they expressed a marked preference for this subtitling approach during the interview. A full list of the reasons motivating the viewers' dis/preferences is presented and discussed in Section 5.4.4.

### *5.4.2 H2: impact of detachment of concurrence between verbal and visual participants*

Hypothesis 2 assumes that removing the original referential information in the target subtitles may cause viewers to pay more visual attention to the visual references such as characters or objects, without affecting their comprehension scores or perception ratings of subtitle quality. This hypothesis was partially confirmed.

Initially, the author assumed that since the verbal information in the EXG subtitles did not explicitly mention the references depicted in the images, EXG viewers might allocate more visual attention to the images in order to identify the referred objects. However, as the statistical results for Clip 2 indicated, both EXG and CG viewers had similar visual attention in terms of DT, DT%, MFD, and FFLIM across different AOIs. Figure 5.8 illustrates the similar gaze patterns of two viewers, one from CG and the other from EXG, during the display of a manipulated subtitle line (for more contextual details, see Figure 5.2). As it shows, when the subtitle appeared, both viewers first looked at Greg's face (the speaking character on the right), then glanced at the subtitle, and finally switched their gaze back to the character's face.

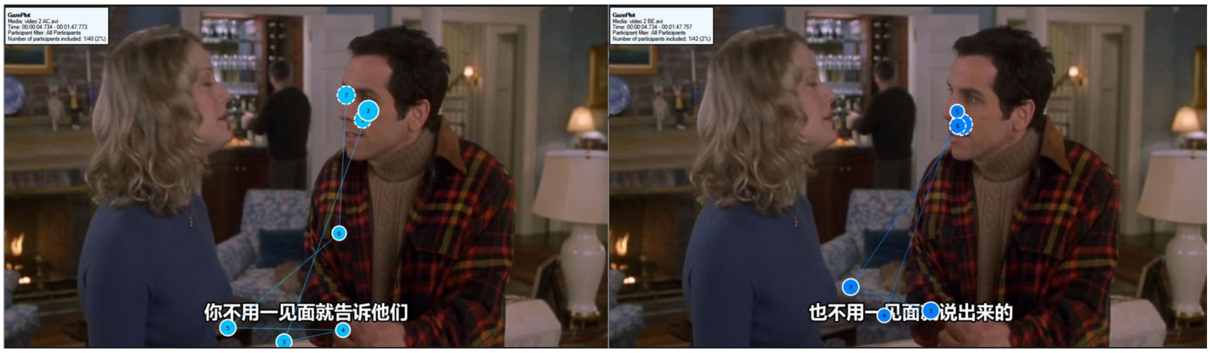

GazePlot
Media: video 2 AC.avi
Time: 00:00:04.734 - 00:01:47.773
Participant filter: All Participants
Number of participants included: 1/40 (2%)
你不用一见面就告诉他们
GazePlot
Media: video 2 BE.avi
Time: 00:00:04.734 - 00:01:47.757
Participant filter: All Participants
Number of participants included: 1/42 (2%)
也不用一见面说出来的

**Figure 5.8** Gaze paths of fixations from CGP119 (left) and EXGP242 (right) on one subtitle line in Clip 2

When asked about this viewing pattern, EXGP242 explained that his initial attention to Greg's face was because the character was speaking. Later, when the subtitle appeared, he naturally looked at it to see what Greg was talking about. More importantly, when EXGP242 looked back at Greg's face for the second time, he did not intend to confirm who was talking to whom. As he commented, "No, no, I was not confirming [who was speaking]…I just noticed him as he was speaking. I heard him speaking…so I could tell who was talking and whom he was talking to, without looking at the characters." This comment suggests that viewers can infer the implicit verbal referential information through other non-verbal channels, such as the visual channel that depicts the presence of the character and the aural channel from which the character's voice is heard. The removal of the original referential information in the target subtitles did not necessarily contribute to more effortful processing.

Similar to EXGP242's comment, many viewers (e.g., CGP107, CGP133, EXGP233) did not think the missing references in the subtitles would hinder their understanding of the narrative. These self-reported assessments were consistent with the viewers' comprehension performance, as there was no significant difference in comprehension scores between the two groups. Additionally, the CG subtitles, which contained clearer referential information, induced significantly higher gain scores for perception of subtitle quality. As CGP128 pointed out, "[I like the CG subtitles because] the references are a bit clearer. This [EXG version] is also okay because, by combining the image, [I] can still understand whom he is talking to, but it may take a little more mental effort to think." In this sense, the CG viewers may not be consciously aware of this issue while watching the subtitles, but they might subconsciously perceive the clarity of the subtitles and thus potentially result in a higher perception rating.

### *5.4.3 H3: impact of expansion of complementarity between verbal and visual circumstances*

Hypothesis 3 posits that reinforcing the visual kinesics of the mise-en-scène with more modifying elements in the target subtitles would lead to reduced visual attention on the visual information, improved comprehension performance, and better perception of subtitle quality. However, the statistical results for Clip 3 do not offer sufficient evidence to support this hypothesis, as no statistically significant inter-groups difference was found in all dependent variables

However, one observed phenomenon was worth noting. Many viewers were unable to answer the third comprehension question for Clip 3, which pertained to what the main character said when he was forced off the plane (for more contextual details, refer to Figure 5.3). While almost all viewers in the experiment correctly answered what happened to the character (i.e., being carried out of the plane), a noticeable number of them could not recall the character's line. Upon replaying the eye-tracking recordings during the follow-up interview, it was discovered that even those who had read this short subtitle line once or twice still could not remember it (e.g., CGP109, CGP118; EXGP216, EXGP231, EXGP234), as shown in some examples presented in Figure 5.9. EXGP234 admitted that she initially glanced at the subtitle and then focused solely on the visual action. Although she also visited the subtitle area for the second time (indicated by Fixation Number 8 in the figure), she believed that she was not attentively processing the text. Similar to EXGP234's view, EXGP231 also acknowledged that the subtitle line was not crucial for the main plot's development and that she usually read subtitles selectively based on the narrative and context. These qualitative explanations further imply that adding more modifying elements to the target subtitles may not enhance viewers' comprehension or memory of narrative details in a film. Not all viewers read subtitles attentively but rather selectively. In particular, when the visual information is highly captivating (such as during a big fight), the importance of subtitles appears to diminish.

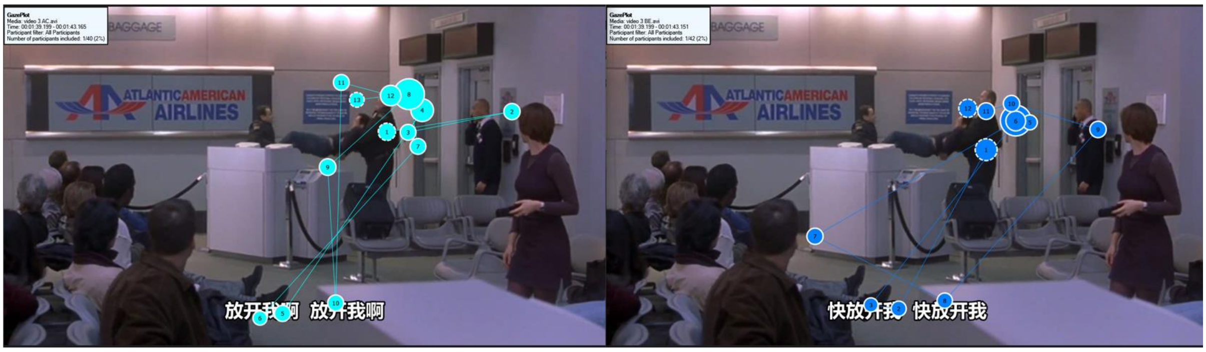

GazePlot
Media: video 3 AC.avi
Time: 00:01:39.199 - 00:01:43.165
Participant filter: All Participants
Number of participants included: 1/40 (2%)
放开我啊 放开我啊
GazePlot
Media: video 3 BE.avi
Time: 00:01:39.199 - 00:01:43.151
Participant filter: All Participants
Number of participants included: 1/42 (2%)
快放开我 快放开我

**Figure 5.9** Gaze paths of fixations from CGP118 (left) and EXGP234 (right) on one subtitle line in Clip 3

### *5.4.4 Viewer preference for translation shifts in text-image relations*

To provide a more comprehensive understanding of the viewers' perspectives on the translation shifts in text-image relations examined in the experiment, some qualitative findings obtained from the semi-structured interviews are presented.

Table 5.18 summarizes the main reasons for the viewers' varying preferences for the subtitling methods, along with some examples quoted from the interviews. Overall, the viewers chose their preferences for four major reasons: linguistic (e.g., violating original meaning, improving naturalness), cognitive (e.g., facilitating comprehension, inducing cognitive burden), multimodal (e.g., sufficient multimodal cues), and filmic (e.g., enhancing characterization).

**Table 5.18** Major reasons and examples for different subtitling preferences

| Clips | Preferences | Major reasons | Examples |
|---|---|---|---|
| Clip 1 (*Expansion of concurrence* between *processes*) | Favored | Facilitating comprehension | "Watching the [EXG] subtitles can assist understanding of what the film is talking about…Maybe sometimes, for example, if you don't see clearly [the visual content] and you don't have the help of the subtitles, you may have to play back [the video]." (EXGP209) |
| | Disfavored | Limiting room for interpretation | "Maybe sometimes you may not have a very deep impression of some details, but I still hope that the [subtitle] version doesn't have to explain all the details so clearly…If some [visual] actions are not clearly stated [in the subtitles], it gives you more space for interpretation and imagination." (CGP110) |
| | | Violating original meaning | "…this is not the original flavor, and it may be a bit different from the intention of the original author or the original screenwriter." (CGP105) |
| | Ambiguous | Sufficient multimodal cues | "If it is combined with the video screen, I don't think [the two versions] matter, because I can know his actions from the image, so I don't necessarily need to watch the subtitles, or maybe just take a glace for a second or two if I don't know what he is talking about. |

|  |  |  |  |
|---|---|---|---|
|  |  |  | Then I can make associations by myself through the images.” (EXGP207) |
| Clip 2 (*Detachment of concurrence* between *participants*) | Favored | Improving naturalness | “Omitting [the referential elements] makes me feel more comfortable…It reads more natural, more like daily life conversation.” (EXGP223) |
|  |  | Improving conciseness | “The omitted version is more concise and [I] can still understand the meaning.” (EXGP212) |
|  | Disfavored | Inducing cognitive burden | “Keeping the referential information may make me understand faster…or give me deeper memory.” (EXGP124) |
|  | Ambiguous | Sufficient multimodal cues | “I have no preference because I can understand [either way]. Additionally, even if I watch the version that retains the referential elements, I don’t necessarily read the entire subtitle line but just grab a few key words, and then look at the image, and then I know what is happening.” (CGP237) |
| Clip 3 (*Expansion of complementarity* between *circumstances*) | Favored | Enhancing characterization | “Just by reading this [EXG subtitle], [I can tell] he is very angry. If you didn’t look at his expression and voice, you couldn’t tell that he is very angry.” (CGP140) |
|  | Disfavored | Over-enhancing characterization | “His emotions can be conveyed through the image and sound, and there is no need to modify them in subtitles.” (CGP129) |

|  |  |  |  |
|---|---|---|---|
|  | Ambiguous | Balancing characterization and formality | “Both [subtitle versions] are good. This [CG version] reads more natural [although not emotionally strong]; this [EXG version] has some emotionally strong words…[but reads] very formal.” (EXGP205) |

Regarding the subtitling method used in Clip 1e, which tended to explicitly incorporate concurrent visual kinesics in the target subtitles, many viewers expressed a preference for this approach, stating that it helped them better understand the narrative details in the film. However, some viewers disliked it because they felt it limited their own individual interpretation and deviated from the original meaning conveyed in the dialogues.

As for the subtitling method applied in Clip 2e, in which referential information such as pronouns and character names were omitted from the source dialogues, many supporters felt that it made the subtitles read more natural and resembled everyday conversation. However, some viewers disliked this approach as it placed a heavier cognitive burden on their viewing experience.

In the case of Clip 3e, where additional elements were added to the target subtitles to describe or modify the visual kinesics, the majority of viewers from both groups favored this approach. They believed it enhanced the traits and emotions of the characters and strengthened the congruence between the text and the visual content. One participant (EXGP237) even expressed disbelief at how plain and emotionless the source dialogues were. Nevertheless, a few viewers disapproved of this approach, considering it to be an over-translation.

For all the three subtitling methods, some viewers believed that both versions were satisfactory. They were tolerant of those linguistic nuances because they believed they could extract the verbal and non-verbal information and combine them to achieve a sufficient understanding of the audiovisual text as a whole. The qualitative analysis reveals the diverse needs and expectations of viewers, suggesting that there is unlikely to be a single optimal subtitle version that satisfies all.

## 5.5 Discussion

### *5.5.1 Summary of experiment findings*

When it comes to audiovisual texts (e.g., films), it is not clear yet how viewers process messages from different modes and which translation methods may achieve the best viewing (Tuominen et al., 2018). The main objective of the main experiment was to obtain empirical evidence and examine viewers' distribution of visual attention across different meaning-making modes and their reception of subtitling as multimodal representation.

As the experimental results for Hypothesis 1 indicated, the reiteration of primary visual information in the target subtitles (i.e., *expansion of concurrence* between verbal and visual *processes*) significantly improved target viewers' comprehension of the audiovisual narrative, enabling them to make sense of the visual information better and faster. As long as the length and speed of the subtitles remained the same, the inclusion of more explicit words in the subtitles did not cause any additional visual attention demands during subtitle reading. In other words, it did not distract viewers from reading the text but facilitated them to understand the ongoing narrative more expediently and accurately.

