More than Meets the Eye

A Reception Study on the Effects of Translation

on Noticing and Memorisation of L2 Reverse Subtitles.

Valentina Ragni

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds

School of Languages, Cultures and Societies

Centre for Translation Studies

November, 2016

Le Parole

Le parole
se si ridestano
rifiutano la sede
più propizia, la carta
di Fabriano, l'inchiostro
di china, la cartella
di cuoio o di velluto
che le tenga in segreto;
le parole
quando si svegliano
si adagiano sul retro
delle fatture, sui margini
dei bollettini del lotto,
sulle partecipazioni
matrimoniali o di lutto;
le parole
non chiedono di meglio
che l'imbroglio dei tasti
nell'Olivetti portatile,
che il buio dei taschini
del panciotto, che il fondo
del cestino, ridottevi
in pallottole;
le parole
non sono affatto felici
di essere buttate fuori
come zambrocche e accolte
con furore di plausi e
disonore;
le parole
preferiscono il sonno
nella bottiglia al ludibrio
di essere lette, vendute,
imbalsamate, ibernate;
le parole
sono di tutti e invano
si celano nei dizionari
perché c'è sempre il marrano
che dissotterra i tartufi
più puzzolenti e più rari;
le parole
dopo un'eterna attesa
rinunziano alla speranza
di essere pronunziate
una volta per tutte
e poi morire
con chi le ha possedute.

Eugenio Montale

Acknowledgements

I dedicate this piece of work to my family and in particular to my father, who, after working for 33 years as a surgeon, is now enrolled at university once again, at the honourable age of 67, to pursue his passion for archaeology. He is a never-ending source of inspiration and motivation. He gently reminded me of a truth I am often prone to forget: *in medio stat virtus*. He is a giant and I am very proud to be looking firmly towards the future while sitting on his shoulders.

This thesis was written under the supervision of Dr Richard Badger (School of Education) and Dr Serge Sharoff (Centre for Translation Studies) at the University of Leeds, to whom I am indebted for the stimulating discussions and precious advice.

I would also like to express my gratitude, in no particular order, to:

All my friends outside academia, who kept me sane and reminded me that there is more to life than work. Without them, my office would have become my home and I would be very vitamin D deficient by now.

The colleagues with whom I shared far more than just an office building, an affiliation or a research interest: Léa Pilsner, Rachel Johnson, Ikhlas Abdul Hadi, Beatrice Ivy, Harris Constantinou, Chris Norton, Joel Millward-Hopkins, Alejandra Ortiz, Gustavo Carvajal, Nicolò Maldina, Alessandro Medici, Bintan Titisari, Luca Antonazzi, Andrea Alvisi, Alexander Frascina and Faustine Roux, amongst others. They made the ebb and flow of writing a bearable, if not even enjoyable, experience. I have learnt a great deal about myself thanks to them all.

Various departments of the University of Leeds: Italian for the recruitment and teaching opportunities, the Centre for Translation Studies for having me in their cohort of Doctoral students, and Linguistics and Phonetics for lighting my passion for quantitative research through fruitful initiatives on statistics, R and eye-tracking.

The impeccable Mike Bellhouse and the Leeds Humanities Research Institute (LHRI) for providing an inspiring space to think, read and write.

Karen Priestley and Ingrid Sharp for their crucial support in the last stages of this project.

A final and special sign of gratitude goes to the Arts and Humanities Research Council (AHRC), the generous funding body without which none of this would have been possible.

Abstract

More than meets the eye: A reception study on the effects of translation on noticing and memorisation of L2 reverse subtitles.

This experimental study addresses one of the least explored audiovisual modes: reverse subtitling (L1 audio, L2 subtitles). Specifically, it investigates the effects of different translational choices on learners' noticing and memorisation of lexical items and grammatical structures. The participants were English (L1) native speakers learning Italian (L2) at an upper-intermediate level (CEFR B2). Formal similarity (literal transfer) and discrepancy (non-literal transfer) between L1 and L2 were compared to establish if and how they affect the learners during subtitle processing and recall. Does one of the two translation conditions yield a better recall rate in a verbatim memory post-test? This constitutes the main research question addressed in this study. The main hypothesis was that there would be a difference in recall by translation condition, with formal equivalence having a facilitative effect on memory and literal subtitles therefore being more accurately recalled by learners. To determine how the different subtitle translations were processed, attention allocation and noticing were investigated through triangulation of eye-tracking, the recall post-test and an open questionnaire, which allowed the subjects' thought processes to also be recorded. Subtitle-specific variables such as corpus frequency and linguistic category (lexicon vs. syntax) were also analysed. While participants watched the reversely subtitled clip, their eye behaviour was recorded using a Tobii X120 eye-tracker. After watching, participants answered the recall post-test followed by the open questionnaire and took part in a working memory control test.

Translation condition was found to influence recall, with literal translations yielding superior recall performance than non-literal ones. The data also showed that participants did notice a translation discrepancy. Eye-tracking findings reveal a complex relationship between language elaboration and memory, whereby comparable amounts of visual attention to two subtitle versions can result in significantly different recall. Moreover, considerable insights were drawn from the open questionnaire, indicating that qualitative data can provide a richer picture of processing and memory attainment and should more regularly support experimental studies. The results demonstrate that subtitle-specific factors like translation can indeed influence the viewer and should therefore be taken into consideration in the design of future subtitle reception studies. The mnemonic potential of reverse subtitles for foreign language learning is also confirmed, strongly suggesting that this subtitling mode should be reconsidered as a valuable tool in language learning and deserves a place in the foreign language classroom.

Future research could build on this study by using a larger sample size and more advanced statistical techniques, such as multilevel modelling. The results obtained highlight the complexity of the language faculty and call for additional reception studies where more fine-grained analyses further assess the effects of translation during the consumption of subtitled material. For instance, using a higher-frequency eye-tracker and considering more eye movement measures in the future will provide more precise insights in the reading process, enabling deeper understanding of information processing and memory retention, both crucial aspects in the development of foreign language skills.

Key words: subtitling, subtitle reception, translation, memory, working memory, recall, noticing, eye tracking, eye movements.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AV – Audiovisual

AVT – Audiovisual Translation

CD – Contextual Diversity

CEFR – Common European Framework of Reference for Languages

CILS – Certificazione di Italiano come Lingua Straniera

CLI – Cross-Linguistic Influence

CLT – Communicative Language Teaching

CPS – Character Per Second

EFL – English as a Foreign Language

ESL – English as a Second Language

FLL – Foreign Language Learning

FLT – Foreign Language Teaching

G1L – Group 1L (order of presentation: literal – non-literal)

G2N – Group 2N (order of presentation: non-literal – literal)

IFL – Italian as a Foreign Language

IPM – Instances Per Million

L – Literal Transfer Condition

LC – Listening Comprehension

LTM – Long-Term Memory

LTS – Long-Term Storage

L1 – First or Native Language

L2 – Second or Foreign Language

MCQ – Multiple Choice Questionnaire

N – Non-Literal Transfer Condition

RQ – Research Question

RST – Reading Span Task

RT – Reaction Time

SL – Source Language

SLA – Second Language Acquisition

ST – Source Text

STM – Short-Term Memory

STS – Short-Term Storage

TL – Target Language

TS – Translation Studies

TT – Target Text

WM – Working Memory

WFE – Word Frequency Effect

WPM – Words per Minute

CHAPTER I
INTRODUCTION

1.1 Chapter One Structure

This thesis presents an experimental reception study that explores the relationship between AVT (Audio-Visual Translation) and the language learner from a cognitive and acquisitional point of view. Specifically, it deals with the reverse subtitling modality[1] (L1 audio, L2 subtitles). In an attempt to measure potential incidental learning, this relationship is addressed through eye-tracking technology during the viewing experience as well as through measures of mnemonic performance immediately after viewing. In this introductory chapter, terms like translation, reverse subtitling, salience, noticing, attention, memory and eye-tracking will be presented in the context of this study. First, I will introduce the reader to the personal reasons that led me to embark on such research project (1.2), then I will outline the context in which the research takes place, its main framework of reference and its two core aims (1.3), and finally I will provide an overview of the structure of the thesis as a whole (1.4).

1.2 Personal Motivation

The idea and inspiration to carry out this research comes from my personal experience. As an EFL learner, I always found using video materials in class thrilling. I fondly remember the bi-weekly session at college where the teacher would take us to the computer lab and have us watch excepts from subtitled VHS tapes, such as Al Pacino's *Looking for Richard* (1996). All of a sudden, the opening of the Shakespearean play *Richard III* not only made more sense, but was much more easily memorised when it came to learning it by heart. As I went on to study translation at university, I developed an interest for a particular type of translation that went hand in hand with video material: subtitling. I enrolled on an MA course in Audiovisual Translation (AVT), during which we were encouraged to watch subtitled films both in our mother tongue and in our foreign languages. For the first time, the Italian films I watched both inside and outside the classroom had to have English subtitles so that others (course mates and friends with no knowledge of Italian) could understand and follow the plot: I was thus watching reversely subtitled video (L1 audio, L2 subtitles). I had never found myself in that viewing

[1] Throughout this piece of work, the terms 'mode', 'modality' and 'type' are used interchangeably and refer to subtitling specifically. For a detailed discussion on modes in AVT from a terminological and taxonomic point of view, see Gambier (2003) and Hernández Bartolomé and Cabrera (2005).

condition before, and I immediately noticed two things: first of all, despite not needing the subtitles, I could not avoid reading them; I had to make a conscious effort to concentrate on the images only. I would not follow them constantly, but at times I would look down without realising and find myself reading the script. Secondly, when I encountered Italian words or expressions that struck my attention (for example idioms, colloquialisms or *realia*[2]) for which I did not know the English equivalent, I would consciously look down to check how they had been translated into English. I would also realise if there were any deviances from the script or mistranslation in the subtitles. Through watching films in the reverse condition, I had learnt new English lexical items as well as grammatical structures and idiomatic expressions, and developed a much stronger awareness of cultural, conceptual, structural and stylistic differences between Italian and English. In both formal (watching films in class) and informal (watching films with friends) settings, this type of translation yielded incidental learning and made me realise aspects of the language input that I had not noticed before. It was through this personal experience that I developed a curiosity towards this specific type of 'unusual' translation modality and turned to the relevant FLL (Foreign Language Learning) literature to ascertain whether these empirical and personal observations were substantiated by research evidence. Why was I able to recall L2 strings encountered while watching reversely subtitled material much more vividly than those I came across while reading L2 print text? Why did I immediately notice lack of correspondences between the SL and the TL, was it due to the nature of my course and personal interest in translation, or would someone with my same language proficiency but no such interest do the same? Could this AVT mode, reverse subtitling, contribute to FLL and Second Language Acquisition (SLA)? If so, what are the underlying reasons of such advantages? What mechanisms are at play when we consume reversely subtitled AVT products and what variables influence FLL during this process?