Hypothesis 2 in the experiment was partially supported, suggesting that the removal of referential information from the original dialogues (i.e., *detachment of concurrence* between verbal and visual *participants*) did not result in different visual attention or comprehension performance. However, it may result in worse perception of subtitle quality, as indicated by the marginally significant result.

Insufficient statistical evidence prevented the confirmation of Hypothesis 3. It was found that reinforcing characterization by adding more linguistic elements to modify or describe the visual kinesics (i.e., *expansion of complementarity* between verbal and visual *circumstances*) did not lead to differences in visual attention, comprehension performance, or perception of translation quality. Nevertheless, interviews conducted with the viewers revealed a significant preference for this subtitling method, as they believed it highlighted the traits and emotions of the characters and strengthened the congruence between the text and the visual content in the audiovisual narrative.

The qualitative analysis further elucidated the underlying factors influencing the observed viewing patterns and reception. As pointed out by the viewers during the interview, their comprehension of the audiovisual narrative was derived from both verbal and non-verbal information. Non-verbal cues were an important source of information, especially when certain content was not explicitly conveyed through subtitles. Viewers' visual attention to subtitles was selective and constantly affected by the co-occurring visual content. Their needs and preferences for subtitling methods were found to be diverse and individualized. It was thus difficult to assume target viewers as a homogenous group (Sasamoto et al., 2021).

### *5.5.2 Subtitle reading as multimodal viewing*

The preceding observations stemming from the experiment underline the potential influence of target subtitles in guiding, or even manipulating, the cognitive attention of target viewers towards the audiovisual content. This study contends that subtitle reading is an integral part of multimodal viewing.

Based on the results for Hypothesis 1, when the target subtitle reiterates the focal visual kinesics in the image (as a case of text-image redundancy), it leads to better comprehension of the audiovisual narrative and faster processing of the corresponding primary visual information. This finding aligns with Bairstow's (2011) observation that "subtitles not only promote dialogue information processing but also the integration of visual data" (p. 217).

The effect of faster processing of primary visual content can be attributed to the reduced cognitive load resulting from the integration of congruent textual and pictorial information (Sweller et al., 1998; Ayres & Sweller, 2005). In audiovisual viewing, information from different meaning-making modes competes for the viewers' cognitive resources and thus causes the effect known as "split attention" (Kalyuga et al., 1999), which "occurs when several sources of information are difficult or impossible to understand in isolation and must be mentally integrated to achieve" (p. 367-368). When watching an audiovisual program in a foreign language, viewers are constantly processing information from various sources, including the original dialogues and sounds (aural mode), the subtitles (verbal mode), and the image (visual mode) (d'Ydewalle & De Bruycker, 2007). In this context, rendering certain visual content more explicitly in the target text is likely to aid viewers in locating and correlating relevant information, thus alleviating the split-attention effect.

The effect of improved comprehension resulting from verbal-visual redundancy provides further evidence for the predictions outlined in Kruger and Liao's (2022) multimodal-integrated language framework. According to the framework, during multimodal viewing, viewers construct their comprehension and memory of the audiovisual content by integrating meanings through their visual and auditory channels that perceive information from multiple modes. Initially, they perceive meaning conveyed independently in each mode (e.g., observing a cat waving its paw in the visual mode and reading the corresponding word "wave" in the verbal mode). Subsequently, they combine and coordinate the perceived information in their minds, ultimately forming their understanding and memory of the multimodal narrative. Verbal-visual redundancy facilitates a process known as "parallel processing" (Kruger & Liao, 2022, p. 30), enabling viewers to simultaneously process the same information from both modes rather than one single mode. For instance, when viewers allocate their attention to the word "wave" in the subtitle, they can at the same time make sense of the corresponding visual action of the cat by using their peripheral vision (i.e., seeing things outside the central vision). Compared to a less specific subtitle which requires viewers to rely solely on the image for certain information, a more explicit subtitle provides dual sources of the same information and thus potentially enhances the cognitive processing of the multimodal narrative. While Liao et al. (2021) discovered that concurrent pictorial content enhanced comprehension despite less time allocated to reading subtitles, the findings of this study further confirm another effect, that is, concurrent textual content facilitated comprehension even when less time was spent on the pictorial content. Verbal-visual redundancy can be a "belt and braces" subtitling method in case viewers miss either the textual or pictorial information in the audiovisual product.

Furthermore, the verbal-visual redundancy information did not impose higher cognitive demands on subtitle reading, as both participant groups in the experiment devoted similar gaze time to the subtitle area. This pattern tallied with previous observations that subtitle reading was effortless (Perego et al. 2010; d'Ydewalle & Gielen, 1992) and selective (Wang & Pellicer-Sánchez, 2022). The target viewers did not diligently read every word of the subtitle, but rather selectively, usually with ease, extracted the essence of the textual information. They would then shift their attention back to the image, following the dynamic multimodal narrative. While the explicitness of the subtitles did not affect the overall subtitle reading process, it could potentially counteract "subtitling blindness" (Romero-Fresco, 2018, p. 252), the phenomenon where the viewer fails to notice or fully appreciate the visual details of the mise-en-scène due to being excessively engaged in reading subtitles. A more explicit subtitle line (or at least regarding to the visual kinesics in the present study) could help viewers interpret the intended visual message more effectively and rapidly, allowing them more time to appreciate other non-verbal components that capture their interest within the audiovisual text. The target subtitle thus functions "as another cinematic tool to influence the viewers' attention and engagement with the film" (Romero-Fresco, 2018, p. 244).

Different from prior studies on subtitling that focused on the impact of verbal-audio redundancy between the target subtitles and the original soundtrack (Bisson et al., 2014; Liao et al., 2022; Perego et al., 2010), the present study analyzed verbal-visual redundancy between the target subtitles and the image. Notably, despite the different operationalizations of multimodal redundancy and the various methods used to measure comprehension (e.g., opened-ended comprehension questions focusing on specific aspects of the audiovisual content in this study vs. three-alternative-choice questions for more general comprehension in Liao et al.'s (2022) study), the present study are consistent with previous research that underscores the positive influence of multimodal redundancy on viewers' reception.

Another significant difference observed in the experiment, albeit marginal, was a more favorable perception of subtitle quality when certain referential information in the original dialogues was maintained, in comparison to when the same information was omitted. In previous AVT research, the subtitling method of reduction was deemed as necessary (Gottlieb, 1992; Georgakopoulou, 2009) and it was assumed to not adversely impact viewers' comprehension due to the presence of accompanying non-verbal cues (Bączkowska, 2011; Taylor, 2004). The corpus analysis in Section 4 further corroborated this prevalent subtitling practice of eliminating referential components such as pronouns and character names. The experimental outcomes also provided evidence to support the previous assumption that this subtitling method did not hinder comprehension. However, the experimental results suggested that this subtitling method seemed not to be applauded by viewers. Instead, some viewers noted during the interviews that they favored subtitles characterized by clarity and transparency. While these viewers were indeed able to deduce implicit referential information from the concurrent non-verbal cues, they expressed a reluctance to expend excessive effort in making sense of the narrative. This sentiment echoed Tuominen's (2011) observation that target viewers tended to engage with subtitles in a superficial manner. The viewers' expressed discontent with reduction in subtitling (or at least the reduction of referential information regarding the *participants* conveyed in both the source dialogues and the image) contradicts the prevalent belief in both the subtitling industry and academia that text condensation is indicative of high-quality subtitles (Szarkowska et al., 2021). Szarkowska et al. (2021) discovered that text reduction in the target subtitles was considered acceptable when viewers were unable to comprehend the original dialogues, but it became problematic when viewers understood the spoken language and were able to discern the reduced content from the original dialogues. The finding of this study further complicated this matter by revealing viewers' negative reception of reduction even when they had limited knowledge of the spoken language. Combining the insights from the primarily text-based perspective in Szarkowska et al. (2021) and the multimodal perspective in the current study, it appears that, irrespective of viewers' proficiency in the source language, their expectations of reduction in subtitles diverge from the current industry practices and the conceptualization in prior AVT research.

Despite target viewers' disfavor for text condensation, they strongly advocated the subtitling method of reinforcing character traits by adding adverbial words or phrases to complement the character's visual kinesics. This observed preference echoed Messerli's (2019) remark that subtitles function as "communicative agents" (p. 532), which, independent of the source dialogues, (re)speak to viewers about the film narrative on behalf of a character, the film director, or the film itself. Some viewers during the interview (e.g., EXGP209, EXGP224) further confirmed this view, stating that the tone of the target subtitles should align with the emotion of the depicted characters based on their non-verbal signs (e.g., gestures, facial expressions), regardless of the wording of the original dialogues. Viewers' preference for this subtitling method can be attributed to the stronger relevance between the verbal and non-verbal information (Ortega, 2011). By enhancing its relevance to the non-verbal content, the target subtitle alleviates viewers' processing effort (Gambier, 2018).

Taken together, it appears that target viewers obtain better reception and enjoyment when subtitles establish a closer and clearer connection with the visual content. Viewers seem to prefer less cognitively demanding subtitles, which aligns with prior research findings (Szarkowska & Gerber-Morón, 2019). They also favor subtitles that enhance characterization by strengthening the traits or emotions of the film characters.

The observed impact of target subtitles on the viewing experience and reception underlines the power of subtitlers. By manipulating the interaction between subtitles and the pictorial content, the subtitlers exercise their "agency", demonstrating their "willingness and ability to act" during the translation process (Kinnunen & Koskinen, 2010, p. 6). In this context, the subtitler's agency does not necessarily imply an ideological standpoint (Tymoczko, 2007) or a political orientation (Baker, 2019). Instead, it emphasizes the subtitler's creativity from a multimodal perspective. Subtitlers and their work are often expected to maintain an invisible presence in AVT viewing (Kuo, 2015; Szarkowska et al., 2021). However, this does not imply that subtitlers must passively adhere to a literal translation approach from the source text. Instead, they may actively employ their linguistic creativity by incorporating other non-verbal meaning-making resources in the target product. To concoct a more effective subtitle design, the subtitlers should keep in mind their role as multimodal representors and their potential power in directing viewers' cognitive attention, not only to the text but to the entire multimodal composition in the audiovisual text (Lautenbacher, 2012).

# CHAPTER 6 CONCLUSIONS

## 6.1 Major findings

### *6.1.1 RQ1: Text-image relations in subtitled audiovisual products*

This thesis examines the intricate relationship between text and image within subtitled audiovisual products. The first research question of the thesis is dedicated to exploring the various text-image relations present in subtitled films. Based on previous frameworks about text-image interactions and drawing on insights from systemic functional grammar and visual grammar, the author proposes a theoretical framework of text-image relations specifically for the study of subtitling. The framework identifies four major categories of text-image relations within audiovisual texts, referred to as the 4 Cs: *Concurrence* (where the meaning of text aligns with that of image), *Complementarity* (where text further modifies or elaborates image), *Condensation* (where text is less specific than image), and *Contradiction* (where text contrasts with image).

As further revealed by the results from the self-built multimodal corpus of subtitled films, the linguistic content (i.e., dialogues) in films appeared to be more semantically specific than that of the visual content (i.e., images), with the observed frequency of *complementarity* relations substantially higher than that of *condensation* relations. Moreover, the linguistic and visual content in films tended to exhibit a high degree of congruence, as evidenced by the substantial number of *concurrence* relations identified. These observed patterns of interaction between text and image provide insights into the complex dynamics that underpin the interplay of language and visuals within the realm of AVT.

### *6.1.2 RQ2: Translation shifts in text-image relations through subtitling*

The second research inquiry within this thesis pertains to the extent to which interlingual subtitling preserves or alters text-image relations in audiovisual products. To address the question, another framework of translation shifts is proposed, with a bottom-up approach based on actual cases of shifts in text-image relation observed in the film corpus. Five major types of translation shifts are identified: *non-shifts*, *obligatory shifts*, *preferential shifts*, *strengthening shifts* and *weakening shifts*. The first three types of shifts are construed as language-induced shifts, resulted from the subtitler's linguistic consideration between the target and source languages. The latter two types are regarded as image-induced shifts, arising from the subtitler's conscious or unconscious attention to non-verbal elements.

The results of the corpus analysis unveiled that interlingual subtitling, functioning as a multimodal representation, tended to **reiterate**, **remove**, or **reinforce** the interactions between text and image in subtitled films. Firstly, the target subtitles were frequently found to reiterate the co-occurring kinesics depicted on the image, making the textual information in the target subtitles more explicit than the original dialogues. This subtitling method strengthened the text-image interplay by creating a new *concurrence* relation in the target products. Secondly, the target subtitles tended to remove from the source dialogues the referential information (e.g., character name, pronouns) that was visually present on the screen. This subtitling method weakened the text-image interaction by erasing the original *concurrence* relation. Lastly, the target subtitles tended to provide additional specification for the visual kinesics. This subtitling approach strengthened the text-image connection by introducing a *complementarity* relation in the target products.

Overall, interlingual subtitling seems to intricately manipulate the interdependency of text and image, either by strengthening or weakening their connection. This multifaceted interplay underscores the dynamic multimodal nature of AVT.

### *6.1.3 RQ3: Reception impact of translation shifts in text-image relations*

The final research question in the thesis explores the reception of subtitling as multimodal representation, examining how translation shifts in text-image relations affect viewers' distribution of visual attention, comprehension, and perception of translation quality. Empirical data were obtained from 82 participants in a between-subject experiment using eye tracking, comprehension tests, perception rating scales, and semi-structured interviews.