1.3 Focus of Investigation and Research Niche

It is with the above questions in mind that this research project was born. A survey of the literature revealed that very little research existed on the use and role of reverse

[2] Words and expressions indicating culture-specific objects, activities, concepts or situations which are typical of a language but do not necessarily have an immediate translation equivalent into another, e.g. *guerrilla, ayatollah*, *mistral* or *puszta*.

subtitles in FLL, the literature being dominated by studies on standard (L2 audio, L1 subtitles) and bimodal input (L2 audio, L2 subtitles), despite the fact that the handful of relevant FLL studies on reverse subtitles carried out in the 1980s and early 1990s all yielded rather promising results (see 2.3). It progressively also became evident that it would have been difficult to determine the individual contribution of reverse subtitle input to FLL without having first an idea of how viewers (in this case, learners) process and relate to such complex input. Subtitling is a type of diamesic translation in polysemiotic media (Gottlieb, 2012: 37), where 'diamesic' refers to the diagonal shift from speech to writing, and 'polysemiotic' refers to the four channels acting simultaneously to convey meaning. To wit, the channels are commonly identified as follows (adapted from Sokoli, 2006; Zabalbeascoa, 2008; Gambier, 2012):

- the acoustic-verbal (dialogues, songs and paralinguistic features such as intonation and accents)
- the acoustic-nonverbal (soundtrack, noises, and paralinguistic features such as voice quality and volume)
- the visual-verbal (intertitles, subtitles, as well as letters, shop signs, street names, hoardings, etc.)
- the visual-nonverbal (the moving images and everything that visually composes them, from lighting and colour to proxemics, gestures, make-up, etc.)

Watching subtitled material is therefore a complex feat, one in which viewers need to integrate information coming from all four channels, thus very much resembling real-life experiences (Lavaur and Bairstow, 2011: 456). It is a rather peculiar type of information processing, where many elements intertwine to produce the intended message, and attention needs to be distributed simultaneously to different components of that message (Ghia, 2012b). Research looking at the cognitive processing of subtitles is crucial to a thorough understanding of how they can contribute to SLA and FLL, since, if a subtitle is not processed (therefore not read), the L2 it contains has no chance of undergoing the further mental processes that may lead to that piece of L2 being memorised. Here, 'cognitive' is intended as related to the selection, encoding, storage and retrieval of linguistic information (see 2.4.2). In order to learn new items or reinforce knowledge of already familiar L2 input through audiovisuals, the L2 needs to be perceived and recognised. In the case of reverse subtitles, the L2 occurs in the subtitles, so it is a case of *written* surface form perception and *visual* word recognition (as opposed to speech perception and aural word recognition). Knowledge about such subtitle reading behaviour can further inform and qualify the findings of acquisitional measures such as vocabulary post-tests, yet we still know little about what happens when we read in multi task situations, for example when we watch a subtitled film (Moran, 2012: 184). This state of affairs is all the more true about viewing a subtitled film in the reverse condition, since the handful of above-mentioned acquisition studies in the 1980s and 1990s did not include a processing component in their analysis, while more recent studies (e.g. d'Ydewalle and DeBruycker, 2007) involving a processing component (eye-tracking), did not include an acquisitional measure. To my knowledge, this is the first piece of work entirely dedicated to reverse subtitles specifically, which takes into account both processing and learning.

When dealing with linguistic input, "[t]he success of the [processing] system in dealing with a given input depends on the characteristics of the input and the information-processing ability (including knowledge and expectancies) of the perceiver." (McLaughlin et al., 1983: 137). Therefore, both the linguistic input and the viewer determine processing outcomes. As we will see in the next chapter, many input-specific characteristics have been found to have an effect on audiovisual processing, e.g. segmentation, video genre and frequency. Amongst these, it has been proposed that translation may also have an impact on the learner (Mariotti, 2002), yet the role of translation strategies in interlingual subtitling has not been addressed explicitly (Ghia, 2012a: 75). In this striking paucity of studies on translational aspects of subtitles in both an acquisitional and a processing perspective, the need for more experimental research assessing the effects of translation choices on the viewer-learner are evident. One exception to this trend is Ghia's (2012a) reception study on the effects of translation strategies on noticing in the context of standard subtitles. Ghia enhanced the linguistic input through translation strategies such as literal transfer, reduction and substitution, as to obtain both literal and non-literal L1 renderings for each L2 audio string. When the written text deviates from the source audio (non-literal translation), Ghia postulates a condition of translational salience (2.7.3), arising when an L2 item gains prominence during the delivery of linguistic input (ibid.: 52). Ghia recorded learners' eye movements to identify reading patterns in the two translation conditions, measured retention via a recognition memory post-test and discussed noticing by triangulating data from eye-tracking, memory scores and an open questionnaire. Within the reverse subtitling mode, the present study also looks at translation specifically within a FLL context in order to explore its potential as a learning tool. Like Ghia, and unlike much research that revolves around either the translated text or the viewer-learner, this study aims to relate the mechanisms of input translation and learner reception in order to establish whether specific subtitling choices are processed and memorised differently from others. In the following chapters, the effects of these translational choices on viewers' perception, noticing and memorisation of lexico-syntactic structures will be investigated. To enable a direct comparison to Ghia's parallel investigation with standard subtitles, formal similarity (literal transfer) and formal discrepancy (non-literal transfer) will be compared.

Alongside translation, a central aspect of the present study is the construct of noticing. Although noticing is considered by many to be central in cognitive approaches to SLA (Smith, 2012: 53), relatively few studies have investigated it experimentally, perhaps because it can be a rather elusive concept, easily confounded with other closely related constructs, or because there is no agreement among researchers yet as to how to define it (see 2.8). This is even more true in AVT, where, to my knowledge, this and Ghia's study (2012a) are the only investigations to tackle noticing and attention in the context of subtitle use for FLL, despite the fact that "[i]ssues associated with attention should (…) be of prime concern for materials designers in cases where multimedia settings oblige learners to process multiple information sources simultaneously as they carry out complex tasks." (Guichon and McLornan, 2008: 4). By looking at noticing, this work aims at triggering further relevant discussion on the topic as well as advancing an original methodological proposal as to how to empirically investigate the construct in AVT.

Drawing parallels between one's results and the findings of other comparable studies is crucial "in order to establish to what extent such results are generalizable to different settings." (Cop et al., 2015: 31). For this reason, Ghia's research on noticing in standard subtitling will be used as the closest point of reference for the present investigation. In the following chapters, her noticing study will be described and her results will be considered in relation to those produced in this study, thus providing comparable data on a parallel investigation carried out in a different setting (namely, that of reverse subtitles). Like Ghia's, this is one of the few studies that innovatively uses eye-tracking as a method to investigate noticing. However, this study expands on the construct of noticing as presented by Ghia by addressing existing noticing and attention research more explicitly, making the construct more central to the discussion, and striving to relate the findings to the existing literature on the topic, e.g. Tomlin and Villa's construct of attention (2.8 and 5.3.5).

In conclusion, the core aims of my research are to establish (a) if viewers (in our case, also Italian FL learners) notice translational discrepancies between ST and TT, and (b) if they recall L2 strings in the reverse subtitles more or less successfully when they share some formal (orthographic or structural) similarity with the L1. Data from eye-tracking, recognition post-tests and questionnaires will be analysed to assess whether the formal similarity (literal transfer) vs. formal discrepancy (non-literal transfer) distinction has a psychological reality and whether salience created through translation enhancement affects noticing of and memory for reverse subtitles. The two core aims of this piece of research will be addressed through five main RQs (Research Questions), which are introduced in section 3.4.

1.4 Overview of Thesis Chapters

This first chapter served as an introduction to the topic of investigation. It described the personal reasons that led to development of the research questions, outlined the central questions that led to the birth of this project, presented some relevant terminology, and introduced the reader to the research niche, context and aims, thus setting the backdrop against which the thesis will develop.

Chapter 2 collocates the study in its wider research surroundings. I will take a closer look at relevant areas of the published literature, analyse the findings and examine how they can inform the design and implementation of this study. The literature review develops along three major trajectories: subtitle translation, language processing and foreign language learning. These elements intertwine throughout the chapter and its different sections, articulating the theoretical framework of investigation. Areas covered in this chapter include translation and subtitling in FLL, formal similarity and CLI (Cross-Linguistic Influence), reading mechanisms and AVT processing, attention and noticing, memory, frequency and eye-tracking.

Chapter 3 addresses the specifics of the study, taking the reader through the different phases of the research implementation. It is at the beginning of this chapter that the research scope, objectives as well as the five RQs are introduced (table 2). Practical information is given on experiment design, video and subtitle materials, participant recruitment as well as data collection and processing. Dedicated sections describe the protocol used to create the two translation conditions (literal vs. non-literal), the recognition memory post-test, the open questionnaire, as well as the control tests for working memory and language proficiency.

Chapter 4 details the process and outcomes of the data analysis carried out using the statistical software R (R Development Core Team, 2015). The relationships among the experimental variables (listed in table 1) as well as those between experimental and dependent variable (scores on the recognition memory post-test) are explored. To provide answers to the five core RQs presented in chapter 3, a number of specific sub-analyses are undertaken. Moreover, control variables such as WM (Working Memory), proficiency, order of stimulus presentation, subtitle length and duration are also addressed herein. The by-subject results are presented first, followed by the by-item results. Details of several analyses are provided, including descriptives, graphics and tables.

Chapter 5 discusses the findings of the previous chapter and presents possible explanations for the patterns emerging from the analysis. The chapter covers the control variables first, and then moves on to the primary research variables, where each RQ is presented in turn, following the order of table 2. In an attempt to account for the behaviour of the variables of interest, the most significant statistical data are considered for the purposes of the discussion and linked back to the relevant literature examined in chapter 2. Limitations of the study and possible methodological improvements are also addressed at the end of this chapter.

Finally, the results are tied together in chapter 6, where a synoptic view and an overall personal interpretation are presented. The chapter first summarises the RQs and the corresponding findings, then it highlights the contributions to knowledge, reflects on how translation and reverse subtitling were found to influence FLL, describes the elements of originality of the study, indicates possible directions for future research and finally collocates the study in a broader context by examining its wider implications.

CHAPTER II
LITERATURE REVIEW

2.1. Introduction

The research project described in this thesis is interdisciplinary in nature, drawing from the fields of AVT, SLA, cognitive psychology, eye-tracking, reading and memory research, amongst others. This chapter presents a selected review of the published literature in these related fields and discusses its implications for the processing and reception of reverse subtitles in FLL, providing the necessary background information leading up to the research questions and the research design that will be described in chapter 3. I offer a critical analysis of the arguments, theories and approaches presented, while considering the main findings of applied studies that are relevant to the purpose, rationale and implementation of the present work, thereby establishing its theoretical framework of reference.

This literature review develops along three major trajectories: subtitle translation, language processing and foreign language learning. After a preliminary explanation of two key terms that characterise this thesis, namely that of 'reception' and 'experimental research' (2.2), a review of studies involving reverse subtitles in FLL is presented (2.3), followed by a brief overview of the history of translation in language teaching and the presentation of an alternative, cognitively motivated view on its role in FLL (2.4). I will argue that translation is not a uniform phenomenon as it has been presented in much of the published psychology literature, and that different kinds of translations can have an impact on language learners. To this purpose, closeness and distance between L1 and L2 will also be addressed (2.5), by analysing research on cognates and the CLI (Cross-Language Influence) effects that varying degrees of formal overlap can have on language processing and learning. After a review of these L1-L2 formal similarities, I will describe the reading process, in both audiovisual and more 'traditional' print texts. The section on reading research (2.6) addresses the mechanisms at work when processing written material, summarises the main findings of the literature on reading and discusses language learning from this perspective. The section on audiovisual processing (2.7) highlights the differences between print text and subtitle reading, provides an overview of this relatively new area of investigation and reviews the variables that influence the processing of audiovisual material, amongst which is subtitle translation. Intrinsically linked to language processing and learning are the cognitive functions of attention and memory. Therefore, two dedicated sections on attention and noticing (2.8) and memory (2.9) respectively will describe how these cognitive functions have been theorised and investigated in the FLL and AVT literature, as well as how they will be operationalised in the present study. A section is then devoted to eye-tracking (2.10) specifically, since this increasingly popular technology has been instrumental to much of the research that will be described throughout this chapter. This section discusses some major eye-tracking measures, presents practical applications of eye-tracking to noticing and AVT research, and explains how this technology will be used in the present work. A section is also dedicated to frequency (2.11), its well-documented effects on language learning and processing, its critical reassessment and the rationale for its inclusion as a variable in the present study. The last section of this chapter addresses research validity, a crucial element of any quantitative, experimental study. The section describes how the concept of validity has been presented in the experimental psychology and linguistics literature, its multiple facets and how these are accounted for in the present investigation (2.12), thus providing a link to the methodology (chapter 3).

2.2 An Experiment on Reception

This study focusses on the reception of an underexplored translation mode, reverse subtitles. Reception can be considered from different viewpoints. Kovačič (1995) distinguishes between four distinct aspects of subtitle reception: *socio-cultural, attitudinal, psychological and perceptual.* The first aspect (socio-cultural) considers how the receiving process is affected by cultural and social variables external to the television dimension (see also Gambier, 2003, 2012) and acts at a supra-individual level (Brems and Ramos Pinto, 2014); the second (attitudinal) deals with viewers' preferences and habits regarding different modes of AVT; the third (psychological) explores readability issues and the psycho-cognitive side of processing translation; finally, the fourth (perceptual) is concerned with the decoding of translation at the behavioural level. These four aspects have more recently been re-categorised by Gambier (2003, 2012) in three types of reception, the so-called 'three Rs': *repercussions* in cultural terms (including attitudes), *reactions* in cognitive terms and *responses* in behavioural terms. The present study addresses reception in the latter two senses. It looks at reception at the individual level of 'real' readers (Brems and Ramos Pinto, 2014), striving to explore the cognitive processes at play during AVT consumption (*reaction*), which includes the study of eye movements as indicators of attention allocation (Orrego-Carmona, 2016). It also explores the perceptual effects of specific linguistic and technical aspects of the translation on the viewer who, in this context, is also a language learner (*response*). These responses include the viewer's general understanding of the content of the video, their subjective judgements of the subtitle translation and their opinions on the difficulty and/or enjoyment of the AV materials (ibid.).

This is also an experimental study. The experimental method "consists of abstracting relevant variables from complex situations in nature and reproducing in the laboratory segments of these situations, varying the parameters involved so as to determine the effect of the experimental variables" (Orne, 1962: 776). More specifically, in psychology research, an experiment is "a systematic research study in which the investigator directly varies some variable (or variables), holds all other factors constant, and observes the results of the systematic variation." (Goodwin, 2008: 163). According to Goodwin (2008), experiments have three core elements: the presence of at least one experimental variable with at least two levels (independent variable), the control of confounding variables (extraneous variables) and the measurement of results (dependent variables). As such, they also involve the direct manipulation of variables (Field and Hole, 2003: 10) as well as randomisation (ibid.: 71), that is to say both random subject selection and random assignment to the experimental groups (Goodwin, 2008: 198). The present study respects all these parameters, details of which are presented in chapter 3. The main independent variable of interest is translation condition, the dependent variable is recall scores in a memory post-test, and extraneous variables are accounted for through experimental design (3.2-3.5, 3.7, 3.11-3.16), participant selection (3.8), control tests (3.9 and 3.10) and data collection (3.17).

Despite the recent surge of interest in AVT processing and reception (see Kruger, 2016), the need for more experimental studies investigating "viewer processing habits, reading strategies, and reception patterns" (Gambier, 2016: 900) in AVT is still felt. The present work starts closing this gap and actively contributes to these research areas by advancing knowledge on what factors affect the relationship between reverse subtitles and the language learner.

2.3 (Reverse) Subtitles in the FLL Literature

From the 1980s onwards, the advantages of using subtitles for FLL purposes have been extensively investigated from different perspectives and for different proficiency levels. Standard subtitling and bimodal input (L2 audio – L2 subtitles) are the two most investigated modalities, both theoretically and experimentally. Bimodal input, in particular, also called same-language subtitling, monolingual, unilingual or intralingual subtitling (Jung, 1990) was found to be beneficial for video comprehension (Price, 1983), linguistic content comprehension in advanced learners (Garza, 1991), oral communicative performance (Borras and Lafayette, 1994), chunking ability, mnemonic retention and language development (Vanderplank, 1988, 1990, 1993), auditory processing (Baltova, 1994), aural word recognition (Markham, 1999), motivation, attention and vocabulary use in post-watching activities (Smith, 1990), reading proficiency and incidental vocabulary learning (Neuman and Koskinen, 1990), reading development (Kothari, 2008), recognition memory (Bird and Williams, 2002) and even L1 literacy development (Kothari 1998).

More recently, standard subtitling has also been the object of research. A number of studies demonstrated positive effects of this aural-written input combination with regard to general language proficiency and productive skills (Ghia, 2012a), acquisition of foreign vocabulary in adults (d'Ydewalle and Pavakanun, 1995, 1996, 1997; Pavakanun and d'Ydewalle, 1992) and children (Koolstra and Beentjes, 1999, d'Ydewalle and Van de Poel, 1999), cognitive effectiveness (Perego et al., 2010), language maintenance and processing (De Bot et al., 1986), text comprehension (Bravo, 2008) and listening and speaking skills (Araújo, 2008). The effects of standard subtitling on the acquisition of grammar rules pertaining to morphology and syntax are still unclear (Van Lommel et al. 2006, Van de Poel and d'Ydewalle, 2001) although more recent studies found significant effects of improved syntactic competence (Ghia, 2007, 2011, 2012a) after prolonged periods of exposure (Ghia, 2012a: 108).

On the other hand, much less attention has been devoted to reverse subtitling, perhaps because watching video in this condition necessarily entails a loss in terms of L2 rich aural input, as the dialogues are in the native as opposed to the foreign language of the viewers. However, listening to dialogues in one's native language can present an advantage during the learning process. Making comprehension effortless, L1 dialogues free the learners from the possible distress arising from having to devote a great deal of attention and concentration to the L2 speech stream and the feeling of struggle and anxiety that might arise if the learners fail to detect the boundaries between words and successfully follow the dialogues (Vanderplank, 1988: 217). When the L2 aural input is very fast or structurally complex and the images do not provide enough contextual clues, the learners' overall cognitive capacities can become overtaxed, which in turn can cause them to lose interest or concentration and eventually give up, ceasing to follow and 'screening' out the input, therefore losing a learning opportunity (ibid.). In the reverse condition, not only does this video content problem not arise as full comprehension of L1 dialogues is achieved, but learners are also allowed to devote their spare cognitive resources to the written L2 input and thus might be more likely to attend and notice salient aspects of the L2. Moreover, it has been proposed that written information tends to leave a deeper memory trace than auditory (Field, 2004: 318). The fact that the L2 appears in written format might therefore make L2 subtitles a valuable tool for the acquisition of L2 skills other than listening. Lambert et al. (1981) provide evidence towards this view in their first pioneering study on the effects of different types of captions on memory retention and comprehension. Nine combinations of spoken dialogue and printed script were tested, amongst which are bimodal input, standard and reverse subtitling. Radio broadcasts were used, alongside word-for-word scripts presented on a TV screen. Grade 5 and 6 functional bilinguals (English native speakers who attended French immersion programmes) were tested on comprehension as well as on contextual meaning, spelling and exact phrasing. Lambert's study confirmed that information is processed more thoroughly when script and audio are matched, as opposed to script- or audio-only conditions, with certain combinations being particularly promising, in particular reverse subtitling. This presentation modality was found to be the most beneficial and produced better results than standard subtitling across all testing conditions, scoring best (out of nine presentation conditions) for contextual meaning and spelling, and second-best for comprehension and exact phrasing. Lambert et al. also argue that listening to the message in the L1 through the more transitory audio channel allowed students to review or discover the L2 format, which would prime them for the testing phase and permit further grasp of L2 details (1981: 139-140). They also suggested that these promising results were in line with the 'depth of processing' view and research on effective reading. Effective reading entails entering a text with a set of hypotheses to test (top-bottom nature of information processing), rather than constructing words from phonemic sub-units to build contextual meaning (bottom-up processing). Thus, the learner would swiftly process the auditory L1 input and fully understand its message, which in turn would allow them to approach the L2 subtitles with a set of expectations about what to look for in the foreign script.

Holobow et al. (1984) carried out a similar study to see if the effectiveness of reverse subtitling would hold over time on a ten-week treatment period. Three conditions were compared: reverse, bimodal and L2-script only. Grade 5 and 6 English native pupils with advanced training in French were tested on comprehension, contextual meaning and memory for exact phrasing. The advantages of reverse subtitling were confirmed as this presentation modality scored best in tests of comprehension and contextual meaning and second-best in memory for phrasing. The L2-script only (no dialogues available) was the least effective modality in all tests. Reverse subtitling and bimodal input were found to be promising devices for second language learning as well as comprehension enhancement, and possible practical and pedagogical applications were suggested. In a third follow-up study, Lambert and Holobow (1984) confirmed and strengthened the above experiments' conclusions as reverse subtitling proved to be the most beneficial conditions also at a lower level of proficiency (participants selected were English native pupils with a much lower mastery of French than the functional bilinguals of the two previous studies). Here L2 script only and bimodal L2 input were the least beneficial methods, suggesting that beginners need some form of help from their native language to fully understand a message.

In these early studies reverse subtitling refers to a presentation condition where aural L1 information is read aloud by the experimenter or played back as recorded from radio broadcasts, while subjects read in silence the provided correspondent L2 script. The interplay is therefore between two channels, the aural and written verbal. No images or video were used. Based on Paivio's bilingual Dual Coding Theory and image-superiority approach (1986), adding another channel (the visual non-verbal, that is, the images) that binds together the other two providing contextual clues and referential information should also be beneficial for memory retention. According to Paivio (1986), memory and cognition are characterised by two discrete systems of symbols, the verbal and the non-verbal memory codes, which are independent but partially interconnected. The independence assumption postulates that the two systems, if acting together, have additive effects on recall. Thus, if items are presented as images plus verbal labels, they will be better recalled than if they are presented just as image or just as words. In the bilingual version of the theory, Paivio and Desrochers (1980) assume that bilinguals have two separate verbal memory representation systems which interact only through translation, and a common imagery system. Due to the independence assumption, bilingually encoded items (translated words) should therefore have additive effects on recall compared to monolingually encoded ones (repeated words). This assumption was evidenced through two experiments (Paivio and Lambert, 1981) where subjects were presented with pictures, French words and English words which they had to encode into English (naming of objects in the pictures, translating from French, simply copying the English word) before being unexpectedly asked to recall the English words. They found that the recall rate for the picture naming task yielded the highest recall rate, followed by the translation task, and then the copying task. Recall rate for translated words was twice that of the copied words, suggesting that the two discrete verbal codes do have an additive effect on memory when interacting (Paivio and Lambert, 1981: 534). The findings sustain the well-established image-superiority hypothesis (Paivio and Csapo, 1973), whereby pictures and images are more likely to be remembered than words, whilst also prove that translation effectively enhances recall compared to a monolingual condition. This finding might also be explained by the levels of processing theory (Craik and Lockhart, 1972), i.e. by proposing that translation requires deeper or more demanding processing than the mere copying of words. Paivio's theories were adduced in explaining the results of another widely cited piece of research carried out by Danan (1992), this time with subtitles added to actual video (moving images). In this study, English native university students with beginner and intermediate knowledge of French were tested on vocabulary recall after watching a video excerpt in the following conditions: French (L2) audio only, reverse subtitles, bimodal L2 input. Students were tested by means of an active recall test in French (fill-in-the-blanks). Correct recall rates in the reverse condition were highest for both proficiency groups, significantly so in the case of beginners, where recall more than doubled compared to the L2-audio only condition (ibid: 515). A background questionnaire was also administered, after which two additional tasks were undertaken. Firstly, pupils had to check previously known items from a list of French words encountered in the test. Then, they had to translate all words in the list into their L1. Through these two additional assignments, Danan was able to provide a measure of prior knowledge (correctly translated checked items) as well as vocabulary learning (correctly translated unchecked items). Translation results were highest in the reverse condition, and for both proficiency groups. The study indicates that reverse subtitling greatly helped activation of the subjects' prior knowledge (in particular for beginners), and that learning did occur with unfamiliar items (ibid: 515). Interestingly, Danan reports that subjects in this group could translate previously unknown words despite not free-recalling them correctly in the memory post-test, strongly suggesting that passive vocabulary learning did indeed occur. In Danan's view, the increased L1-L2 mapping and deeper encoding in this condition allowed subjects to recognise and understand items in the translation task. As in the Lambert's study, processing *written* L2 strings (in both reverse and bimodal conditions) clearly improved memory recall. This is in line with the common visual-superiority views and with studies on subtitling where encoding and retrieval where found to be most successful when input is given visually rather than aurally (d'Ydewalle and Gielen, 1992).