As the results indicated, viewers exhibited significantly better comprehension and allocated less gaze time to the primary visual information when watching films with target subtitles that provided more explicit information about the visual actions. The verbal-visual reiteration seemed to reduce the viewers' cognitive effort, facilitating more effective and rapid processing of the audiovisual narrative. Moreover, viewers tended to perceive lower subtitle quality when target subtitles contained less referential information related to the visual objects. While this subtitling method did not necessarily hinder viewers' comprehension or affect their distribution of visual attention, it received a substantial amount of negative feedback from the control group and split feedback from the experimental group. Some viewers expressed during the interviews that they favored subtitles with clarity and transparency and that they were reluctant to invest excessive effort in deciphering the narrative. Finally, both groups strongly favored the subtitling method that provided more descriptive information to modify the visual kinesics. As perceived by the viewers, this approach could enhance the traits and emotions of the characters and strengthen the congruence between the text and the visual content. When considering the three investigated subtitling methods in the experiment, the viewers explained their preferences on four major grounds, i.e., linguistic features, cognitive effort, multimodal composition, and filmic narration. It appeared that viewers' needs and expectations for subtitling methods were varied and individualized.

These experimental findings highlight the potential influence of subtitles in guiding the cognitive attention of target viewers towards the audiovisual content. The author contends that subtitle reading is an integral part of multimodal viewing. Target viewers may obtain better reception and enjoyment when subtitles establish a closer and clearer connection with the visual content. Viewers are more inclined towards subtitles that are less cognitively demanding. For a more effective subtitle design, the subtitlers should bear in mind their potential to guide viewers' cognitive attention, not only towards the text but the overall multimodal composition within the audiovisual product.

## 6.2 Contributions

Subtitles that are literally transferred from the source language are only *halve vertalingen* (Kriek, 2002, cited in Díaz Cintas & Remael, 2007, p. 57), or in other words, "half-way translations". In terms of theorization, this project attempts to (re)conceptualize subtitled products as multimodal ensembles by highlighting the interplay between image and subtitles. By synthesizing research insights from corpus-based multimodal analysis and audience reception, it contributes to moving translation studies in the direction of multimodality and expands the exclusive focus of traditional translation studies on texts that are "verbal only" (Gottlieb, 2012, p. 38). Its theoretical contribution also lies in that it extends the line of AVT research by seeking to develop two theoretical frameworks to systematically describe the interplay of verbal-visual modes in subtitled films. In most previous AVT research on text-image interplay, the analytical unit of visual-verbal interaction has received little attention. For instance, it usually remains unclear whether a word or a phrase in the subtitle interacts with the entire image, or whether it is the entire sentence that relates to only a part of the image. To address this gap, this project draws on transitivity systems in systemic functional grammar and visual grammar to propose frameworks that clarify the basic comparative unit for analyzing text-image relations, namely, the verbal and visual *participants*, *processes*, and *circumstances*. Furthermore, in contrast to previous analyses of translation shifts in AVT studies, which have primarily concentrated on monomodal shifts within language alone, the framework of translation shifts proposed in this project considers other non-verbal elements. Through the establishment of a well-defined analytical unit and the incorporation of a multimodal shift analysis, the two frameworks address "[t]he challenge of (…) designing a systematic framework for the analysis of multimodal texts" in AVT studies (Remael & Reviers, 2019, p. 260).

The methodological contribution of this project is threefold. First, it is one of the first few attempts that combine corpus and experimental evidence in audiovisual reception studies. The corpus findings encapsulate real-life subtitle practices while the results from the experiment and interviews represent the voice of subtitle viewers. The combination of both approaches facilitates the elucidation of potential disparities between academic theories, industrial practices, and end-user expectations (Wu & Chen, 2022). Second, this project has developed a coding scheme (see Figure 4.1) for annotating multimodal corpora of subtitled films. The coding scheme will provide a set of heuristics for the multimodal analysis of subtitled films in future studies. AVT researchers have tended to select film materials based on convenience sampling, the generalizability of which is limited. Corpus-informed findings concerning the types of text-image relations and their distribution patterns will enable researchers to ascertain whether the semiotic phenomena are representative and important enough to call for in-depth experimental studies. Third, one research product of this project is a corpus of subtitled films with multimodal annotation. It demonstrates that multimodal corpus analysis is applicable to the context of English-Chinese subtitling, thus serving as an important addition to existing multimodal corpora that have been primarily sourced from films subtitled in European languages (Bruti, 2020).

The experiment conducted in this project will also yield practical implications, which offer some important insights into the mechanism of translating non-verbal elements in films. The findings from the experiment have further highlighted the potential power of subtitlers to direct and influence target viewers' cognitive attention towards the audiovisual content (Lautenbacher, 2012, 2015). In this light, audiovisual translators (or subtitlers) can be more aware of the critical non-verbal elements in audiovisual products, so as to achieve a more comprehensible and effective work of translation (e.g., explicating the visual kinesics in the target subtitles; preserving the verbal referential information relevant to the visual objects).

Moreover, as for translation trainees, when dealing with multiple semiotic resources in audiovisual products, they can use the two theoretical frameworks to justify their multimodal representation in the target subtitles.

## 6.3 Limitations

This thesis is subject to several limitations that should be considered when interpreting the results. Due to limited resources, the representativeness of the findings in the thesis may be challenged by the subjectivity of multimodal analysis, the relatively small size of the film corpus, the comparability of subtitles annotated in the corpus, and the ecological validity of the experiment.

First of all, it is crucial to acknowledge the inherent subjectivity of multimodal analysis. Different observers approaching a text-image interaction from different perspectives may arrive at varying interpretations. Although two coders were involved in annotating text-image relations and translation shifts within the multimodal corpus, the inter-coder agreement was not exceptionally high (80.97%). To mitigate the extent of subjectivity in the ultimate analysis, it is advisable to develop a comprehensive annotation guideline (e.g., Pastra, 2008). As for the identification of translation shifts in text-image relations, some image-induced shifts may also be interpreted as language-induced shifts. For example, an instance of *detachment of concurrence* for the transitivity component of *participants* (see Figure 5.2) may not be solely attributed to the subtitler's consideration of the co-occurring visual object. It was possible that the pronouns were omitted not because they were redundant given the characters' visibility on screen, but rather because the target sentences would sound more natural in Chinese without those pronouns. In this sense, the case of *detachment of concurrence* could be identified as a *preferential shift* instead (see definition in Section 3.4.3). The potential for multiple interpretations thus adds complexity to the application of the proposed frameworks.

Moreover, text-image relations may not be exclusively confined within clear categorical boundaries; rather, the categorization might manifest a continuum or gradience, which could be further described through the lens of prototype theory (Halverson, 2000). For instance, consider a film scene that shows a Serbian husky with distinctive pointed ears and a black and white coat. If the subtitle mentions "Husky", it can be said that the text establishes a concurrence relation with this visual entity because the depiction aligns with the prototypical features of a husky. On the other hand, if the text simply states "a dog", the text-image relation can be seen as falling somewhere between a *concurrence* and a *condensation* relation. This is because (a) the textual dog directly corresponds to the visual husky, indicating a degree of *concurrence*, but (b) a husky may not be the prototype of dogs, making the text less specific than the image and thus exhibiting a case of *condensation*.

Regarding the corpus-based study, it has to be admitted that the size of the multimodal corpus compiled for this thesis is arguably not large, with 30 annotated scenes sampled from 10 subtitled films. However, the corpus was deemed to be representative and sufficiently comprehensive to address the research questions in this project. The corpus consisted of 10 films spanning a 17-year period, from 2002 to 2019, thereby capturing subtitling patterns across different timeframes. In addition, the annotated films encompassed diverse genres, including action (e.g., *The Dark Knight*), adventure (e.g., *The Lord of the Rings: The Return of the King*), science fiction (e.g., *Avengers: Endgame*), and animation (e.g., *Coco*). The inclusion of various genres ensured the examination of distinct subtitling styles across different films. Moreover, the selected films boasted a substantial viewer base, with most of the subtitled films garnering millions of views, thus suggesting the potential extensive impact of the target subtitles. As Soffritti (2019) asserted, "even small and only partially annotated MMC [multimodal corpora] for AVT have proved to be useful as an empirical basis for very many contributions" (p. 345). The results from the corpus analysis also indicated that the corpus size was large enough "to examine what is typical, as well as what is rare in the language" (Baker, 2018, p. 169), for example, the disparities in frequency between different text-image relations. This corpus was thus arguably representative of various film genres and film subtitling practices over the last two decades, guaranteeing a thorough examination of recent subtitling strategies as multimodal representation (if any) in the AVT industry. Nevertheless, if more resources were available, augmenting the corpus size by incorporating a wider array of film genres could serve to enhance the representativeness of the corpus.

As for the bilingual subtitles analyzed in the corpus study, they were not fully comparable across the 10 selected films due to their different textual attributes, such as word counts and subtitle lines (see Table 4.1 in Section 4.1.2). Since one inclusion criterion for the corpus was based on the scene length, it was inevitable that different speech rates or styles in those scenes might lead to different textual features in the subtitles. This may potentially create an issue of over/under-representation. For example, a larger/smaller number of subtitles were included for the corpus analysis because of a faster/slower speech rate, thus skewing the frequency counts. While the amount of verbal content (i.e., total words of subtitles) of the annotated films was not directly comparable, at least the amount of visual content (i.e., scene duration) was proportionate. Since the focus of this study is on the relations between text and image, it is reasonable to ensure the comparability of either the text or the image between the annotated materials when it is impossible to have it both ways. In this study, it was more viable to guarantee the comparability of the amount of visual content and thus the difference of the subtitle features between the annotated films was assumed to be acceptable.

Ecological validity stands as a core concern in experimental studies, which often involve a tradeoff between control and generalizability. The video clips used as audiovisual stimuli in this eye-tracking experiment were very short, each lasting about two minutes, which deviated from a usually longer viewing in real-world scenarios. To maintain manageable control in the experiment, the author sacrificed "mundane realism" in the design (Mellinger & Hanson, 2022, p. 9). Additionally, the use of short video stimuli is not uncommon in AVT experiments (e.g., Black, 2022; Božović, 2023). Another potential concern of the experiment is the relatively small sample size. Due to a relatively high data attrition rate during data processing and limitations in financial and temporal resources for recruiting more participants, each group's sample size ranged from 25 to 30 for the final eye-tracking data analysis. It is noteworthy that this sample size is only slightly above the threshold recommended by Orero et al. (2018), who suggest that "[s]ample sizes of fewer than 25 per group are unlikely to yield statistical power" (p. 110).

### 6.4 Avenues for future research

Over the past decade, a notable turn in research focus within the field of Translation Studies has been observed, moving "from linguistic factors to extra-linguistic factors" (Huang & Liu, 2019, p. 50). In AVT studies, some particular emphases have been placed on the role of subtitles in building a multimodal cohesion in the audiovisual text (e.g., Chen, 2019; Lautenbacher, 2015; Taylor, 2016). This thesis attempts to synthesize research insights from corpus-based multimodal analysis and eye-tracking experiments. Considering the acknowledged research limitations, the thesis opens up avenues for future research to gain a more comprehensive understanding of the characteristics and impact of interlingual subtitling from a multimodal perspective.

Concerning the proposed theoretical framework of text-image relations, the analytical units for identifying text-image relations have been confined to three primary transitivity components (*process*, *participant*, and *circumstance*). However, a more granular analysis of transitivity categories also warrants consideration depending on the research questions. For example, a researcher may want to explore how film characters are constructed through different transitivity features. To illustrate, in a film, Character A is often encoded as Actor in the material *processes*, while Character B is frequently portrayed as Existent in the existential *processes*. In this way, A is represented as a more powerful character than B, and it will be interesting to explore whether the text-image interplay between the verbal and visual *processes* undergoes any shifts (e.g., visual Actor vs. verbal Existent). In such a case, the analysis of text-image relations between the verbal and visual *processes* can be further extended to the subtypes of *processes*, such as Material *processes*, Mental *processes*, and Relational *processes* (see Noverino et al., 2020). This more refined categorization holds the potential to offer fresh insights into the nuanced interactions between textual and visual elements in audiovisual products.

As the two frameworks proposed in the thesis consider only the verbal and visual modes, there is potential to extend the frameworks to encompass the auditory dimension, analyzing the soundtrack alongside text-image relations. This is worth conceptualization because a translation shift in a text-image relation may also be influenced by the speech rate of the portrayed characters. For instance, in the cases of *detachment of concurrence*, where the target subtitle omits the original textual reference to the co-occurring visual *participant*, it is possible that the texts are condensed because the original speech rate is high, which causes a time constraint and allows little room for a full representation of meaning conveyed in the source text. Thus, an additional aspect worth considering in the corpus analysis is the spatio-temporal aspects of the subtitles (Díaz Cintas & Remael, 2021), such as the maximum possible subtitle speed allowed by the industrial standard and the actual speed of a subtitle line. This approach could elucidate translation decisions not only in consideration of text-image interaction but the time and space constraints on subtitling.

Another aspect worth exploring in multimodal corpora of subtitled products is the (semi-)automatic annotation in light of the advancements in AI technology over the past few years. While the annotation process in this thesis is entirely manual, the adoption of more sophisticated visual recognition technology could significantly reduce the human resources required for the annotation of multimodal corpora. A recent attempt has been made by Baker and Collins (2023), who used the automatic tagger Google Cloud Vision to annotate a small multimodal corpus of British news articles. The recent launch of Gemini, a large multimodal model developed by Google, has further expanded the possibilities for AI-assisted multimodal analysis, as Gemini is reportedly capable of analyzing not only texts and static images but real-time dynamic video data (Google Gemini Team, 2023). More exploration of this kind could potentially streamline the annotation workflow and enhance the efficiency of corpus-based AVT studies.