These advantageous results of reverse subtitles in Danan's study were explained through the dual coding approach, where bilingually encoded words (translated items) are better recalled than their monolingually encoded counterparts. The above considerations would explain why the reverse subtitling condition was overall more effective than the bimodal condition, where no translation was involved. The fact that the reverse mode ranked much higher than the standard (included by Danan in a first pilot but subsequently substituted by the bimodal condition) despite both modes involving translation, could be explained by the translation direction, as in the four major studies addressed above. If coupled with foreign script, listening to dialogues in one's native language does indeed appear to present an advantage in terms of L2 memory retention, and this is regardless of whether extra non-verbal contextual information is simultaneously presented through video images. Effective and automatic aural processing of the L1 seems to provide the learners with an ideal semantic framework of reference enabling a more focussed approach to the foreign input where hypotheses and expectations on the L2 form can be tested. Moreover, as previously mentioned, full understanding of the story line prevents possible frustration arising from unsuccessful message content comprehension which might screen off the L2, as posited by the affective filter hypothesis (Krashen, 1985). Thus reverse subtitling can be seen as having an advantage over its standard counterpart, where students are more likely to be disappointed by not understanding the audio and having to rely on subtitles; in a situation where comprehension is not an issue and anxiety levels are under control, a student can approach the new input with a more positive attitude and might be more likely to remember how a certain word or expression is rendered in L2 writing.

The studies described above clearly show that using reverse subtitles can be a successful strategy for L2 encoding, retention and maintenance, memory for exact phrasing, activation of prior knowledge, and vocabulary learning. Even Vanderplank, in his renowned article on the effects of (monolingual) teletext subtitling on EFL learners, deems it worth mentioning how, over a two-year period he himself spent in Finland, where he was watching L1 English and American programmes with Finnish subtitles (reverse condition), his "knowledge and understanding of Finnish was helped through reading the Finnish subtitles" (1988: 281). Despite the anecdotal and experimental evidence reviewed above, very little attention has been reserved to this AVT mode since the late 1980's, with the exception of a handful of studies (e.g. Čepon, 2011; D'Ydewalle and De Bruycker, 2007). The present investigation therefore reassesses reverse subtitling in light of the benefits just discussed and further explores its potential. Moreover, most studies concerned with the use of reverse subtitles are comparative in nature, having addressed the topic mainly in terms of (dis)advantage(s) in relation to other modalities. Unlike with standard subtitling and bimodal input, where initial contrastive research led to several separate individual-condition analyses in relation to further specific theoretical and practical questions, at the time of writing virtually no investigations specifically concerned with particular aspects of research (e.g. exposure and learning) within the reverse subtitling modality *per se* have been proposed, despite evidence suggesting rather robust reasons to carry out this type of research (with one single exception to date, reviewed immediately below). Thus, central to the goals of the present analysis is to focus on this type of audiovisual translation in order to examine its effects on language learners, discuss theoretical and practical implications for FLL and FLT and, if in order, propose an argument for its reconsideration.

The only exception to the 'contrastive research trend' in reverse subtitling mentioned above is a recent study by Čepon (2011), who tackled the impact of this AVT mode on a number of FLL-related measures, namely lexicon and grammar acquisition, reading comprehension (receptive), spelling and writing skills (productive). Through surprise tests administered on these measures before and after viewing the experimental subtitled video, Čepon addressed the question of whether any incidental learning of L2 English in 77 L1 Slovenian university students (B2 level) through reading reverse subtitles would be possible. After the post-tests, he also held in-depth interviews with the subjects to collect their personal opinions on the viewing experience. He found some post-test improvement due to video exposure in all measures (as evidenced by significant χ^2 tests), with greater effects on writing compared to reading comprehension. Interestingly, the fewer the mistakes in the pre-tests, the smaller the improvement in the post-tests, and vice-versa, the more incorrect answers students provided in the pre-tests, the greater the improvements in the post-tests. Vocabulary showed the greatest improvement, while grammar was found to be the least affected measure (ibid.:20). However, based on his results, Čepon states that reverse subtitling "may potentially enhance already existing FL grammar knowledge and exert mild effects on grammar acquisition/revision" (ibid.). In this sense, he argues for grammar what Danan (1992) found for vocabulary, i.e. that reverse subtitling helps reactivate previous knowledge. His interviews also revealed that more than half the subjects approved of reverse subtitles as a novel and promising FL activity. Čepon underlines how his study is the first to deal with reverse subtitling through a combination of productive and receptive skills. In line with his research, the present study also considers both, i.e. investigates visual word *recognition* memory (receptive skill) through a MCQ (Multiple Choice Questionnaire) post-test, as well as active, unprompted *recall* (productive skill) though an open questionnaire[3]. Taken together, these measures provide a more comprehensive picture of the effects of reverse subtitles on memory, allowing to draw more precise insights into some of the processes at play in the learner's mind. Finally, Čepon reports that "quite a few [students] were able to recollect the exact moments when they started paying attention to reading reverse subtitles – namely, when they were *especially interested in how certain Slovenian words were translated*, either new words or the ones they were just not familiar with" (2011: 16-17, my italic). It thus emerges that students can indeed concentrate on L2 subtitles and did appreciate being able to immediately check L1 words that they found striking against the L2 subtitles. In other words, translation seems to be able to trigger focused attention and interest, potentially making it a useful tool exploitable by language teachers. However, as we will see, this potential has been ignored for several years and is still not fully accepted in much of the current FLT literature. This research project therefore also positions itself in and contributes to a much wider debate on the role of translation in the learning and teaching of a foreign language, the topic of the next section.

2.4 Translation in the FLL Literature

2.4.1 Translation in SLA and FLT

Translation is one of the more traditional techniques used in language teaching (Scheffler, 2012: 604). However, it seems to have been criticised or at best neglected in first and second language acquisition studies (Anderman and Rogers, 1996: vii), SLA Theory, and communicative language teaching (CLT). Perhaps for rhetorical effect, Cook goes as far as saying that translation has been "treated as a pariah in almost all the fashionable high-profile language teaching theories of the 20th century" (2010: xv). Historically, with the exception of the unpopular grammar-translation method, in early teaching approaches such as the direct method and the audiolingual method translation was indeed to be excluded from the classroom. Then the development of more meaning-oriented and cognitive approaches in the 1970s paved the way for the functional-notional and communicative approaches widely used today, within which translation may be tolerated but is generally discouraged or at least not emphasised (Hummel, 1995: 448-449).

[3] For more information on recognition and recall memory tests, see section 2.9 in this chapter.

According to Cook (2010), the reasons for this ostracism are manifold: pedagogic, in that translation was considered tedious and demotivating; cognitive, as translation was believed to hinder rather than support L2 acquisition and processing; and practical, as it was not thought to be an activity needed by learners in the real world. In spite of this, translation remains central in many educational contexts all around the world (Benson, 2000): it is still part of syllabuses and examinations, and it still practised in many classrooms, especially at university-level (Malmkajær, 2004). Despite the attention reserved by contemporary L2 teaching approaches to cognitive and meaningful communicative activities, translation has not generally been identified as having potential value in the foreign classroom. However, the role of translation in the language classroom is slowly being reappraised in works such as that of Leonardi (2010, 2011). Her fundamental assumption is that translation is a natural cognitive activity happening in the mind of the learner and that, at least in the early stages of acquisition, learners do filter the L2 input via their L1. Since it is not possible to stop learners from translating, Leonardi proposes to add translation as the fifth skill to complement the four core skills of language learning, namely reading, writing, listening and speaking (2011: 18). She calls this practice 'pedagogical translation', where translation is intended both as the activity of translating texts as well as the use of already translated materials such as glossaries and the like. Leonardi also cites psycholinguistic evidence when she states that "the stored translation equivalents in the brain have a stronger basis in terms of imaginary representation and thus, they are remembered better." (ibid.). As we have seen in section 2.3, and as we will also see from a different perspective in 2.5, this psycholinguistic evidence comes from different experiments showing that using translation can have certain cognitive advantages, which would make it a valuable pedagogical tool, or at the very least one whose potential needs exploring.

2.4.2 Translation: A Cognitive Approach

In her review of the psycholinguistic support to translation, Hummel (1995) touches upon the levels of processing framework (Craik and Lockhart, 1972), according to which the processing activity carried out on a stimulus leads to memory traces which are qualitatively different in terms of level or depth, with meaningful, semantic information being more deeply encoded compared to the shallow processing of non-meaningful input. The elaborateness of processing view (Anderson, 1990) is a later model which expanded on the original version with the intention of accounting for differences in recall for input processed at the same depth or level. Traces within a single level could still differ in terms of elaboration, and it is posited that more extensive analysis of a stimulus results in a more elaborate trace that is more likely to be remembered. In this view, both input and the related elaborations are encoded into an internal network of propositions, and activation of these traces is necessary for the subject to remember the information. If the subject generates a mnemonic episode with multiple and partially redundant elaborations for the same information, they will have more chances of recalling that piece of information. The model thus suggests that inferential redundancy or interconnectedness of propositions is an important factor to improve memory retention, and that quantitative differences also count, as more elaborated traces will carry with them additional information used to produce the interconnections (Hummel, 1995: 452). In the rich context of watching audiovisual material, translation can be seen to establish meaningful semantic connections between the two verbal systems, which can be further strengthened by the constant presence of extra referential information in the video. Translation can also be seen as a way of creating a more elaborate link between words, as the language learner receives a coupled set of elaborations for concepts conveyed in the source text (e.g. a sentence in the L1 audio will be coupled with the L2 subtitle). Bilingually encoded words would therefore allow for greater elaboration during processing, which in turn would make information more memorable. In this respect, therefore, this view converges with the Dual Coding Theory (see 2.3). Hummel also quotes support from theories of information processing in reading and active hypothesis testing, which were also mentioned in the Lambert experiment. The active testing of hypotheses occurring in this specific direction (from L1 audio to L2 script) allowed by reverse subtitling would explain why the other condition in the Lambert experiment where script and audio are matched (standard subtitling) was not as effective: it is in the reverse direction that the top-down nature of information processing is best embodied, as listening to the L1 dialogues enables viewers to quickly process L1 information and efficiently spend the remaining time trying to map the more easily understood message content with its related L2 written form. I would therefore argue that mapping known to unknown information can be a more successful strategy than mapping unknown to known information in this setting. If the audio is in the L1, it is very unlikely that information will be missed out, so the learner will have a clear picture of what is happening in the clip and it is with this certainty that they will approach the L2 written input to look for ways of expressing those concepts in the L2 and test their hypotheses. If the audio is in the L2, viewers might miss or not understand parts of the speech stream. They are exposed to the L2 in the more transient and less memorable aural form, and they then have to make sense of this possibly patchy L2 information by using L1 subtitles. It is understandable how students might simply naturally end up relying heavily on the native-language subtitles and ignoring the foreign audio, especially if the level of the clip is not carefully tailored to their level of proficiency. Moreover, even assuming that the audio is not ignored and the learner actively tries to match what they hear (L2) with they see (L1), if an aural string is lost (i.e. not understood or heard properly) then the mapping to its equivalent L1 form would not be as stable, in that one of the two terms of the connection (the L2 term provided aurally) will not be established with certainty, leaving the incontrovertible written L1 input to be matched to a piece of less clearly identified L2 input. This could be particularly valid with a language like L2 English, where the relation between phonetics and orthography is not always straightforward, and the same string of sounds (e.g. /breɪk/) can sometimes refer to two words with very different meanings (*break*, e.g. a coffee break, and *brake*, e.g. a car brake), adding a level of potential ambiguity for the ESL learner who is trying to process an English term or expression given aurally. Thus, it could be the case that the increased mapping yielded by bilingual input is not enough *per se*, that both terms of the connections need to be meaningful and make sense to the learner in order for multiple memory paths to result in more elaborate (and less likely to be forgotten) connections. Bilingual memory traces established through translation during consumption of standard AV material (from L2 audio to L1 script) could therefore be less effective than bilingual traces established from L1 audio to L2 script (reverse), especially when the L2 input conveyed aurally in the standard subtitling condition is not fully comprehended by the learner.

In recent years, some AVT scholars have also adopted a cognitive approach to translation (Pavesi, 2002; Perego, 2012; Bairstow, 2012; Ghia, 2007, 2012a; Caimi, 2005, 2006, 2012), where cardinal cognitive principles underscore the link between thought, meaning and linguistic structure, and language is considered to directly reflect cognition (Lee, 2001). Drawing from Biehler and Snowman (1993) Caimi states that cognitive science:

> describes how information is processed, the ways in which individuals pay attention to it, their being aware of whether the information input is true or false, their ability to select and encode it and, finally, their capability to store it in memory for various lengths of time and to retrieve it from storage and use it for communicative purposes (2012: 117).

Within a cognitive approach to AV language processing, reception studies aim at understanding how exactly a film and its translation are perceived by the viewer (Bairstow, 2012). It is within this framework of reference that the present study investigates whether the status of translation as *persona non grata* in FLL can be reviewed in light of the potential cognitive benefits it may bring to the learner. In Danan's experiment, translation seemed to be able to link the two verbal systems during processing, thus allowing, together with contextual visual clues, the establishment of multiple paths for retrieval, and leaving a deeper memory trace. Whilst Danan's study highlighted the need to further explore the potential of translation in an audiovisual setting, it was not followed by any specific investigation on how translation is perceived by the viewer in this modality. As the cognitive processing of subtitles is still an under-investigated topic (Perego et al., 2010: 244), further research on how particular features of subtitled material affect processing and memorisation strategies is called upon. And it is precisely these translational features of L2 subtitles that this piece of research is concerned with.

2.4.3 What Translation?

What none of the cognitive models and studies described above specify is what is meant exactly by translation. Translation is not a uniform phenomenon in the sense that the same piece of L1 input (be it a noun, phrase, sentence or text) can result in several correct and perfectly acceptable L2 translations. So if an L1 sentence is given to *n* professional translators, potentially *n* different but equally valid translations could be produced as output. However, in the accounts analysed in the previous sections, translation is treated as a 'uniform' phenomenon, as if only a one-to-one L1-L2 equivalence and one output was possible. This is clearly not always the case, and the way in which translation is carried out is very likely to have an effect on the way bilingual input is perceived by the viewer, especially in a situation like subtitling where both languages are used simultaneously and can therefore be compared on a moment-by-moment basis. Moreover, in most psychology studies on translation, the investigation has been concerned with individual words at a micro-level (often taken completely out of context, for example as part of a vocabulary list) rather than phrases and sentences at a macro-level, although the latter is also crucial to achieve a thorough understanding of how translation impacts the learner. Despite the above, "No explicit attention has been devoted to the role of translation in interlingual subtitling" (Ghia: 2012a: 75). In other words, very few accounts have tried to explain if and how different translation strategies (resulting in different translational outputs) are perceived by viewers. This is herein deemed essential if we are to assess the role played by translation in the acquisition process. In Danan's, Lambert's and Holobow's studies, would results have changed if different translational outputs had been chosen for testing, for example an L2 version more formally distant from the original L1? When watching audiovisual material, learners tend to compare or relate aural and written information (Ghia, 2012a: 35). It seems therefore reasonable to assume that the manipulation of aural or written input can affect encoding during subtitle consumption, where encoding is defined as the first process in forming a memory, in which information from sensory input is changed into a form that the memory system can deal with and store for later use (McLeod, 2013). In AVT practice, several strategies are commonly used by professional subtitlers to find the formally and semantically closest equivalent in the TL while reducing and converting the ST from speech to script. However, very little experimental research has so far addressed these strategies in terms of impact on and usability for language learners (Perego and Ghia, 2011: 179). In cognitive terms, for example, is processing more taxing with translations diverging from or similar to the soundtrack? And during the mapping process, does a literal rendering leave a deeper memory trace than a much freer gist translation? Or is the opposite true? Would the learner notice the L1-L2 discrepancy with translations more distant from the original? It is with these questions in mind that the experiment presented in chapter 3 was designed. Translation is the focus of this investigation, and in particular its degree of discrepancy from the source as created by translational strategies. In this sense, the study is exploratory, as it aims at assessing whether the distinction between closeness vs. distance from the source in audiovisual translation has a psychological reality in terms of perception and memory, and whether it can therefore be a viable route to investigate the effect of linguistic differences in the subtitles on the viewer.

Distance from the original, equivalence, accuracy and literal translation are all concepts thoroughly debated in the fields of Translation Studies and Translation Theory. The very concept of equivalence, cardinal to much discussion in these fields, has been conceptualised from a different, much narrower angle in cognitive psychology research. For example, in their analysis of cognates and interlingual homographs[4], Dijkstra et al. (1999) consider translation equivalents words displaying both similarity of form (e.g. similar spelling) *and* semantic identity (same meaning), i.e. they address L1-L2 equivalence

[4] On cognates and interlingual homographs, see 2.5 (notes 7 and 8).

from a purely orthographic and semantic point of view, and only for single words taken in isolation. A translation scholar like Nida, on the other hand, for whom translation reproduces "in the receptor language the closest natural equivalent of the source-language message, first in terms of meaning and secondly in terms of style" (Nida and Taber, 1982: 12), clearly addresses the matter from a rather different view-point. As an investigation concerned primarily with translation, my work starts from the notion of equivalence[5] as discussed in TS rather than in psychology. Definitions given in the latter field rarely allow to go beyond the word-level, yet this study is concerned not only with individual words presented as a list, but phrases and structures occurring in context. Using Dijkstra et al.'s definition (1999) would prove too narrow and inappropriate to capture the macro-level changes often necessary in AVT. Therefore, inspiration to conceptualise the closeness vs. distance distinction was drawn from key concepts in TS, such as formal and functional equivalence (Nida and De Waard, 1986). Since the early age of pre-linguistics writings on translation, opinion has "swung between literal and free, faithful and beautiful, exact and natural translation" (Newmark, 1981: 38). In fact, the distinction between word-for-word and sense-for-sense (i.e. literal and free) dates back to Cicero and St. Jerome (Munday, 2008: 19). Literal translation is therefore a widely understood concept, yet its quantification for the experimental purposes described above does present a challenge, as the natural distance between two languages requires that, in order to preserve the content of the message, its form must sometimes be changed. Given the lack of preceding studies in AVT exploring the effects of types of translation on the viewer (except Ghia, 2012a and 2012b), and considering the wide range of in-built differences between languages, an established general routine for operationalising translation in this sense simply does not exist. In subtitling, the term 'literal transfer' was first introduced by Gottlieb (1992) and refers to a target text matching the source as closely as possible in terms of both meaning and form, sometimes resulting in a word-for-word rendering, yet retaining L2 grammaticality. In this study, formal similarity and discrepancy in translation are thus operationalised as literal and non-literal transfer, and in terms of degrees of formal and semantic distance from the source. The closest degree of literalness was achieved first (literal transfer) and a more formally distant translation was thereafter produced in order to create a discrepancy (non-literal transfer). A range of translation procedures was used to achieve this. Central to this process was Vinay and

[5] For a detailed description of how the concept of equivalence is used in practice in this study, see 3.11.

Darbelnet's model (1995), where two main strategies are identified: direct and oblique translation, which refer back to the literal vs. free distinction mentioned above. They identify seven procedures: borrowing, calque, literal translation, transposition, modulation, equivalence and adaptation. Deslile et al.'s (1999) distinction between word-for-word translation and one-to-one translation was also used, as it was found to better describe the degrees of differences in the side of the spectrum closer to literalness. In his categorisation, the former indicates a transfer of SL primary meanings of words, grammar structure and word order into the TL (e.g. *I go to the cinema – Io vado al cinema*), whereas in the latter the SL word or expression has a correspondent in the TL, but these two words primary meanings may differ (e.g. *take* and *fare* in the following: *take an exam – fare un esame* [do an exam]). By primary we mean the meaning of a word taken in isolation, i.e. out of context. A detailed discussion on the translation strategies deployed is given in the translation protocol (3.11), and the two subtitle versions are reported in full in Appendix B.1.