While the corpus analysis in this thesis examines *what* and *how* text-image relations are altered through interlingual subtitling, it would also be interesting to explore *why* they are changed. In the discussion of the corpus findings (see Section 4.4), the author has postulated possible explanations for the translation shifts frequently observed, such as to enhance the multimodal cohesion in the audiovisual text (Remael & Reviers, 2019; Taylor, 2016), to adhere to the "the minimalist approach" (Taylor, 2012, p. 27) by condensing the texts and letting the viewers derive meaning from other semiotic resources, and to reinforce the characterization of the portrayed character (Messerli, 2019). Yet, these postulations require further validation, for example, through interviews with subtitlers to inquire about their translation decisions during the subtitling process. Exploring the reasons behind subtitling methods that consider non-verbal elements could lead to more interesting implications for future researchers and practitioners.

Regarding experimental reception studies on subtitling, there is still limited knowledge about the cognitive processing of subtitles with the concurrence of other visual and auditory information (Liao & Kruger, 2023). The experiment undertaken in this thesis investigates the impact of three prevalent types of translation shifts frequently observed in the multimodal corpus. To extend the scope of inquiry, further research could delve into the potential impact of other types of translation shifts on target viewers' reception and viewing experience. It is also advisable to conduct replication studies with a larger sample size to validate the findings of the current experiment. Additional empirical evidence will contribute to a more comprehensive understanding of subtitling as a form of multimodal representation and the subsequent multimodal viewing experience.

# REFERENCES

Abuczki, Á., & Ghazaleh, E. B. (2013). An overview of multimodal corpora, annotation tools and schemes. *Argumentum*, *9*(1), 86-98.

Aleksandrowicz, P. (2019). Subtitling song lyrics in films–pilot reception research. *Across Languages and Cultures 20*(2), 173-195.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*(3), 258-290.

Ayres, P., & Sweller, J. (2005). The Split-attention principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 135–146). Cambridge University Press.

Bączkowska, A. (2011). Some remarks on a multimodal approach to subtitles. *Linguistics Applied*, *4*, 47-65.

Bairstow, D. (2011). Audiovisual processing while watching subtitled films: A cognitive approach. In A. Serban, A. Matamala, & J. M. Lavaur (Eds.), *Audiovisual translation in close-up: Practical and theoretical approaches* (pp. 205–219). Peter Lang publishing.

Baker, M. (2019). Audiovisual translation and activism. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation* (pp. 453-467). Routledge.

Baker, P. (2018). Corpus methods in linguistics. In L. Litosseliti (Ed.), *Research methods in linguistics* (2nd ed., pp. 167-191). Bloomsbury.

Baker, P., & Collins, L. (2023). Creating and analysing a multimodal corpus of news texts with Google Cloud Vision's automatic image tagger. *Applied Corpus Linguistics*, *3*(1), 100043.

Baldry, A., & Thibault, P. J. (2006). *Multimodal transcription and text analysis: A multimedia toolkit and coursebook*. Equinox Pub.

Bateman, J. A. (2014a). Using multimodal corpora for empirical research. In C. Jewitt (Ed.), *The Routledge handbook of multimodal analysis* (pp. 238-252). Routledge.

Bateman, J. A. (2014b). *Text and image: A critical introduction to the visual/verbal divide*. Routledge.

Baumgarten, N. (2008). *Yeah, that's it!:* Verbal reference to visual information in film texts and film translations. *Meta*, *53*(1), 6-25.

Bisson, M., Van Heuven, W. J. B., Conklin, K., & Tunney R. J. (2014). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics*, *35*(2), 399-418.

Black, S. (2022). Could integrated subtitles benefit young viewers? Children's reception of standard and integrated subtitles: A mixed methods approach using eye tracking. *Perspectives, 30*(3), 1-17.

Blum-Kulka, S. (1986/2000). Shifts of Cohesion and Coherence in Translation. In L. Venuti (Ed.), *Translation studies reader* (pp. 298-313). Routledge.

Bogucki, Ł. (2020). *A relevance-theoretic approach to decision-making in subtitling*. Palgrave Macmillan.

Bonsignori, V. (2018). Using films and TV series for ESP teaching: A multimodal perspective. *System*, *77,* 58-69.

Bordwell, D. (2006). *The way Hollywood tells it: Story and style in modern movies*. University of California Press.

Bordwell, D., Thompson, K., & Smith, J. (2017). *Film art: An introduction* (7th ed.). McGraw-Hill Education.

Božović, P. (2019). How should culture be rendered in subtitling and dubbing?: A reception study on preferences and attitudes of end-users. *Babel*, *65*(1), 81-95.

Brookes, G., & McEnery, T. (2020). Corpus linguistics. In S. Adolphs & D. Knight (Eds.), *The Routledge handbook of English language and digital humanities* (pp. 378-404). Routledge.

Bruti, S. (2020). Corpus approaches and audiovisual translation. In Ł. Bogucki & M. Deckert (Eds.), *The Palgrave handbook of audiovisual translation and media accessibility* (pp. 381-396). Palgrave Macmillan.

Burczynska, P. (2017). *Investigating the multimodal construal and reception of irony in film translation: An experimental approach*. [Doctoral dissertation, The University of Manchester].

Caffrey, C. (2008). Viewer perception of visual nonverbal cues in subtitled TV Anime. *European Journal of English Studies*, *12*(2), 163-178.

Calzada Pérez, M. (2007). *Transitivity in translating: The interdependence of texture and context* (vol. 8). Peter Lang.

Caniato, M., Crocco, C., & Marzo S. (2015). Doctor or Dottore? How well do honorifics travel outside of Italy. *The Journal of Specialised Translation*, (23), 131-204.

Catford, J. C. (1965). *A linguistic theory of translation: An Essay in applied linguistics*. Oxford University Press.

Chaume, F. (2004). Film studies and translation studies: Two disciplines at stake in audiovisual translation. *Meta*, *49*(1), 12-24.

Chen, L. (2018). Subtitling culture: The reception of subtitles of fifth-generation Chinese films by British viewers. [Doctoral dissertation, University of Roehampton]. https://www.semanticscholar.org/paper/Subtitling-culture-%3A-the-reception-of-subtitles-of-Chen/f16d15a4d7c8e16b73588ee7c56348171a3e442b

Chen, X. (2022). Taboo language in non-professional subtitling on Bilibili.com: A corpus-based study. *Languages*, *7*(2), 138.

Chen, Y. (2019). *Translating film subtitles into Chinese: A multimodal study*. Springer.

Chen, Y., & Wang, W. (2019). Semiotic analysis of viewers' reception of Chinese subtitles: A relevance theory perspective. *Journal of Specialised Translation*, (32), 194-216.

Chen, Z. (2020). An audience reception study on multimodal relations in subtitled films: The case of *Coco*. [Master's thesis, Guangdong University of Foreign Studies].

Chesterman, A. (1998). Causes, translations, effects, *Target*, *10*(2): 201–230.

Chuang, Y. T. (2006). Studying subtitle translation from a multi-modal approach. *Babel*, *52*(4), 372-383.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Coolican, H. (2019). *Research methods and statistics in psychology* (7th ed.). Routledge.

Cui, Y., Liu, X., & Cheng, Y. (2023). Attention-consuming or attention-saving: an eye tracking study on punctuation in Chinese subtitling of English trailers. *Multilingua*, *42*(5), 739-763.

Cyrus, L. (2009). Old concepts, new ideas: Approaches to translation shifts. *MonTI*, *1*, 87-106.

d'Ydewalle, G., & Gielen, I. (1992). Attention allocation with overlapping sound, image, and text. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 415-427). Springer-Verlag.

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

d'Ydewalle, G., Praet, C., Verfaillie, K., & Rensbergen, J. V. (1991). Watching subtitled television: Automatic reading behavior. *Communication Research*, *18*(5), 650-666.

d'Ydewalle, G., Van Rensbergen, J., & Pollet, J. (1987). Reading a message when the same message is available auditorily in another language: The case of subtitling. In K. O'Regan & A. Lévy-Schoen (Eds.), *Eye movements: From physiology to cognition* (pp. 313-321). Elsevier Science Publishers.

da Silva, L. M. (1998). Character, Language and Translation: A linguistic study of a cinematic version of A Streetcar Named Desire. *Cadernos de Tradução*, *1*(3), 339-368.

Danan, M. (1992). Reversed subtitling and dual coding theory: New directions for foreign language instruction. *Language Learning*, *42*(4), 497-527.

Desilla, L. (2014). Reading between the lines, seeing beyond the images: An empirical study on the comprehension of implicit film dialogue meaning across cultures. *The Translator*, *20*(2), 194-214.

Díaz Cintas, J., & Remael, A. (2007). *Audiovisual translation: Subtitling*. St. Jerome Publishing.

Díaz Cintas, J., & Remael, A. (2021). *Subtitling: Concepts and practices*. Routledge.

Dicerto, S. (2018). *Multimodal pragmatics and translation: A new model for source text analysis*. Springer International Publishing.

Doherty, S., & Kruger, J. L. (2018). The development of eye tracking in empirical research on subtitling and captioning. In T. Dwyer, C. Perkins, S. Redmond, & J. Sitat (Eds.), *Seeing into screens: Eye tracking and the moving image* (pp. 46-64). Bloomsbury.

Edley, N., & Litosseliti, L. (2018). Critical perspectives on using interviews and focus groups. In L. Litosseliti (Ed.), *Research methods in linguistic* (2nd ed., pp. 195-225). Bloomsbury.

Fernández, A., Matamala, A., & Vilaró, A. (2014). The reception of subtitled colloquial language in Catalan: an eye-tracking exploratory study. *Vigo International Journal of Applied Linguistics*, (11), 63-80.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.

Fievez, I., Montero Perez, M., Cornillie, F., & Desmet, P. (2023). Promoting incidental vocabulary learning through watching a French Netflix series with glossed captions. *Computer Assisted Language Learning*, *36*(1-2), 26-51.

Fox, W. (2018). *Can integrated titles improve the viewing experience?: Investigating the impact of subtitling on the reception and enjoyment of film using eye tracking and questionnaire data*. Language Science Press.

Fresno, N., & Sepielak, K. (2022). Subtitling speed in Media Accessibility research: Some methodological considerations. *Perspectives*, *30*(3), 415–431.

Gambier, Y. (2006). Multimodality and audiovisual translation. In *MuTra 2006-audiovsiual scenarios: Conference proceedings* (pp. 1-8). MuTra.

Gambier, Y. (2018). Translation studies, audiovisual translation and reception. In E. Di Giovanni & Y. Gambier (Eds.), *Reception studies and audiovisual translation* (pp. 43-66). John Benjamins Publishing.

Georgakopoulou, P. (2009). Subtitling for the DVD Industry. In J. Díaz Cintas & G. Anderman (Eds.), *Audiovisual Translation: Language Transfer on Screen* (pp. 21-35). Palgrave Macmillan.

Gerber-Morón, O., & Szarkowska, A. (2018). Line breaks in subtitling: An eye tracking study on viewer preferences. *Journal of Eye Movement Research*, *11*(3), 1-22.

Gerber-Morón, O., Szarkowska, A., & Woll, B. (2018). The impact of text segmentation on subtitle reading. *Journal of Eye Movement Research, 11*(4), 1-18.

Germeys, F., & d'Ydewalle, G. (2007). The psychology of film: Perceiving beyond the cut. *Psychological Research*, *71*(4), 458-466.

Ghia, E. (2012). The impact of translation strategies on subtitle reading. In E. Perego (Ed.), *Eye tracking in audiovisual translation* (pp. 157-182). Aracne.

Ghoneam, N. (2015). The effect of subtitling on the enhancement of EFL learners' listening comprehension. *Arab World English Journal (AWEJ)*, *6*(4), 275-290.

Godfroid, A. (2020). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge.

Google Gemini Team. 2023. Gemini: A Family of Highly Capable Multimodal Models. https://storage.googleapis.com/deepmindmedia/gemini/gemini_1_report.pdf.

Gottlieb, H. (1992). Subtitling-a new university discipline. In C. Dollerup & A. Loddegaard (Eds.), *Teaching translation and interpreting: Training, talent and experience* (pp. 161-170). John Benjamins Publishing.

Gottlieb, H. (2001). Subtitling. In M. Baker (Ed.), *Routledge encyclopedia of translation studies* (2nd ed., pp. 317-323). Routledge.

Gottlieb, H. (2012). Subtitles - Readable dialogue?. In E. Perego (Ed.), *Eye tracking in audiovisual translation* (pp. 37-81). Aracne.

Gutt, E. (2000). *Translation and relevance: Cognition and context* (2nd ed.). St. Jerome Pub.

Halliday, M.A.K. (1994). *An introduction to functional grammar*. Arnold.

Halliday, M. A. K., & Matthiessen, C. M. (2014). *Halliday's introduction to functional grammar* (4th ed.). Hodder Arnold.

Halverson, S. (2000). Prototype effects in the "translation" category. In A. Chesterman, N. Gallardo San Salvador, & Y. Gambier (Eds.), *Translation in context: Selected papers from the EST Congress, Granada 1998* (pp. 3-16). John Benjamins Publishing.

Han, C., & Wang, K. (2014). Subtitling swearwords in reality TV series from English into Chinese: A corpus-based study of *The Family*. *The International Journal for Translation & Interpreting Research*, *6*(2), 1-17.

Hart, C. (2020). Multimodal discourse analysis. In C. Hart (Ed.), *Researching discourse: A student guide* (pp. 143-179). Routledge.

Hefer, E. (2013a). Television subtitles and literacy: Where do we go from here?. *Journal of Multilingual and Multicultural Development, 34*(7), 636-652.

Hefer, E. (2013b). Reading second language subtitles: A case study of Afrikaans viewers reading in Afrikaans and English. *Perspectives, 21*(1), 22-41.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Hu, K., O'Brien, S., & Kenny, D. (2020). A reception study of machine translated subtitles for MOOCs. *Perspectives, 28*(4), 521-538.