2.5 Formal Similarity and CLI

As we have seen in the previous section, translation will be operationalised as literal and non-literal. Part and parcel with the concept of 'literalness' is that of similarity. Linguistic similarity can be investigated at the higher, language-system level or at the lower, sub-language level. The former distinguishes between language families, with, for example, English and Danish belonging to the Germanic family, while Italian and Portuguese to the Romance family of Indo-European languages. The latter distinguishes between sub-language features, mainly orthographic, semantic and phonological similarities at word- and sentence-level. Research in psycholinguistics and SLA has shown that these similarities can have an effect on language acquisition and processing. For example, it has been found that semantic similarity with the L1 affects L2 verbs learning (Yu, 1996), and the degree of similarity or 'closeness of translation' (Laufer, 2000: 188) between idioms in L1 and L2 determines the avoidance rate of L2 idioms in elicited production. In her critical review, Koda (1996) convincingly argued that the acquisition of L2 word recognition skills is facilitated if the L1 and L2 writing systems share structural similarities. After Koda's paper, several studies have further confirmed that L1 orthographic backgrounds play a role in L2 visual word recognition and lexical processing (Wang et al., 2003; Koda, 1997, 1999; Muljani et al., 1998) as well as L2 word-form learning (Laufer, 1997; Hamada and Koda, 2011). A plethora of research on cognates has also added evidence to this view that language similarities have an influence on the learner. For example, it has been found that bilinguals[6] are faster in reading cognate words[7] than control words (Dijkstra et al., 1999). Through lexical decision and progressive demasking tasks, Dijkstra and colleagues discovered that, while phonological similarity has inhibitory effects on RTs, both orthographic and semantic similarity have facilitating effects, with cognates and interlingual homographs[8] being "identified faster than matched controls because they share lexical and sublexical orthographic representations across languages. This sharing leads to stronger activation of the orthographic representations during recognition and therefore to faster RTs." (ibid.: 511). The researchers also proposed that formal similarity *alone* (i.e. just partial letter overlap) can be sufficient to create these facilitation effects (ibid.: 513), which was confirmed in a study on translation recognition tasks in Spanish-English bilinguals (Talamas et al., 1999). Talamas et al. presented participants with a word in one language followed by a word in the other. Participants had to decide if the second word was the correct translation of the first word, and their RTs were recorded in the process. The translation pairs were manipulated as to have correct translation pairs (*garlic-ajo*), false form-related pairs (*garlic-ojo* [eye]), false semantically related pairs (*garlic-cebolla* [onion]) and false unrelated pairs (*garlic-lana* [wool]). It was found that less fluent learners tend to be more affected by

---

[6] As it often is the case in the psychology and psycholinguistics literature, in this study bilinguals are defined as "native speakers of Dutch who had learned English as a foreign language at school for at least 6 years and used English regularly during their study." (Dijkstra et al., 1999: 501). Thus, the authors adopt the common Grosjean's (1992) definition, which emphasises the regular *use* of languages rather than the *fluency* level attained. In this view, bilinguals are individuals able to express themselves with ease in two or more languages, rather than people who have been exposed to and speak fluently two languages *from birth* (bilingual first language acquisition).

[7] Dijkstra et al. limit the definition of cognates to "translation equivalents with *completely identical* orthographies" (1999: 500, my italics) and shared semantics, such as in the pair *sport-sport* (Italian-English). However, several authors from different research areas (Cop et al., 2015; Talamas et al., 1999; Hall, 2002; De Groot and Nas, 1991; Van Assche et al., 2009, Partington, 1998 amongst others) define cognates more broadly, as "words that are translation equivalents but also show some degree of form overlap" (Cop et al., 2015: 3), thus adopting the more common definition of words with similar – rather than identical – sound and spelling. Consequently, it is in this latter meaning that I use the word throughout this piece of work. 'Translation equivalence' throughout this section is intended in the psycholinguistic sense discussed in 2.4.3.

[8] Dijkstra et al. call 'interlingual homographs' what is commonly known in non-specialist terminology as 'false friends', i.e. words in different languages that share exactly the same orthographic form but without meaning overlap (1999: 497), e.g. *burro*[butter]-*burro*[donkey] (Italian-Spanish). Again, they restrict their definition to *identity* of form, whereas false friends commonly include words with different meaning and *similar* spelling (Laufer, 1989: 12; Partington, 1998), e.g. *libreria*[bookshop]-*library* (Italian-English).

formal similarities (longer RTs) and made more classification errors overall (lower accuracy), but showed less interference from semantic similarity, while the reverse was true for fluent learners, although the situation was more complex for this latter group. Overall, the findings suggest that in the initial stages, learners rely on lexical associations from L2 to L1 and therefore are prone to confusion when encountering orthographically similar words, regardless of whether they share semantics. This situation progressively changes as proficiency increases, with learners becoming more likely to mediate L2 words through concepts, although mediations at the lexical level do not disappear altogether but remain as a form of interlanguage connection (Talamas et al., 1999: 46). Significant L1-L2 overlaps such as cognates also affect the development and organization of the L2 mental lexicon[9] in early stages of vocabulary development (Hall, 2002). According to Hall's Parasitic Strategy, when the learner first encounters an unknown word:

> the key to learning the word is first to establish a form representation, i.e. construct a memory trace of the pronunciation and/or spelling, and then to make the right connections with existing lexical and conceptual knowledge. The strategy claims that after registering the form, learners will immediately identify a translation equivalent, should one be available, through overt translation into L1, by an L1 or L2 definition, by some icon (e.g. a picture or mime), contextual clues or by whatever other medium. (2002: 72).

This is consistent with the view that translation equivalents (such as *garlic-ajo*) "appear to have a different and closer cognitive status than within-language synonyms" (Francis, 2005: 251), perhaps precisely because these immediate lexical associations at the representation level are established upon the first encounter with a new L2 word. Thus, L2 learners economically maximise already established language structures (i.e. entries in the mental lexicon) to connect L2 words to a translation equivalent, and this may be particularly true in a situation like that of interlingual subtitling, where the L1 and its L2 translation are available *simultaneously* and *at all times*. Making use of this previous knowledge (lexicon entries), however, can lead to facilitation effects but also identification errors, such as taking *to second* [to support (a speaker, a proposition) in a

---

[9] For a discussion on the mental lexicon, word representations and other core theoretical concepts related to language processing, see 2.6.3 later in this chapter.

debate[10]] to mean *assecondare*[11] [to please someone in their demands or wishes] in Italian, due to the formal overlap between the items. Laufer (1989) called this phenomenon 'deceptive transparency'. Words are transparent "if they can be interpreted on the basis of intra- and/or inter-lingual clues in the word itself" (Laufer, 1989: 11), and they are deceptively so when these clues seem to assist in inferring the meaning of the word, but in fact they do not. Transparency occurs when a L2 word is formally similar to its translation equivalent in a language familiar to the learner (usually but not necessarily the L1) or when a part of that word is similar to a familiar word in another language and another part bears meaning in the L2 itself (ibid.). For example, the Italian *fatale* will be familiar to anyone acquainted with its translation equivalent *fatal*, and *costruzione* will be recognised as *construction* if one recognises the similarity between the two roots *costr-* and *constr-* and knows that the Italian suffix *-zione* usually corresponds to *-tion* in English. However, even cognate pairs with substantial formal overlap like these two examples may not always be reliable translation equivalents. Transparency can be seen to act along a cline starting from *true cognates*, also sometimes referred to as true friends, i.e. reliable translation equivalents in a large number of contexts, through to *partial cognates*, true translation equivalents in a smaller number of contexts, to *false friends*, where transparency is most deceptive and, in fact, other lexical choices from the same semantic field would be preferable in the translation. Partington (1998) provides a fitting example of deceptive transparency with the pair *correct–corretto* (English–Italian). The formal overlap in this pair is almost total (two letter difference), yet a corpus analysis of the words in question reveals that considering them "excellent friends" (1998: 56), i.e. assuming them to have the same meaning and being used in the same way in all contexts, would be a mistake. Although they can be and sometimes are translation equivalents, frequency and collocation analyses show that *corretto* is relatively more common in Italian than *correct* is in English, and the latter has a rather different pattern of co-occurrence from the former, often appearing in collocations such as *the correct time* or *the correct weight*, a collocational behaviour that *corretto* does not share, with *the correct time* often being translated by *l'ora esatta* [the exact time] and *the correct weight* being translated by *il peso forma* [the 'fitness' weight] (ibid.: 54-55). This partial semantic overlap in deceptive transparency can cause problems to experienced translators, let alone language learners. Transparency does not

[10] OED, 2016.

[11] Sabatini-Coletti, 2016.

emerge only from semantic overlap, however. Laufer states that deceptive transparency can be semantically, phonologically or morphologically motivated, and distinguishes between five categories: words with a deceptive morphological structure, idioms, false friends, words with multiple meanings and synforms (1989: 12). Synforms are pairs or groups of formally similar lexical items in the same language, which can cause some problems during word recognition and learning, e.g. *economic/economical* or *acute/cute* (ibid.: 13). The role played by deceptive transparency is also evidenced in a study on form-meaning mappings by Spada and colleagues, who confirm that "the linguistics features learners are most likely to have long-term difficulty acquiring are those in which there is a misleading similarity between the L1 and the L2" (Spada et al., 2005: 199). Therefore, formal similarity can cause both interference and facilitation depending on a number of factors, amongst which are the task at hand, the type of similarity and the proficiency level of the learner. Among facilitation effects, the well-established cognate effect described above for L2 reading was found to operate even in L1 reading, both in isolated word recognition (Van Hell and Dijkstra, 2002) and sentence contexts (Van Assche et al., 2009). The latter study attested an influence of the weaker second language, learned in adolescence, on the highly automated skill of reading in the learners' native language. Participants read cognate words and control words in sentences while their eye movements were recorded. Data on first fixation durations (the duration of the first fixation on a target word), gaze durations (total fixation durations from the moment the eyes land on a target word until they move away from that word), and regression-path durations (the time period starting with the first encounter with a word and ending when a subsequent word is fixated) was collected (ibid.: 925). Results showed that the durations of all three reading-time measures analysed decreased significantly as orthographic overlap increased. Thus, the presence of a cognate effect in L1 sentence processing strongly suggests that representations of a non-dominant L2 are activated even when L2 cognate words are not relevant for L1 text comprehension, and this influence is strong enough to change L1 reading by affecting L1 word recognition (ibid.: 926).

Collectively, the phenomena described above have been referred to as Cross-Linguistic Influence (CLI), language transfer, linguistic interference, native language influence, and language mixing, among others (Odlin, 2003). In fact, Odlin defined transfer as "the influence resulting from similarities and differences between the target language and any other language that has been previously (and perhaps imperfectly) acquired" (1989: 27). Positive transfer then arises in situations where formal transparency would facilitate acquisition, whereas negative transfer occurs when deceptive transparency interferes with the acquisition process. Since the word 'transfer', as intended by TS scholar Gottlieb (1992), already is a central linchpin for this thesis (see 2.4.3), in order to avoid terminological confusion I will use the term CLI to refer to the phenomena addressed in this section. Although in SLA there has been some controversy around this topic, stemming from its association with unpopular behaviourist theories of habit-formation and contrastive analysis, there is little doubt amongst researchers that CLI is real and that "when a highly similar language is the target, the native language can greatly facilitate acquisition" (Odlin, 2003: 478). CLI occurs in virtually all language sub-systems (Kellerman and Sharwood Smith, 1986; Hamada and Koda, 2011), including pragmatics and rhetoric, semantics, syntax, morphology, phonology, phonetics and orthography (Odlin, 2003: 437).

Based on the above review, like Laufer (2000) and many others, in this investigation I start from the assumption that CLI plays a role in FLL, and may therefore have an impact on the processing of subtitles, which constitutes a special case of reading. From Laufer I also take the definition of formal similarity and difference: "Formal similarity occurs when two languages use the same linguistic elements to express the intended meaning, while formal difference occurs when the meaning is expressed by different linguistic means" (2000: 187). Therefore, formal similarity has a broad meaning that allows for symmetries and correspondences between L1-L2 syntactical structures to be included in the discussion. As outlined above, on the other hand, formal transparency has a more specific meaning, referring to word orthography and L1-L2 similarities at sub-word level specifically. Although these orthographic similarities will also be considered in this study, the term 'transparency' as it has been used in the literature is too narrow to account for the whole range of translational phenomena addressed herein. Moreover, in the psycholinguistic literature, it is mostly semantic transparency that has been examined, in particular in word compounding research. To avoid confusion with other areas of investigation and ensure an appropriate choice of terminology that can fully cover the translational phenomena analysed in 3.11, I therefore prefer the broader term 'formal similarity' over 'formal transparency'.

In AVT, CLI and the effects of formal similarity have virtually never been tackled directly. However, scholars have occasionally observed it in the discussion of their findings. For example, Van de Poel and d'Ydewalle recognise CLI at the higher language system level when they state that "similarity with the native language may cause interference in acquiring a foreign language; however, the similarity could also facilitate the acquisition process." (2001: 271). In their comparative study on standard and reverse subtitles, they use two foreign languages, French and Danish, with Dutch L1 participants. Danish, like Dutch, is a more closely related Germanic language, while French is more distant, being a Romance language (ibid.: 262). Aside from finding that reverse subtitling consistently resulted in more vocabulary acquisition in adult learners comparted to standard subtitling (ibid.: 272), Van de Poel and d'Ydewalle obtained acquisition effects typically stronger for Danish than French, and concluded that similarity at the language system level can have facilitation effects, as shown by their vocabulary and syntax acquisition tests.

In their study on the effects of multimodality[12] on L2 learning, Guichon and McLornan (2008) found L1-L2 interference at the lexical level. They tested French native undergraduate students (with an intermediate level of L2 English) for comprehension of an English authentic video where information was presented through different channels, including two conditions with subtitles, in the L1 (standard) and L2 (bimodal) respectively. The participants took notes in L1 or L2 during watching and then produced a small written summary of the video content in the L2. Reading L1 French subtitles while watching an English L2 clip affected their terminology use in written production. For example, the term *self-employed* (heard aurally) was rendered in the L1 subtitles as *indépendant*, the French translation equivalent. Half the subjects in the standard subtitling condition used the English word *independent* in their summaries, *de facto* ignoring the L2 target heard during viewing. Moreover, in the analysis of the students' notes, the authors note the presence of numerous Gallicisms, e.g. *propriety* for owners, with influence from the French *propriétaires*, or *vacancy* for holiday, with influence from *vacances*, and this was particularly true when the students were exposed to standard subtitles. Guichon and McLornan conclude that "it seems quite evident that interference between French and English impaired lexical accuracy as students concentrated on L1 subtitles rather than on the oral message in the L2" (2008: 7).

A situation similar to Guichon and McLornan's for L2 production occurred in Ghia's (2012a) investigation for L2 reception (reviewed in detail in section 2.7.3 in this chapter), where she looked at translation effects on learners' L2 memorization within a standard subtitling condition (English L2 audio, Italian L1 subtitles). The post-test consisted of a

---

[12] In their study, multimodality was defined as a condition that "[m]akes sensory information accessible in diverse semiotic codes and offers the opportunity to comprehend information through different channels" (Guichon and McLornan, 2008: 2).

MCQ on verbatim recognition of L2 wordings as heard in the foreign audio. Ghia found that, in most cases of misrecognition, learners chose the back-translation of the expressions appearing in the subtitles, again suggesting that L1 interference occurred, resulting in higher focus on L1 input to the detriment of L2 mnemonic accuracy.

As the review above clearly shows, previous psycholinguistic research on CLI has concentrated on vocabulary, either in out-of-context lists or, at best, in isolated sentence contexts. As far as syntax is concerned, research seems rather sparse, yet some studies have attested CLI in this sub-system. For instance, Hall and Schultz (1994) analysed written productions by Mexican natives learning L2 English at beginner level, and found that the majority (74%) of syntactic errors were explained by the behaviour of Spanish structure equivalents. Some CLI effects for syntax have also been found in cleft sentences (Flippula, 1999) and in the use of prepositions in written production (Jarvis and Odlin, 2000). In an AVT scenario like that of the present study, where the mnemonic effects of literal and non-literal translations are compared, would having both source and target available at all times affect CLI effects on individual words? And would the pattern of results for vocabulary hold for syntax as well? It could be postulated that this scenario either (a) causes more interference (resulting in less accurate recall) with non-literal translations, since the potentially confusing formal difference between L1-L2 is constantly highlighted by having both languages simultaneously on screen; or (b) clarifies the 'non-standard' L1-L2 correspondence in non-literal translations, resulting in equal recall performance between the two conditions. In Ghia's (2012a) above-mentioned study, literal and non-literal L1 subtitles were compared, with findings revealing that literal (formally similar) subtitles were recalled consistently more accurately than non-literal ones, and this happened regardless of linguistic category (lexicon or syntax). Given that CLI has been shown to occur in L2 memorization and learning as well (Yu, 1996; Laufer, 1997; Hamada and Koda, 2011), the mnemonic advantages in Ghia's findings seem to support (a), namely a situation where non-literal renderings create inhibition effects, while literal ones facilitate recall due to their similarities with the source, both for lexical and syntactic items. In Ghia's study, literal translations involved no main ST-TT changes, with word order and vocabulary preserved between audio and subtitles, often resulting in word-for-word renderings, while diverging, non-literal translations involved lexical and syntactic substitutions, pragmatic substitutions and reductions in the L1 subtitles (2012a: 78). If, as we have seen, translation equivalence in psycholinguistic terms has a closer cognitive status in the learner's mind, and if L2 learners use existing entries in the mental lexicon (2.6.3) to connect L2 words to a translation equivalent via lexical links, it makes sense that when translations deviate "from what could have been perceived as the 'norm' (i.e. literal translation)" (ibid.: 82) this correspondence between translation pairs is violated and the cognitive perception of this mismatch results in inhibitory effects on recall. This situation may also be enhanced when L1-L2 'norm-defying' translations can only be focused on for as long as the subtitle remains on screen (typically no more than 6 seconds), which may be not long enough for learners to register and internalise the novel L1-L2 correspondence to an extent that allows retrieval after just one exposure. Moreover, CLI effects could be particularly evident in cases where it is not enough to recall the gist or content of what was said in the L2, but memory for exact phrasing is called upon. Verbatim recognition memory requires to retrieve the word form as it appeared during viewing. A situation where both SL and TL appear simultaneously on screen (fostering L1-L2 comparisons and increased mapping) appropriately lends itself to assess which combination of semiotic channel (aural, written), language (L1, L2) and type of translation (formally similar or diverging) causes CLI facilitation or inhibition effects. For these reasons, a verbatim memory test equivalent to Ghia's MCQ was chosen as the testing procedure to assess whether formal similarity in reverse translation has an effect on learning.

2.6 Reading Research

2.6.1 The Mechanics of Reading

Subtitle processing is basically a form of reading, albeit a very peculiar one. Therefore, addressing how *static* texts are read is clearly relevant to information processing in subtitling, and enables a more informed discussion of how *dynamic* texts – such as audiovisual texts – are read. Reading is a complex cognitive process (Koda, 1996: 450) and, some argue, "the most important and ubiquitous skill that people acquire for which they were not biologically programmed" (Reichle et al., 1998: 125). When reading under normal circumstances, the eyes do not fall on every word, let alone every letter. They move through quick, jerk-like, ballistic movements called *saccades*, which happen several times per second, last approximately 25-60ms and move the eyes around 7-9 character spaces (Schotter and Rayner, 2012). Between saccades, the eyes stay relatively still[13] for short periods called *fixations*, which last around 200-250ms on average (Rayner and Pollatsek, 1989). Saccadic movements are necessary to redirect the gaze to a new place in order to identify the visual stimulus to be processed, but it is mainly during fixations that the information is extracted from the stimulus (in the case of reading, word forms). During saccades, sensitivity to visual input is diminished (a phenomenon called 'saccadic suppression') and no meaningful information is gathered (Cop et al., 2015). Many words receive more than one fixation, while many – approximately a third of the total – are skipped altogether, especially if they are very common, highly predictable or if they are function words (Schotter and Rayner, 2012: 87). The reason we move our eyes in this fashion while reading is because vision has the most acuity (resolution) in the fovea, an area of around 2 degrees of visual angle in diameter, found in the centre of the retina (ibid.: 84). In the parafoveal area and further out in the so-called periphery, visual acuity drops sharply, so in order to best gather information during reading, the fovea must be repositioned over the words to be processed. Fixations and saccades are not the only eye movements made by the oculo-motor system, however. *Smooth pursuits* are much slower, voluntary movements of the eyes designed to maintain a moving stimulus on the fovea (Purves et al., 2001). *Regressions* can be defined as backward movements where the eye goes back to a previously read or skipped part of the text (Schotter and Rayner, 2012: 87), they move along a horizontal axis (Ghia, 2012a) and typically indicate integration difficulty (Rayner, 1998; Cop et al., 2015). In the context of subtitling specifically, *attentional shifts* are back-and-forth movements occurring between visual scene and subtitles, proceed along a vertical axis and involve regressions and second-pass fixations (Ghia, 2012a: 80). Second-pass reading "includes all fixations after a regressive eye movement on those parts of the text that the eye had already passed during the first pass" (Wotschack, 2009: 5). Being concerned with processing in an acquisitional perspective, the present investigation will consider fixation measures (specifically, duration and count), since this is where linguistic information is extracted from surface word form and submitted to further mental elaboration processes that might lead to subsequent retention. To investigate these eye movements in subtitle reading, the present work

[13] Even during fixations, the eyes are not perfectly still: miniature eye movements such as flicks, drifts and tremors (nystagmus) happen in order to constantly stimulate the fovea and prevent the fixated image from fading (Steinman et al. 1973; Rayner, 1998).

exploits eye-tracking. For a specific account of the use of this technology, the reader is referred to section 2.10.

2.6.2 Assumptions and Core Findings

The main underlying assumption in most reading research derives from what Just and Carpenter (1980) called 'the eye-mind hypothesis': "the eye remains fixated on a word as long as the word is being processed. So the time it takes to process a newly fixated word is directly indicated by the gaze duration" (ibid.: 330), and spending more time on a word indicates the word requires more processing. This is the current view in psycholinguistics, where eye movements are considered indicative of moment-to-moment visual and cognitive processing (Rayner, 1998; Ghia, 2012a: 73) and fixation durations are "a marker of the ease of accessing the meaning of a word and integrating this into the current sentence." (Cop et al., 2015: 2). Although some processing can happen in the parafoveal and peripheral view (Schotter et al., 2012) and attention can be dissociated from gaze position under certain circumstances (Posner, 1980), for the most part the position of the eyes reflects where the mind is attending to (Schotter and Rayner, 2012: 85), and attention and eye movements are tightly coupled in virtually all tasks (Deubel and Schneider, 1996). Therefore, in the current study, the presence of eye fixations to subtitled input will be considered a proxy measure of attention allocation (for a detailed discussion of attention and noticing see 2.7).

Although foveal vision is where most of the visual information (be it print or images) comes from, some information is extracted from the parafovea: paravoveal processing can indeed facilitate foveal processing, as shown by research on the *perceptual span* (McConkie and Rayner, 1975). This is a window around the point of fixation, typically spanning 3-4 character letters to its left and 14-15 to its right (in left-to-right writing systems), which determines the area where information can be accessed at any one point during reading. Moreover, the area where words can be accurately identified from a fixation location (*word identification span*) is even smaller, around 7-8 characters to the right of the fixation (Rayner, 1998: 380). Experiments using a series of gaze-contingent display paradigms have shown that very little information is gathered outside the perceptual span. In these paradigms, the reader is free to move the eyes wherever they wish, but the amount of information available at any one time is manipulated experimentally (ibid.: 379). In fact, if two words to the right of the fixated word are available for the eyes to see, but text beyond this area is somewhat perturbed (usually by replacing the letters of the words with Xs or other letters), the subjects whose eye movements are monitored do not even recognise that something is odd about the text (i.e., that only a few characters at the time are visible) and reading proceeds in a normal fashion (Schotter and Rayner, 2012: 89). Text within the perceptual span, on the other hand, affords what is called *preview benefit*, namely a processing advantage resulting from having a valid preview of the upcoming word. That is to say, if word *n+1*, to the right of word *n* (the item currently being fixated) is available *before* the eyes fixate it, this leads to fixations around 30-50ms shorter on *n+1* (ibid.: 90). Interestingly, and rather relevantly to our previous discussion of formal similarity, researchers believe that the speeded processing obtained through the preview is mainly due to orthographic and phonological effects (ibid.). Words that come after the fixated word (parafoveal stimuli) therefore can affect reading by shortening fixation durations.