Huang, C. T. J. (1984). On the distribution and reference of empty pronouns. *Linguistic Inquiry, 15*(4), 531-574.

Huang, J., & Wang, J. (2023). Post-editing machine translated subtitles: Examining the effects of non-verbal input on student translators' effort. *Perspectives*, *31*(4), 1-21.

Huang, Q., & Liu, F. (2019). International translation studies from 2014 to 2018: A bibliometric analysis and its implications. *Translation Review, 105*(1), 34-57.

Ivir, V. (1981). Formal correspondence vs. translation equivalence revisited. *Poetics today*, *2*(4), 51-59.

Jennings, M. A., & Cribbie, R. A. (2016). Comparing pre-post change across groups: Guidelines for choosing between difference scores, ANCOVA, and residual change scores. *Journal of Data Science, 14*(2), 205-229.

Jewitt, C., Bezemer, J., & O'Halloran, K. L. (2016). *Introducing multimodality*. Routledge.

Ji, M., & Oakes, M. P. (2012). A corpus study of early English translations of Cao Xueqin's Hongloumeng. In M. P. Oakes & M. Ji (Eds.), *Quantitative methods in corpus-based translation studies* (pp. 177-208). John Benjamins Publishing.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *13*(4), 351-371.

Karim, K., & Nassaji, H. (2020). The revision and transfer effects of direct and indirect comprehensive corrective feedback on ESL students' writing. *Language Teaching Research, 24*(4), 519-539.

Kinnunen, T., & Koskinen, K. (2010). *Translators' agency*. Tampere University Press.

Kipp, M. (2001). Anvil: A generic annotation tool for multimodal dialogue. In *The 7th European Conference on Speech Communication and Technology* (pp. 1367-1370).

Klaudy, K. (2009). Explicitation. In M. Baker & G. Saldanha (Eds.), *Routledge encyclopedia of translation studies* (pp. 104-108). Routledge.

Kress, G., & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Edward Arnold.

Kress, G., & Van Leeuwen, T. (2021). *Reading images: The grammar of visual design* (3rd ed.). Routledge.

Kruger, J. L., & Liao, S. (2022). Establishing a theoretical framework for AVT research: The importance of cognitive models. *Translation Spaces*, *11*(1), 12-37.

Kruger, J. L., & Steyn, F. (2014). Subtitles and eye tracking: Reading and performance. *Reading Research Quarterly*, *49*(1), 105-120.

Kruger, J. L., Hefer, E., & Matthew, G. (2014). Attention distribution and cognitive load in a subtitled academic lecture: L1 vs. L2. *Journal of Eye Movement Research*, *7*(5), 1-15.

Kruger, J. L., Wisniewska, N., & Liao, S. (2022). Why subtitle speed matters: Evidence from word skipping and rereading. *Applied Psycholinguistics*, *43*(1), 211-236.

Kuo, A. S. Y. (2015). Professional realities of the subtitling industry: The subtitlers' perspective. In R. Baños Piñero & J. Díaz Cintas (Eds.), *Audiovisual translation in a global context: Mapping an ever-changing landscape* (pp. 163-191). Palgrave Macmillan.

Kuscu-Ozbudak, S. (2022). The role of subtitling on Netflix: An audience study, *Perspectives*, *30*(3), 1-15.

Künzli, A. (2021). From inconspicuousness to flow-the CIA model of subtitle quality. *Perspectives*, *29*(3), 326-338.

Lautenbacher, O. P. (2012). From still pictures to moving pictures: Eye tracking text and image. In E. Perego (Ed.), *Eye tracking in audiovisual translation* (pp. 135-155). Aracne.

Lautenbacher, O. P. (2015). Reading cohesive structures in subtitled films: a pilot study. In E. Perego & S. Bruti (Eds.), *Subtitling today: Shapes and their meanings* (pp. 33-56). Cambridge Scholars Publishing.

Lee, M., Roskos, B., & Ewoldsen, D. R. (2013). The impact of subtitles on comprehension of narrative film. *Media Psychology*, *16*(4), 412-440.

Lertola, J., & Mariotti, C. (2017). Reverse dubbing and subtitling: Raising pragmatic awareness in Italian English as a second language (ESL) learners. *The Journal of Specialised Translation*, (28), 103-121.

Liao, M. H. (2011). Interaction in the genre of popular science: Writer, translator and reader. *The Translator*, *17*(2), 349-368.

Liao, S., & Kruger, J. L. (2023). Cognitive processing of subtitles: Charting the future by mapping the past. In A. Ferreira & J. W. Schwieter (Eds.), *The Routledge handbook of translation, interpreting and bilingualism* (pp. 161-176). Routledge.

Liao, S., Kruger, J. L., & Doherty, S. (2020). The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension. *The Journal of Specialised Translation*, (33), 70-98.

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.

Liao, S., Yu, L., Reichle, E. D., & Kruger, J. L. (2021). Using eye movements to study the reading of subtitles in video. *Scientific Studies of Reading*, *25*(5), 417-435.

Lim, K. H., Benbasat, I., & Ward, L. M. (2000). The role of multimedia in changing first impression bias. *Information Systems Research*, *11*(2), 115-136.

Lindstromberg, S. (2016). Inferential statistics in Language Teaching Research: A review and ways forward. *Language Teaching Research*, *20*(6), 741-768.

Liu, Y., Zheng, B., & Zhou, H. (2019). Measuring the difficulty of text translation: The combination of text-focused and translator-oriented approaches. *Target*, *31*(1), 125-149.

Maoyan. (2022, January 7). *China Theatrical Market Report 2021*. https://piaofang.maoyan.com/feed/news/162480

Martin, J. R., Quiroz, B., & Wang, P. (2023). *Systemic functional grammar: A text-based description of English, Spanish and Chinese*. Cambridge University Press.

Martinec, R., & Salway, A. (2005). A system for image–text relations in new (and old) media. *Visual communication*, *4*(3), 337-371.

Marzban, A., & Zamanian, M. (2015). The impact of the subtitling task on vocabulary learning of Iranian EFL learners." *Journal of Applied Linguistics and Language Research*, *2*(1), 1-9.

Matthew, G. (2021). Do additional, visual elements in recorded lectures influence the processing of subtitles?. *Southern African Linguistics and Applied Language Studies*, *39*(1), 66-81.

Matthiessen, C. M. I. M. (2007). The multimodal page: A systemic functional exploration. In T. D. Royce & W. L. Bowcher (Eds.), *New directions in the analysis of multimodal discourse* (pp. 1-62). Lawrence Erlbaum Associates.

Mattsson, J. (2009). The subtitling of discourse particles: A corpus-based study of *well*, *you know*, *I mean*, and *like*, and their Swedish translations in ten American films. [Doctoral dissertation, University of Gothenburg].

Mellinger, C. D., & Hanson, T. A. (2017). *Quantitative research methods in translation and interpreting studies*. Routledge.

Mellinger, C. D., & Hanson, T. A. (2022). Considerations of ecological validity in cognitive translation and interpreting studies. *Translation, Cognition & Behavior*, *5*(1), 1-26.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*(1), 89-115.

Messerli, T. C. (2019). Subtitles and cinematic meaning-making: Interlingual subtitles as textual agents. *Multilingua*, *38*(5), 529-546.

Miko, F. 1970. La théorie de l'expression et la traduction. In J.S. Holmes (Ed.), *The nature of translation* (pp. 61-77). Mouton de Gruyter.

MOE. (2013, June 1). *Table of General Standard Chinese Characters*. Ministry of Education of the People's Republic of China. Retrieved August 1, 2023, from http://www.moe.gov.cn/jyb_sjzl/ziliao/A19/201306/t20130601_186002.html

Mubenga, K. S. (2010). Investigating norms in interlingual subtitling: A systemic functional perspective. *Perspectives*, *18*(4), 251-274.

Munday, J. (1998). A computer-assisted approach to the analysis of translation shifts. *Meta*, *43*(4), 542-556.

Munday, J. (2016). *Introducing translation studies: Theories and applications*. Routledge.

Muñoz, C. (2017). The role of age and proficiency in subtitle reading. An eye-tracking study. *System*, *67*, 77-86.

Nedergaard-Larsen, B. (1993). Culture-bound problems in subtitling. *Perspectives*, *1*(2), 207-240.

Negi, S., & Mitra, R. (2020). Fixation duration and the learning process: An eye tracking study with subtitled videos. *Journal of Eye Movement Research*, *13*(6). 1-15.

Neumann, S., Freiwald, J., & Heilmann, A. (2022). On the use of multiple methods in empirical translation studies: A combined corpus and experimental analysis of subject identifiability in English and German. In S. Granger & M. Lefer (Eds.), *Extending the scope of corpus-based translation studies* (pp. 98-129). Bloomsbury Academic.

Newcomb, H. (2004). Narrative and genre. In J. D. H. Downing, D. McQuail, P. Schlesinger, & E. Wartella (Eds.), *The SAGE Handbook of Media Studies* (pp. 413-428). Sage.

Noverino, R., Nababan, M. R., Santosa, R., & Djatmika. (2020). The realization of experiential meaning in Indonesian subtitles of *The Kingdom* (2007): Cases of transitivity system clausal constituents reduction. *Pertanika Journal of Social Sciences & Humanities, 28*(2), 1015-1033.

O'Halloran, K. L. (2015). The language of learning mathematics: A multimodal perspective. *The Journal of Mathematical Behavior*, *40,* 63-74.

Orero, P., Doherty, S., Kruger, J. L., Matamala, A., Pedersen, J., Perego, E., Esteva, S. R., Vilageliu, O. S., & Szarkowska, A. (2018). Conducting experimental research in audiovisual translation (AVT): A position paper. *The Journal of Specialised Translation*, (30), 105-126.

Orrego-Carmona, D. (2016). A reception study on non-professional subtitling: Do audiences notice any difference?. *Across Languages and Cultures*, *17*(2), 163-181.

Orrego-Carmona, D. (2019). Audiovisual translation and audience reception. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation* (pp. 367-382). Routledge.

Ortega, E. S. (2011). Subtitling and the relevance of non-verbal information in polyglot films. *New Voices in Translation Studies*, *7*(1), 19-34.

Pastra, K. (2008). COSMOROE: A cross-media relations framework for modelling multimedia dialectics. *Multimedia Systems*, *14*(5), 299-323.

Pavesi, M. (2019). Corpus-based audiovisual translation studies: Ample room for development. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation* (pp. 315-333). Routledge.

Pedersen, J. (2011). *Subtitling norms for television: An exploration focussing on extralinguistic cultural references*. John Benjamins Publishing.

Pedersen, J. (2018). From old tricks to Netflix: How local are interlingual subtitling norms for streamed television?. *Journal of Audiovisual Translation*, *1*(1), 81-100.

Pekkanen, H. (2007). The duet of the author and the translator: Looking at style through shifts in literary translation. *New Voices in Translation Studies*, *3*(1), 1-18.

Perego, E. (2003). Evidence of explicitation in subtitling: Towards a categorisation. *Across languages and cultures*, *4*(1), 63-88.

Perego, E., Del Missier, F., Porta, M., & Mosconi, M. (2010). The cognitive effectiveness of subtitle processing. *Media psychology*, *13*(3), 243-272.

Perego, E., Orrego-Carmona, D., & Bottiroli, S. (2016). An empirical take on the dubbing vs. subtitling debate: an eye movement study. *Lingue e Linguaggi*, *19,* 255-274.

Pérez-González, L. (2014a). Multimodality in translation and interpreting studies: Theoretical and methodological perspectives. In S. Bermann & C. Porter (Eds.), *A companion to translation studies* (pp. 119-131). John Wiley & Sons.

Pérez-González, L. (2014b). *Audiovisual translation theories, methods and issues*. Routledge.

Ragni, V. (2020). More than meets the eye: An eye-tracking study of the effects of translation on the processing and memorisation of reversed subtitles. *The Journal of Specialised Translation*, (33), 99-128.

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.

Ramos Pinto, S. (2018). Film, dialects and subtitles: An analytical framework for the study of non-standard varieties in subtitling. *The Translator*, *24*(1), 17-34.

Ramos Pinto, S., & Mubaraki, A. (2020). Multimodal corpus analysis of subtitling: The case of non-standard varieties. *Target*, *32*(3), 389-419.

Reiss, K. (1977). Text types, translation types and translation assessment. Trans. by A. Chesterman. In A. Chesterman (Ed.), *Readings in translation theory* (pp. 105-115). Finn Lectura.

Remael, A. (2003). Mainstream narrative film dialogue and subtitling: A case study of Mike Leigh's 'Secrets & Lies'(1996). *The Translator*, *9*(2), 225-247.

Remael, A., & Reviers, N. (2019). Multimodality and audiovisual translation: Cohesion in accessible films. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation* (pp. 260-280). Routledge.

Romero-Fresco, P. (2018). Eye tracking, subtitling and accessible filmmaking. In T. Dwyer, C. Perkins, S. Redmond, & J. Sitat (Eds.), *Seeing into screens: Eye tracking and the moving image* (pp. 235-258). Bloomsbury.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). John Wiley & Sons.

Saldanha, G., & O'Brien, S. (2014). *Research methodologies in translation studies*. Routledge.

Sasamoto, R., Doherty, S., & O'Hagan, M. (2021). The 'hookability' of multimodal impact captions: A mixed-methods exploratory study of Japanese TV viewers. *Translation, Cognition & Behavior*, *4*(2), 253-280.

Secară, A. (2011). R U ready 4 new subtitles? Investigating the potential of social translation practices and creative spellings. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, *10,* 153-171.

Silva, B. B., Orrego-Carmona, D., & Szarkowska, A. (2022). Using linear mixed models to analyze data from eye-tracking research on subtitling. *Translation Spaces, 11*(1), 60-88.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. (2003). *Reading concordances*. Longman.

Smith, T. J. (2013). Watching you watch movies: using eye tracking to inform film theory. In A. P. Shimamura (Ed.), *Psychocinematics: Exploring cognition at the movies* (pp. 165-191). Oxford University Press.

Soffritti, M. (2019). Multimodal corpora in audiovisual translation studies. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation* (pp. 334-349). Routledge.

Sotelo Dios, P. (2011). *Using a multimedia parallel corpus to investigate English-Galician subtitling*. Paper presented at Supporting Digital Humanities, Copenhagen. https://www.semanticscholar.org/paper/Using-a-Multimedia-Parallel-Corpus-to-Investigate-Dios/e8add030828e650e20396f330d33ee12ef1105dc

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational psychology review, 10*, 251-296.

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.

Szarkowska, A., Díaz Cintas, J., & Gerber-Morón, O. (2021). Quality is in the eye of the stakeholders: what do professional subtitlers and viewers think about subtitling?. *Universal Access in the Information Society*, *20*(4), 661-675.

Talaván, N., & Rodríguez-Arancón, P. (2014). The use of reverse subtitling as an online collaborative language learning tool. *The Interpreter and Translator Trainer, 8*(1), 84-101.

Taylor, C. (2003). Multimodal transcription in the analysis, translation and subtitling of Italian films. *The Translator*, *9*(2), 191-205.

Taylor, C. (2004). Multimodal text analysis and subtitling. In E. Ventola, C. Charles, & M. Kaltenbacher (Eds.), *Perspectives on multimodality* (pp. 153-172). John Benjamins Publishing.

Taylor, C. (2012). Multimodal texts. In E. Perego (Ed.), *Eye tracking in audiovisual translation* (pp. 13-35). Aracne.

Taylor, C. (2013). Multimodality and audiovisual translation. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of translation studies Volume 4* (pp. 98-104). John Benjamins Publishing.

Taylor, C. (2016). The multimodal approach in audiovisual translation. *Target, 28*(2), 222-236.

Thibault, P. (2000). The multimodal transcription of a television advertisement: Theory and practice. In A. Baldry (Ed.), *Multimodality and multimediality in the distance learning age* (pp. 311-385). Palladino Editore.

Thompson, G. (2014). *Introducing functional grammar* (3rd ed.). Routledge.

Tirkkonen-Condit, S. (2005). The monitor model revisited: Evidence from process research. *Meta*, *50*(2), 405-414.

Tirkkonen-Condit, S., & Mäkisalo, J. (2007). Cohesion in subtitles: A corpus-based study. *Across Languages and Cultures, 8*(2): 221–230.

Tortoriello, A. (2011). Semiotic cohesion in subtitling: The case of explicitation. In A. Serban, A. Matamala, & J. M. Lavaur (Eds.), *Audiovisual translation in close-up: Practical and theoretical approaches* (pp. 61-74). Peter Lang publishing.

Toury, G. (1980). *In search of a theory of translation*. The Porter Institute for Poetics and Semiotics, Tel Aviv University.

Tuominen, T. (2011). Accidental reading? Some observations on the reception of subtitled films. In A. Serban, A. Matamala, & J. M. Lavaur (Eds.), *Audiovisual translation in close-up: Practical and theoretical approaches* (pp. 189-204). Peter Lang publishing.

Tuominen, T., Jiménez Hurtado, C., & Ketola, A. (2018). Why methods matter: Approaching multimodality in translation research. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *17*, 1–21.

Tymoczko, M. (2007). *Enlarging Translation, Empowering Translators*. St. Jerome Publishing.

Unsworth, L. (2006). Towards a metalanguage for multiliteracies education: Describing the meaning-making resources of language-image interaction. *English teaching: Practice and critique*, *5*(1), 55-76.

Unsworth, L. (2007). Image/text relations and intersemiosis: Towards multimodal text description for multiliteracies education. In L. Barbara & T. Berber Sardinha (Eds.), *Proceedings of the 33rd IFSC: International Systemic Functional Congress* (pp. 1165-1205). Pontificia Universidade Catolica de Sao Paulo.

Valeontis, K., & Mantzari, E. (2006). The linguistic dimension of terminology: Principles and methods of term formation. In *1st Athens International Conference on Translation and Interpretation Translation: Between Art and Social Science* (pp. 13-14).

Van Leuven-zwart, Kitty M. (1989). Translation and Original: Similarities and Dissimilarities, I. *Target*, *1*(2), 151-181.

Van Leuven-zwart, Kitty M. (1990). Translation and Original: Similarities and Dissimilarities, II. *Target*, *2*(1), 69-95.

Van Lommel, S., Laenen, A., & d'Ydewalle, G. (2003). Foreign-grammar acquisition while watching subtitled television programs. *Psychological Reports, 76*, 243-258.

Vinay, J.-P., & Darbelnet, J. (1958). *Stylistique Comparée du Français et de l'Anglais: Méthode de Traduction*. Didier.

Vinay, J.-P., & Darbelnet, J. (1995). *Comparative Stylistics of French and English: A Methodology for Translation*. trans and eds by Juan Sager and Marie-Jo Hamel. John Benjamins.

Vitucci, F. (2017). The semiotic cohesion of audiovisual texts. Types of intersemiotic explicitations in the English subtitles of Japanese full-length films. *Studia Translatorica, 8*, 83-104.

Wallis, S. (2021). *Statistics in corpus linguistics research: A new approach*. Routledge.

Wang, A., & Pellicer-Sánchez, A. (2022). Incidental vocabulary learning from bilingual subtitled viewing: An eye-tracking study. *Language Learning*, *72*(3), 765-805.

Wang, L., Tu, Z., Zhang, X., Li, H., Way, A., & Liu, Q. (2017). A novel approach to dropped pronoun translation. *Machine Translation*, *31*, 65–87.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation* (pp. 1556-1559). European Language Resources Association.

Wu, Z. (2017). The making and unmaking of non-professional subtitling communities in China: a mixed-method study. In L. Orrego-Carmona, D. Orrego-Carmona, & Y. Lee (Eds.), *Non-professional subtitling* (pp. 115-143). Cambridge Scholars.

Wu, Z., & Chen, Z. (2021). A systematic review of experimental research in audiovisual translation 1992–2020. *Translation, Cognition & Behavior, 4*(2), 281-304.

Wu, Z., & Chen, Z. (2022). Towards a corpus-driven approach to audiovisual translation (AVT) reception: A case study of YouTube viewer comments. *The Journal of Specialised Translation, 38*, 128-154.

Zabalbeascoa, P. (2008). The nature of the audiovisual text and its parameters. In J. Díaz Cintas (Ed.), *The didactics of audiovisual translation* (pp. 21-37). John Benjamins.

Zhang, M., & Feng, D. W. (2020). Introduction: Multimodal approaches to Chinese-English translation and interpreting. In M. Zhang & D. Feng (Eds.), *Multimodal approaches to Chinese-English translation and interpreting* (pp. 1-15). Routledge.

Zhang, M., & Pan, L. (2009). Introducing a Chinese Perspective on Translation Shifts: A Comparative Study of Shift Models by Loh and Vinay & Darbelnet. *The Translator*, *15*(2), 351-374.

Zheng, B., & Xie, M. (2018). The effect of explanatory captions on the reception of foreign audiovisual products: A study drawing on eye-tracking data and retrospective interviews. *Translation, Cognition & Behavior*, *1*(1), 119-146.

# APPENDICES

## Appendix 1. List of top 100 popular and highly rated English films between 2000 and 2020

| Rank | Film title | Year | North America grossing score | China grossing score | IMDb rating score | Douban rating score | Mean score |
|---|---|---|---|---|---|---|---|
| 1 | *The Dark Knight* | 2008 | 239 | 0 | 247 | 224 | 177.5 |
| 2 | *The Lord of the Rings: The Return of the King* | 2003 | 205 | 0 | 244 | 220 | 167.3 |
| 3 | *Avengers: Endgame* | 2019 | 249 | 244 | 175 | 0 | 167.0 |
| 4 | *Avengers: Infinity War* | 2018 | 246 | 231 | 188 | 0 | 166.3 |
| 5 | *Inception* | 2010 | 155 | 26 | 238 | 242 | 165.3 |
| 6 | *Avatar* | 2010 | 248 | 221 | 0 | 174 | 160.8 |
| 7 | *Coco* | 2017 | 53 | 185 | 179 | 221 | 159.5 |
| 8 | *Zootopia* | 2016 | 192 | 212 | 0 | 232 | 159.0 |
| 9 | *The Lord of the Rings: The Two Towers* | 2002 | 193 | 0 | 236 | 203 | 158.0 |
| 10 | *Interstellar* | 2014 | 11 | 146 | 223 | 239 | 154.8 |
| 11 | *The Lord of the Rings: The Fellowship of the Ring* | 2001 | 172 | 0 | 241 | 199 | 153.0 |
| 12 | *The Dark Knight Rises* | 2012 | 229 | 0 | 178 | 137 | 136.0 |
| 13 | *WALL·E* | 2008 | 84 | 0 | 189 | 236 | 127.3 |
| 14 | *Up* | 2009 | 157 | 0 | 127 | 217 | 125.3 |
| 15 | *Toy Story 3* | 2010 | 220 | 0 | 129 | 110 | 114.8 |
| 16 | *Jurassic World* | 2015 | 244 | 204 | 0 | 0 | 112.0 |
| 17 | *Jurassic World: Fallen Kingdom* | 2018 | 221 | 219 | 0 | 0 | 110.0 |
| 18 | *Avengers: Age of Ultron* | 2015 | 230 | 209 | 0 | 0 | 109.8 |
| 19 | *Furious 7* | 2015 | 196 | 232 | 0 | 0 | 107.0 |
| 20 | *The Pianist* | 2002 | | | 217 | 200 | 104.3 |
| 21 | *Aquaman* | 2018 | 187 | 227 | 0 | 0 | 103.5 |
| 22 | *Spider-Man: Far From Home* | 2019 | 210 | 203 | 0 | 0 | 103.3 |
| 23 | *Harry Potter and the Deathly Hallows: Part 2* | 2011 | 208 | 0 | 36 | 164 | 102.0 |
| 24 | *Captain America: Civil War* | 2016 | 217 | 187 | 0 | 0 | 101.0 |
| 25 | *Harry Potter and the Sorcerer's Stone* | 2001 | 174 | 0 | 0 | 218 | 98.0 |
| 26 | *Captain Marvel* | 2019 | 225 | 163 | 0 | 0 | 97.0 |
| 26 | *Star Wars: Episode VII - The Force Awakens* | 2016 | 250 | 138 | 0 | 0 | 97.0 |

**List (continued)**

| Rank | Film title | Year | North America grossing score | China grossing score | IMDb rating score | Douban rating score | Mean score |
|---|---|---|---|---|---|---|---|
| 28 | *The Lion King* | 2019 | 240 | 140 | 0 | 0 | 95.0 |
| 29 | *Frozen II* | 2019 | 233 | 143 | 0 | 0 | 94.0 |
| 29 | *The Prestige* | 2006 | 0 | 0 | 203 | 173 | 94.0 |
| 31 | *Joker* | 2019 | 188 | 0 | 181 | 0 | 92.3 |
| 32 | *Transformers: Dark of the Moon* | 2011 | 195 | 169 | 0 | 0 | 91.0 |
| 33 | *The Jungle Book* | 2016 | 200 | 159 | 0 | 0 | 89.8 |
| 34 | *Inside Out* | 2015 | 198 | 0 | 77 | 83 | 89.5 |
| 35 | *Iron Man 3* | 2013 | 218 | 124 | 0 | 0 | 85.5 |
| 36 | *Transformers: Age of Extinction* | 2014 | 115 | 226 | 0 | 0 | 85.3 |
| 37 | *Black Panther* | 2018 | 247 | 92 | 0 | 0 | 84.8 |
| 38 | *Pirates of the Caribbean: The Curse of the Black Pearl* | 2003 | 165 | 0 | 0 | 159 | 81.0 |
| 39 | *The Fate of the Furious* | 2017 | 86 | 235 | 0 | 0 | 80.3 |
| 40 | *Spider-Man: Homecoming* | 2017 | 186 | 130 | 0 | 0 | 79.0 |
| 41 | *Green Book* | 2019 | 0 | 14 | 119 | 180 | 78.3 |
| 42 | *Guardians of the Galaxy Vol. 2* | 2017 | 209 | 100 | 0 | 0 | 77.3 |
| 43 | *Despicable Me 3* | 2017 | 137 | 165 | 0 | 0 | 75.5 |
| 44 | *Beauty and the Beast* | 2017 | 236 | 64 | 0 | 0 | 75.0 |
| 45 | *Django Unchained* | 2012 | 0 | 0 | 193 | 106 | 74.8 |
| 46 | *The Avengers* | 2012 | 243 | 54 | 0 | 0 | 74.3 |
| 47 | *Wonder Woman* | 2017 | 219 | 75 | 0 | 0 | 73.5 |
| 48 | *A Beautiful Mind* | 2001 | 0 | 0 | 102 | 190 | 73.0 |
| 49 | *Thor: Ragnarok* | 2017 | 171 | 120 | 0 | 0 | 72.8 |
| 50 | *Monsters, Inc.* | 2001 | 150 | 0 | 21 | 115 | 71.5 |
| 50 | *Ready Player One* | 2018 | 0 | 202 | 0 | 84 | 71.5 |
| 52 | *Whiplash* | 2014 | 0 | 0 | 208 | 76 | 71.0 |
| 53 | *Venom* | 2018 | 59 | 224 | 0 | 0 | 70.8 |
| 54 | *Finding Nemo* | 2003 | 207 | 0 | 74 | 0 | 70.3 |
| 55 | *Memento* | 2000 | 0 | 0 | 196 | 82 | 69.5 |
| 56 | *Mission: Impossible - Fallout* | 2018 | 81 | 188 | 0 | 0 | 67.3 |
| 57 | *Hachi: A Dog's Tale* | 2009 | 0 | 0 | 27 | 241 | 67.0 |
| 58 | *Shutter Island* | 2010 | 0 | 0 | 96 | 171 | 66.8 |
| 59 | *Life of Pi* | 2012 | 0 | 55 | 0 | 211 | 66.5 |
| 60 | *Batman v Superman: Dawn of Justice* | 2016 | 182 | 78 | 0 | 0 | 65.0 |

**List (continued)**

| Rank | Film title | Year | North America grossing score | China grossing score | IMDb rating score | Douban rating score | Mean score |
|---|---|---|---|---|---|---|---|
| 61 | *Catch Me If You Can* | 2002 | 0 | 0 | 54 | 202 | 64.0 |
| 62 | *Harry Potter and the Chamber of Secrets* | 2002 | 136 | 0 | 0 | 119 | 63.8 |
| 62 | *Harry Potter and the Prisoner of Azkaban* | 2004 | 119 | 0 | 0 | 136 | 63.8 |
| 64 | *Guardians of the Galaxy* | 2014 | 184 | 68 | 0 | 0 | 63.0 |
| 64 | *The Hobbit: The Battle of the Five Armies* | 2015 | 125 | 127 | 0 | 0 | 63.0 |
| 66 | *Rogue One: A Star Wars Story* | 2017 | 238 | 13 | 0 | 0 | 62.8 |
| 67 | *Big Hero 6* | 2015 | 82 | 41 | 0 | 127 | 62.5 |
| 68 | *Captain America: The Winter Soldier* | 2014 | 131 | 112 | 0 | 0 | 60.8 |
| 69 | *Star Wars: Episode VIII - The Last Jedi* | 2017 | 242 | 0 | 0 | 0 | 60.5 |
| 70 | *Incredibles 2* | 2018 | 241 | 0 | 0 | 0 | 60.3 |
| 71 | *Jumanji: welcome to the Jungle* | 2018 | 214 | 25 | 0 | 0 | 59.8 |
| 72 | *Despicable Me* | 2010 | 123 | 0 | 0 | 114 | 59.3 |
| 72 | *Star Wars: Episode IX - The Rise of Skywalker* | 2019 | 237 | 0 | 0 | 0 | 59.3 |
| 74 | *How to Train Your Dragon* | 2010 | 73 | 0 | 40 | 122 | 58.8 |
| 74 | *Logan* | 2017 | 88 | 117 | 30 | 0 | 58.8 |
| 76 | *Finding Dory* | 2016 | 234 | 0 | 0 | 0 | 58.5 |
| 77 | *The Grand Budapest Hotel* | 2014 | 0 | 0 | 62 | 167 | 57.3 |
| 77 | *The Pursuit of Happyness* | 2006 | 0 | 0 | 0 | 229 | 57.3 |
| 79 | *Shrek 2* | 2004 | 228 | 0 | 0 | 0 | 57.0 |
| 80 | *Flipped* | 2010 |  | 0 | 0 | 226 | 56.5 |
| 80 | *Toy Story 4* | 2019 | 226 | 0 | 0 | 0 | 56.5 |
| 82 | *Harry Potter and the Goblet of Fire* | 2005 | 149 | 0 | 0 | 75 | 56.0 |
| 82 | *The Hunger Games: Catching Fire* | 2013 | 224 | 0 | 0 | 0 | 56.0 |
| 84 | *Pirates of the Caribbean: Dead Man's Chest* | 2006 | 223 | 0 | 0 | 0 | 55.8 |
| 85 | *Doctor Strange* | 2016 | 96 | 122 | 0 | 0 | 54.5 |
| 86 | *Gladiator* | 2000 | 9 | 0 | 207 | 0 | 54.0 |
| 86 | *The Hunger Games* | 2012 | 216 | 0 | 0 | 0 | 54.0 |
| 88 | *Spider-Man* | 2002 | 215 | 0 | 0 | 0 | 53.8 |

**List (continued)**

| Rank | Film title | Year | North America grossing score | China grossing score | IMDb rating score | Douban rating score | Mean score |
|---|---|---|---|---|---|---|---|
| 89 | *X-Men: Days of Future Past* | 2014 | 99 | 115 | 0 | 0 | 53.5 |
| 90 | *Transformers: The Last Knight* | 2017 | 0 | 213 | 0 | 0 | 53.3 |
| 91 | *Transformers: Revenge of the Fallen* | 2009 | 212 | 0 | 0 | 0 | 53.0 |
| 92 | *Frozen* | 2013 | 211 | 0 | 0 | 0 | 52.8 |
| 93 | *Warcraft* | 2016 | 0 | 210 | 0 | 0 | 52.5 |
| 94 | *Fast & Furious Presents: Hobbs & Shaw* | 2019 | 0 | 207 | 0 | 0 | 51.8 |
| 95 | *Ant-Man and the Wasp* | 2018 | 67 | 139 | 0 | 0 | 51.5 |
| 95 | *Star Wars: Episode III - Revenge of the Sith* | 2005 | 206 | 0 | 0 | 0 | 51.5 |
| 97 | *The Departed* | 2006 | 0 | 0 | 205 | 0 | 51.3 |
| 98 | *Spider-Man 2* | 2004 | 204 | 0 | 0 | 0 | 51.0 |
| 99 | *The Secret Life of Pets* | 2016 | 202 | 0 | 0 | 0 | 50.5 |
| 100 | *Despicable Me 2* | 2013 | 201 | 0 | 0 | 0 | 50.3 |
| 100 | *Inglourious Basterds* | 2009 | 0 | 0 | 166 | 35 | 50.3 |
| 100 | *Spider-Man: Into the Spider-Verse* | 2018 | 16 | 0 | 185 | 0 | 50.3 |

## Appendix 2. Subtitles designed for treatment conditions in the main experiment

Clip 1: Dialogues in the scene selected from 00:08:18 to 00:11:58 in *Meet the Parents*

| ST | TT for CG | TT for EXG | Treatment instance No. |
|---|---|---|---|
| You make it sound like they're really hard to please. | 你爸好像很难侍候似的 | 你爸好像很难侍候似的 | |
| No, not at all! He's the sweetest man in the whole world. Just relax! | 不是的　他是世界上最可爱的男人 | 不是的　他是世界上最可爱的男人 | |
| He's gonna love you. I promise. | 他会喜欢你的　我保证 | 他会喜欢你的　我保证 | |
| As much as he loves Dr. Bob? | 像喜欢你妹夫那样吗 | 像喜欢你妹夫那样吗 | |
| Take it easy on the sarcasm. Humor is entirely wasted on my parents. | 别太嘴贫　我爸妈不欣赏幽默 | 别太嘴贫　我爸妈不欣赏幽默 | |
| What, are they Amish? | 什么　他们这么保守的吗 | 什么　他们这么保守的吗 | |
| | 别开玩笑啦 | 别开玩笑啦 | |
| OK, no jokes. | 好　不开玩笑 | 好　不开玩笑 | |
| What are you doing? | 你在干嘛呢？<br>[What are you doing?] | 你干嘛？抽烟？<br>[What you doing? Smoking?] | 1 |
| What? | 什么 | 什么 | |
| I told you my dad sees smoking as a sign of weakness. | 我爸认为这是懦弱的表现 | 我爸认为这是懦弱的表现 | |
| OK, all right, I'll leave 'em in the car. | 好　我放在车上 | 好　我放在车上 | |
| No, no, no, no, he'll check there. | 不行　他会检查车子的 | 不行　他会检查车子的 | |
| – Oh, gosh. – What··· | 什么啊<br>[What?] | 扔了吗<br>[Throw away?] | 2 |
| Yeah, the roof is probably a better idea. | 行　屋顶上也许更好<br>[Yeah, on the roof probably better.] | 行　那就扔到屋顶吧<br>[Yeah, then throw onto the roof.] | 3 |
| Hey. Oh, and··· | 还有⋯ | 还有⋯ | |
| ···we're not living together. | 我俩没有同居 | 我俩没有同居 | |
| I thought you said you told him. | 我以为你告诉他们了 | 我以为你告诉他们了 | |
| Well··· | 这个. . . | 这个. . . | |
| – Hi, Daddy! Hi! – Sweet pea! | 嗨　爸爸 | 嗨　爸爸 | |
| – I missed you so much, Pamcake. | 我好想你啊　女儿 | 我好想你啊　女儿 | |
| I missed you too, Flapjack. | 我也好想念你　爸爸<br>[I missed you too, dad.] | 我也好想你　亲一个<br>[I missed you too. Kiss.] | 4 |
| Oh, boy, oh, boy, oh, boy, oh, boy, oh, boy, oh, boy! | 天呐　天呐　天呐<br>[Oh, boy, oh, boy, oh, boy.] | 抱一抱　抱一抱<br>[Hug, Hug.] | 5, 6 |
| Short stack, short stack, coming up. | 一叠一叠　大小手<br>[Short stack, short stack, big and small hands.] | 一叠一叠　叠起手<br>[Short stack, short stack, stack the hands.] | 7 |
| Where's my «wittle» girl? | 我的乖女呢 | 我的乖女呢 | |
| Mommy! Mom! | 嗨　妈妈 | 嗨　妈妈 | |
| You look so beautiful. | 你好美啊 | 你好美啊 | |
| So do you. Look at you. | 你也是　你看看你 | 你也是　你看看你 | |
| Oh, I'm sorry. Mom, Dad, this is Greg. | 对不起　爸妈　这是阿基 | 对不起　爸妈　这是阿基 | |

| | | | |
|---|---|---|---|
| -Hi, Greg. I’m Pam’s father, Jack Byrnes. Good meeting you. -Great to finally meet you. | -我是白梅的爸爸 白杰克 -幸会 | -我是白梅的爸爸 白杰克 -幸会 | |
| And I’m Dina. Welcome to Oyster Bay. | -我是迪娜 欢迎来到蚝湾 -谢谢 | -我是迪娜 欢迎来到蚝湾 -谢谢 | |
| And anyway, Greg, meanwhile, anything you need, just ask. | 你需要什么说一声就行了 | 你需要什么说一声就行了 | |
| That’s right. Mi casa es su casa. | 对　把我家当作你家 | 对　把我家当作你家 | |
| Oh, thanks, Jack. You too. | 谢谢　杰克 | 谢谢　杰克 | |
| Yeah. | 好　快进屋吧 | 好　快进屋吧 | |
| Hey, Mom, the house looks great. | 这房子好棒 | 这房子好棒 | |
| We like it. | 我们很喜欢 | 我们很喜欢 | |
| Beautiful. | 真漂亮 | 真漂亮 | |
| Oh, I’ll get it, honey. | 宝贝　让我来 | 宝贝　让我来 | |
| Oh, thanks, Mom. | 谢谢妈妈 | 谢谢妈妈 | |
| There he is. There’s our little guy. | 它来了 | 它来了 | |
| Jinxy, come here, boy. | 是黑仔呀 | 是黑仔呀 | |
| Come here, baby. Come to Daddy, Jinxy. Come on. Come on. Jinxy. | 黑仔过来　到爸爸这儿 | 黑仔过来　到爸爸这儿 | |
| Come here. Come to Daddy. | 过来　黑仔 | 过来　黑仔 | |
| Jinxy. | 黑仔呀 | 黑仔呀 | |
| Come on. | 来　到爸爸这儿<br>[Come on. To daddy here.] | 来　跳到爸爸这<br>[Come on. Jump to daddy here.] | 8 |
| Taught him that in one week. | 我花了一星期时间教它这个<br>[I spent a week teaching it this.] | 我教了一星期　它就会跳上来<br>[I taught for a week and it can jump up.] | 9 |
| This is Pam’s cat, Jinxy. | 这是白梅的猫　黑仔 | 这是白梅的猫　黑仔 | |
| Jinxy, say hello to Greg. | 黑仔　跟阿基打个招呼 | 黑仔　跟阿基打个招呼 | |
| – Wave to Greg. – Hello, Jinx. | 你好呀　黑仔 | 你好呀　黑仔 | |
| Attaboy. That took me another week. | 我又花了一个星期教它这个<br>[I spent another week teaching it this.] | 我又教了一星期　它就会挥手<br>[I taught for another week and it can wave.] | 10 |
| Oh, my gosh. | 我的天呐 | 我的天呐 | |
| Pam, I didn’t know you had a cat. | 我不知道你养猫呢 | 我不知道你养猫呢 | |
| I left him here when I moved to Chicago. | 我搬到芝加哥时留下的 | 我搬到芝加哥时留下的 | |
| Your daddy’s found his new best friend. | 它是你爸的新挚友 | 它是你爸的新挚友 | |

Clip 2: Dialogues in the scene selected from 00:13:55 to 00:15:40 in *Meet the Parents*

| **ST** | **TT for CG** | **TT for EXG** | **Treatment instance No.** |
|---|---|---|---|
| You know, I wish you hadn’t told your parents I hate cats. | 真希望你没说我讨厌猫 | 真希望你没说我讨厌猫 | |
| But you do hate cats. | 你真的讨厌猫啊 | 你真的讨厌猫啊 | |
| You didn’t have to tell them right when we met. | 你不用一见面就告诉他们<br>[You didn’t have to tell them at the first meeting.] | 也不用一见面就说出来的<br>[Didn’t have to say at the first meeting.] | 1, 2 |
| I know. I’m sorry. It just kinda slipped out. | 对不起　我只是顺口<br>[Sorry, I just slipped out.] | 对不起　就只是顺口<br>[Sorry, just slipped out.] | 3 |
| Get your red-hot pu-pus. | 你们快来吃热腾腾的点心 | 快点来吃些热腾腾的点心 | 4 |

| | | | |
|---|---|---|---|
| | [You two, come and get red-hot pu-pus now.] | [Come and get some red-hot pu-pus now.] | |
| My goodness, what is that? | 我的天　那是什么 | 我的天　那是什么 | |
| Oh, that’s just a little something from me. | 那只是我的一点心意<br>[That’s just a little something from me.] | 那只是一点心意而已<br>[That’s just a little something.] | 5 |
| Go ahead. Open it up. | 来吧　拆开看看 | 来吧　拆开看看 | |
| Look, honey, Greg brought us a present. | 亲爱的　阿基给我们买了礼物<br>[Honey, Greg brought us a present.] | 亲爱的　阿基还买了一份礼物<br>[Honey, Greg brought a present.] | 6 |
| Oh, isn’t that nice? | 真漂亮 | 真漂亮 | |
| Oh, look at this. It's a flowerpot with the dirt in it. | 看　是个花盆　里面还有泥 | 看　是个花盆　里面还有泥 | |
| Actually, the real gift is what’s planted in the soil. | 真正的礼物在泥里面 | 真正的礼物在泥里面 | |
| The bulb of a Jerusalem tulip. | 耶路撒冷的郁金香球茎 | 耶路撒冷的郁金香球茎 | |
| Which I was told | 我听说是<br>[Which I was told is] | 据说这是<br>[Which was said to be] | 7 |
| is one of the rarest and most beautiful flowers in existence. | 最罕见和美丽的花卉之一 | 最罕见和美丽的花卉之一 | |
| Oh, right, right, the Jerusalem··· | 对　对　耶路撒冷··· | 对　对　耶路撒冷··· | |
| genus. Yes, yes. | 属于耶路撒冷郁金香花属 | 属于耶路撒冷郁金香花属 | |
| Anyway, yeah, the guy said with regular watering, | 不管是什么　听说定期浇水的话 | 不管是什么　听说定期浇水的话 | |
| it should bloom in about six months, so··· | 六个月内便会开花 | 六个月内便会开花 | |
| Oh, we’ll look forward to that, Greg. | 我们期待它开花　阿基<br>[We look forward to it in blossom, Greg.] | 期待它能够开花　阿基<br>[Looking forward to it in blossom, Greg.] | 8 |
| So, Greg, how’s your job? | 阿基　最近工作怎么样 | 阿基　最近工作怎么样 | |
| Good, Pam. Thanks for asking. | 很好　白梅　谢谢关心 | 很好　白梅　谢谢关心 | |
| I···I recently got transferred to triage. | 我最近调到分流部<br>[I recently got transferred to triage.] | 最近调到了分流部<br>[Recently got transferred to triage.] | 9 |
| Oh, is that better than a nurse? | 那比做护士好吗 | 那比做护士好吗 | |
| No, Mom, triage is a unit of the E.R. | 分流部是急症室的一个部门 | 分流部是急症室的一个部门 | |
| It’s where all the top nurses work. | 那里都是高级护士 | 那里都是高级护士 | |
| Well··· | 这个嘛··· | 这个嘛··· | |
| No, they do. | 是真的 | 是真的 | |
| Not many men in your profession, though, are there, Greg? | 这行业男人不多吧　阿基<br>[Not many men in your profession, Greg?] | 这个行业的男人不多吧<br>[Not many men in your profession?] | 10 |
| No, Jack, not traditionally. | 不多　杰克　传统上不多 | 不多　杰克　传统上不多 | |

Clip 3: Dialogues in the scene selected from 01:33:35 to 01:35:16 in *Meet the Parents*

| **ST** | **TT for CG** | **TT for EXG** | **Treatment instance No.** |
|---|---|---|---|
| Excuse me. | 不好意思 | 不好意思 | |
| | 麻烦让一下 | 麻烦让一下 | |
| OK, where’s the fire, huh? | 急什么 | 急什么 | |

| | | | |
|---|---|---|---|
| You’re gonna have to check that. | 先生　你的行李要托运 | 先生　你的行李要托运 | |
| I got it. | 我能塞进去 | 我能塞进去 | |
| No, I’m sorry. That bag won’t fit. | 不行　你的行李太大了 | 不行　你的行李太大了 | |
| No, I’m not··· hey. I’m not checking my bag, OK? | 我不要托运行李<br>[I’m not checking the luggage.] | 打死我都不托运<br>[Definitely, I’m not checking.] | 1 |
| There’s no need to raise your voice. | 你用不着那么大声音 | 你用不着那么大声音 | |
| I’m not raising my voice. | 我没有大声 | 我没有大声 | |
| This would be raising my voice to you, OK? | 这样才是大声 | 这样才是大声 | |
| I don’t want to check my bag. | 我的行李不托运<br>[My luggage is not to be checked.] | 打死我都不托运<br>[Definitely, I’m not checking.] | 2 |
| By the way, your airline, you suck at checking bags. | 你们公司不懂怎么托运<br>[Your company don’t know how to check.] | 你们公司根本不懂托运<br>[Your company don’t know how to check at all.] | 3 |
| Because I already did that once, and you lost it, | 我托运过　你们弄丢了 | 我托运过　你们弄丢了 | |
| and then I had everything screwed up very badly for me. OK? | 我什么都搞砸了 | 我什么都搞砸了 | |
| I can assure you that your bag will be placed safely below deck with the other luggage. | 我保证你的行李会安放在行李舱内 | 我保证你的行李会安放在行李舱内 | |
| Oh, yeah? How do you know my bag will be safe below with the other luggage? | 你怎么知道行李在舱内会安全 | 你怎么知道行李在舱内会安全 | |
| Are you physically gonna take my bag beneath the plane? | 你会亲自拎去那里吗 | 你会亲自拎去那里吗 | |
| Gonna go with the guys with the earmuffs and put it in there? | 你现在会下机去放行李吗 | 你现在会下机去放行李吗 | |
| No. | 不会 | 不会 | |
| No? OK. Then shut your pie hole | 那就闭上嘴<br>[Then, shut up the mouth.] | 那赶紧闭嘴<br>[Then, shut up the mouth right now.] | 4 |
| and listen to me | 听我说 | 听我说 | |
| when I say that I am finished with the checking-of-the-bags conversation! | 行李这件事情没得商量<br>[This luggage issue is non-negotiable.] | 行李的事绝对没得商量<br>[This luggage issue is absolutely non-negotiable.] | 5 |
| Sir, we have a policy on this airline that if a bag is this large, we··· | 先生　这么大的行李··· | 先生　这么大的行李··· | |
| Get your grubby little paws | 放开你那双肮脏的手<br>[Get your grubby hands off] | 赶紧放开你肮脏的手<br>[Immediately get your grubby hands off] | 6 |
| off of my bag, OK? | 别碰我的行李 | 别碰我的行李 | |
| Not like I have a bomb in here, it’s not like I want to blow up the plane. | 我又不是匿藏炸弹炸毁飞机 | 我又不是匿藏炸弹炸毁飞机 | |
| I wanna stow my bag according to your safety regulations. | 我只是想按规定放好行李 | 我只是想按规定放好行李 | |
| Sir, sir··· | 先生！先生！ | 先生！先生！ | |
| If you would take a second and take the little sticks out of your head and clean out your ears, | 你把你的耳朵掏干净听着<br>[Clear out your ears and listen,] | 你把耳朵掏干净好好听着<br>[Clear out your ears and listen carefully,] | 7 |

|  |  |  |  |
|---|---|---|---|
| maybe you would see that I’m a person who has feelings, | 也许你就会明白我想要做的 | 也许你就会明白我想要做的 |  |
| and all I have to do is do what I wanna do! All I wanna do is hold onto my bag and not listen to you! | 只是想保住行李　不想听你啰嗦 | 只是想保住行李　不想听你啰嗦 |  |
| The only way that I would ever let go of my bag | 你别想拿到我的行李 | 你别想拿到我的行李 |  |
| would be if you came over here now and tried to pry it from my dead, lifeless fingers. | 我会这样抓得紧紧的 | 我会这样抓得紧紧的 |  |
| OK? If you can get it from my kung fu grip, then you can come and have it. | 你要是能抢走　那就来抢 | 你要是能抢走　那就来抢 |  |
| OK? Otherwise, step off, bitch. | 不然就滚开得了　贱人<br>[Otherwise, step off, bitch.] | 不然就赶紧滚开　贱人<br>[Otherwise, step off right now, bitch.] | 8 |
| Get off of me! Get off of me! | 放开我啊　放开我啊<br>[Get off of me. Get off of me.] | 快放开我　快放开我<br>[Immediately get off of me. Immediately get off of me.] | 9, 10 |

## Appendix 3. Comprehension tests for each stimulus clip and the translated versions

### 片段 0 理解题

1. 片段中，男主角的行李为什么不能随身提上飞机？
2. 男主角的行李被弄丢了，里面有什么重要物品？
3. 男主角的衣服在机场被什么弄脏了？

### 片段 1 理解题

1. 片段中，刚下车的时候，女主角制止了男主角做什么事情？她是如何制止的？
2. 女主角见到爸爸后，两人激动地做了哪些动作？
3. 爸爸教会了小猫“黑仔”哪些动作？

### 片段 2 理解题

1. 片段中，男主角讨厌猫这件事，谁告诉了谁？
2. 片段中，谁给谁买了什么礼物？
3. 片段中，谁提问了谁什么行业的男人不多？

### 片段 3 理解题

1. 片段中，男主角的情绪如何？哪些言行细节能反映出他的情绪？
2. 男主角对托运行李的态度如何？在争执过程中，他对空姐说了哪些不客气的话？
3. 男主角最后有什么遭遇？他对此说了什么？

### Comprehension test for Clip 0

1. In the clip, why couldn't the man's luggage be carried on the plane with him?
2. The man's luggage was lost. What important item was in it?
3. What made the man's clothes dirty at the airport?

### Comprehension test for Clip 1

1. In the clip, when the two characters got out of the car, what did the woman stop the man from doing? How did she stop it?
2. After the woman saw her father, what actions did the two excitedly do?
3. What behaviors did the father teach his cat "Jinxy"?

### Comprehension test for Clip 2

1. In the clip, who told whom about the man's dislike for cats?
2. In the clip, who bought a gift to whom? What is the gift?
3. In the clip, who asked whom about a profession in which not many men work? What is that profession?

### Comprehension test for Clip 3

1. In the clip, how was the man's emotion? What details in his words and actions could reflect his emotion?
2. What was the man's attitude towards checked luggage? What unkind words did he say to the flight attendant during their dispute?
3. What happened to the man in the end? What did he say about it?

## Appendix 4. Perception scale for subtitle quality for each stimulus clip and the translated version

### 感受题

1. 根据你的感受，对以下陈述表达意见。

| 观点陈述 | 非常不同意 | 不同意 | 比较不同意 | 比较同意 | 同意 | 非常同意 |
|---|---|---|---|---|---|---|
| 我觉得字幕表达自然 | ○ | ○ | ○ | ○ | ○ | ○ |
| 我觉得字幕表达清晰 | ○ | ○ | ○ | ○ | ○ | ○ |
| 我觉得字幕表达简洁 | ○ | ○ | ○ | ○ | ○ | ○ |
| 我觉得字幕速度合适 | ○ | ○ | ○ | ○ | ○ | ○ |
| 我觉得字幕内容与画面内容一致 | ○ | ○ | ○ | ○ | ○ | ○ |
| 我觉得字幕表现出了角色的特点和情感 | ○ | ○ | ○ | ○ | ○ | ○ |

### Perception scale

1. Based on your feeling, express your opinion on the following statements.

| Statements | Strongly disagree | Disagree | Somewhat disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| I think the subtitles are natural. | ○ | ○ | ○ | ○ | ○ | ○ |
| I think the subtitles are clear. | ○ | ○ | ○ | ○ | ○ | ○ |
| I think the subtitles are concise. | ○ | ○ | ○ | ○ | ○ | ○ |
| I think the subtitle speed is appropriate. | ○ | ○ | ○ | ○ | ○ | ○ |
| I think the content of the subtitles is congruent (matched) with the content of the image. | ○ | ○ | ○ | ○ | ○ | ○ |
| I think the subtitles bring out the character’s traits and emotions. | ○ | ○ | ○ | ○ | ○ | ○ |

## Appendix 5. Personal information form and the translated version

个人信息搜集表

1. 姓名：__________

2. 年龄：__________

3. 专业：__________

4. 性别：
   ○男
   ○女

5. 观看外语影视节目时，在多大程度上你会看字幕？
   ○从不
   ○偶尔
   ○有时
   ○经常
   ○大部分
   ○总是

6. 你的泰语听力水平如何？
   ○完全不懂
   ○初级
   ○中级
   ○高级

7. 你刚刚观看的片段都是电影选段，你看过这些电影吗？
   ○看过（请写出电影名字）_________________
   ○不确定（请简单解释）_________________
   ○没看过

Personal Information Form

1. Name：___________
2. Age: ___________
3. Major: ___________
4. Gender:
   ○ Male
   ○ Female
5. When watching audiovisual programs in a foreign language, how often do you read the subtitles?
   ○ Never
   ○ Occasionally
   ○ Sometimes
   ○ Often
   ○ Almost always
   ○ Always
6. How is your Thai listening ability?
   ○ No proficiency
   ○ Elementary
   ○ Intermediate
   ○ Advanced
7. The clips you just watched are excerpts from a film. Did you ever watch this film before?
   ○ Yes, I did. (Please write down the film title) __________________
   ○ I’m not sure (Please briefly explain why) __________________
   ○ No, I didn’t.