The average reading speed for adult, fluent reading is 200-350 words per minute (Hulstijn, 2001: 264), which roughly corresponds to 5 words per second. However, reading time varies considerably depending on a multitude of linguistic, contextual and personal factors. Amongst other variables, eye movements are affected by word class (Ghia, 2012b; Godfroid et al., 2010), polysemy or lexical ambiguity (Duffy et al., 1988; Clifton et al., 2007), concreteness (Godfroid et al., 2010), predictability (Ehrlich and Rayner, 1981; Balota et al., 1985; Rayner and Well, 1996; Staub and Rayner, 2007), sematic relations with prior words in a sentence (Morris 1994), pragmatic features such as the link between word semantics and co-text (Reichle et al., 2007), syntactic complexity (Clifton et al. 2007; Just and Carpenter, 1980), contextual diversity, i.e. how many different contexts a word appears in (Norris, 2009; Adelman et al., 2006), familiarity (Williams and Morris, 2004; Chaffin et al., 2001), age of acquisition of a word (Juhasz and Rayner, 2003; Clifton et al., 2007), reader's education (Just and Carpenter, 1980), and reading skills (Rayner, 1986).

Two further particularly crucial factors that affect fixation and overall reading times are frequency and word length. One of the most evident findings is that short words are much more likely to be skipped than longer words, especially if they are function words (Schotter and Rayner, 2012; Cop et al., 2015). Word length also affects parafoveal preview (Pollatsek, et al., 2008: 3), resulting in 2-3-letter words being fixated roughly 25% of the times, while longer words (8 letters or more) are virtually always fixated (Rayner, 1998: 375). Moreover, frequent words are fixated less while rare lexical items are fixated for longer (Pollatsek et al., 2008: 2; Rayner et al., 1989). Another striking finding is that gaze duration on a noun tends to be longer when preceded by a low-frequency adjective as opposed to a high-frequency one, a phenomenon called 'spill-over effect' (Rayner and Duffy, 1986; Pollatsek et al., 2008), which demonstrates how local sentence context also plays a role in information processing.

Therefore, a reader's viewing pattern and reading rate are characterised by a complex interplay of variables. Amongst these, however, frequency effects on reading stand out as one of the most discussed in the literature. Given the crucial role that this variable has been deemed to play both in reading and acquisition, frequency was included as a control variable in the present experiment, and is addressed in more detail both in 2.11 and 3.14.

2.6.3 The Architecture of the Mental Lexicon

However, reading is not just a matter of surface perception, but of integrating information in the current words and sentence into wider units of meaning, in order to achieve overall input comprehension. How do we internalise language information from reading? The tools currently used to investigate these complex cognitive processes, e.g. eye-tracking, can only provide an indirect insight on the type of cognitive processing that ensues from reading (Smith, 2012), so the answer to this question ultimately depends on what assumptions are made about the language faculty. What follows is a brief introduction to these main assumptions, where the key terminology discussed comprises the mental lexicon, representation, access, encoding, activation and word recognition. There is much debate in the literature about these issues, and different currents of thought exist. This is a general introduction, with the sole aim of allowing the reader to, firstly, understand the terminology used at various points in this thesis, and, secondly, appreciate some of the underlying assumptions of the research on processing, reading, formal similarity and acquisition presented in this literature review.

In psycholinguistics, the existence of a *mental lexicon* is posited for both monolinguals and bilinguals. This is a large mental database where we store knowledge of words (Dijkstra, 2007: 252), including phonological and orthographic information, and from which we retrieve words during comprehension and production. It has been proposed that monolinguals possess a mental lexicon of roughly 50,000 words, from which information can be retrieved as fast as in 1/3 of a second (ibid.). Bilinguals[14], therefore, must have

[14] For a definition of 'bilingual', see note 6 in section 2.5.

thousands more words in store, and although they have been found to recognise words and make lexical decisions slower than monolinguals (Ransdell and Fischler, 1987), the cognitive effort required to access and use these lexical entries still seems overall rather small, given the ease with which bilinguals handle the two language codes on a daily basis. Words must be *represented* in the mental lexicon with respect to spelling, sound, meaning as well as other characteristics, e.g. morphology, pragmatics and language membership. Thus, the mental lexicon is usually thought of as "a multidimensional storage space set up along these dimensions" (Dijkstra, 2007: 252) and contains representations, i.e. "knowledge structures in the mind of the learner" (Harrington, 2001: 99), which can be lexical and conceptual. The former comprise linguistic constructions and the latter non-linguistic, conceptual knowledge about the world and common human experiences (Seilhamer, 2010). During reading, in order for the representations to be accessed, the words need to be somehow 'perceived', the first step in this process being the encounter with a word and the activation of its orthographic form (Dijkstra, 2007: 252). Thus, *word recognition* during reading primarily proceeds from orthographic information. Phonological representations also emerge during the process, but they are not always utilised in lexical access (Taft, 1993: 91, in Hulstijn, 2001). Thus, "[w]ord recognition takes place in an interactive way via the activation of sub-lexical units ranging from (components of) letters and morphemes" (Hulstijn, 2001: 265). Upon print information perception, words are recognised and their form is believed to be mapped onto their corresponding representation in the mental lexicon. This process is called *encoding*. Thus, "encoding refers to the transformation of linguistic information at one stage of processing for use in the next" (Doughty, 2001: 211). The mental lexicon must therefore be 'entered', in order to make these connections between word forms and lexical representations. The process "by which the basic sound-meaning connections of language, i.e., lexical entries, are activated" (Pylkkänen, 2016) is called *lexical access*, and *activation* is seen almost as a sort of electric impulse connecting the entries in the mental lexicon (Field, 2003: 16). Upon encountering a word, a reaction is triggered in the mind of the reader/listener, whereby other words become highlighted (that is to say, activated). For example, when coming across the word *teacher*, words with a semantic (e.g. *learning*, *classroom*) or formal (e.g. *preacher*, *teaching*) connection to *teacher* become activated in such a way that can be tested through priming and affects RTs in various experimental tasks. Activation is automatic, i.e. it cannot be turned on or off, and decays quite fast, i.e. words do not remain activated for long (ibid.).

2.6.4 Reading and Incidental Language Acquisition

Knowing about how words are stored and accessed in the mental lexicon is relevant also because it strives to describe what further elaboration processes happen in the reader's mind during reading that might lead to subsequent retention. Language can be acquired through reading both in monolingual (Nagy et al., 1985) and bilingual circumstances (Pitts et al., 1989). In fact, many believe that a large amount of vocabulary is learnt incidentally through silent reading (e.g. Krashen, 1989; DeKeyser, 1998), and this is in line with the research on formal similarity in word recognition and learning reviewed in section 2.5. However, more recently, researchers have argued that "incidental uptake of vocabulary through extensive reading alone is much slower a process than was once believed" (Godfroid et al., 2010: 176; see also Laufer, 2005), where uptake refers to stimuli that are committed to long-term memory (Godfroid et al., 2010: 170). Vocabulary acquisition is an incremental process (Nation, 2001), and "since it is the quality and the frequency of the way in which new words are processed that determine their acquisition, it may be too simple to conclude that the only thing students need is extensive reading" (Hulstijn, 2001: 272). Multiple encounters in a relatively short amount of time are needed for a word to start leaving a trace in memory, with studies reporting that two to four encounters of unknown words led to roughly 60% accuracy in identification of word meaning (Rott, 1999, in Hamada and Koda, 2011). On the other hand, as Godfroid et al. remind us, "we cannot deny the possibility that some memory trace, however feeble, is created on a first encounter with a word" (2010: 176). Therefore, it can be postulated that some incidental learning through L2 subtitle reading may occur during processing of filmic material, especially if one considers the lower affective filter (Krashen, 1985) and generally high engagement of this activity, which may contribute to retention of word forms after just one exposure. Generally speaking, AVT research is in line with the view that vocabulary can be learnt through reading in multimodal contexts in just one exposure, since, as we have seen in 2.3, vocabulary gains have been demonstrated through reading L2 subtitles, both reverse (Danan, 1992; Holobow et al. 1984; Čepon, 2011) and bimodal input (Vanderplank, 1988; Neuman and Koskinen, 1990; Bird and Williams, 2002). Hulstijn also recognises that attention plays a role in this process, and argues that L2 acquisition during reading "will be substantially enhanced when learners' attention is oriented towards unfamiliar words, i.e. when the meaning of unfamiliar words is given by means of marginal glosses or has to be looked up in a dictionary (...)" (2001: 274). One could argue that reading reverse subtitles in learning contexts may present a situation where the L1 audio itself acts as a series of constant clarifying glosses or parallel dictionary entries, providing the basis for attention to be directed towards unknown or unusual words, and for noticing to occur. The present study considers attention and noticing processes specifically, to assess whether increased, clear L1-L2 mappings (see 2.4.2) through reading subtitles support the establishment or reinforcement of mental traces between translation equivalents in the learner's mind, potentially leading to incidental mnemonic gains.

As a complex skill, reading entails several components. These components are usually categorised into two broad areas: lower-level and higher-level processes (Fender, 2001). According to Yamashita, at the lower-level, basic linguistic processes are involved, such as "letter identification, word recognition, syntactic parsing, and proposition encoding" (2013: 52), while the higher-level comprises (meta)cognitive processes such as "integrating information within a text, activating and utilizing background knowledge in text meaning construction, making elaborative inferences, monitoring comprehension, and strategic processing." (ibid.). In lower-level sentence processing, readers parse (i.e. analyse) a text automatically, proceed by chunking (i.e. by grouping words into syntactic units) and recognize particular syntactic configurations during online comprehension (Perego, 2008: 213). Researchers seem to concur on the fact that lower-level skills rely on more inflexible, automatized processes, whereas higher-level skills entail more flexible, adjustable processing (Hulstijn, 2001: 266) and generally require attentional resources (Grabe, 2009). As we have seen in 2.5, a wealth of studies have been investigating lower-level skills in vocabulary processing (for a review, see Koda, 1996). Research on higher level skills is, on the other hand, still scarce, especially involving acquisition through the natural reading of long passages. In fact, what claims to be the first study to investigate incidental learning of grammatical features from text reading was published very recently (Bordag et al., 2016). The study looked at two grammar properties of verbs (subcategorization and [ir]regularity), and found that contextual syntactic complexity had a positive influence on the incidental acquisition of new L2 words, triggering shifts in learner attention from lower word-level to higher text-level. Aside from this example, however, much L2 reading research investigated processing at word level, and the few studies that looked at sentences in fact still focused on the recognition of individual words embedded in a sentence context (Cop et al., 2015: 2). Intuitively, in terms of FLL, reading individual word lists differs from reading words in isolated sentences, which in turn differs from reading sentences in context. In this last case, a reader has to assess how the words fit into sentences, build a mental representation of the syntactic relationships as well as integrate them in the wider text context. Research shows that all this happens rather quickly, often while the gaze is fixated on a word (Clifton and Staub, 2011: 897). The present study contributes to filling this gap by addressing subtitle processing, noticing and memorisation at the level of both words and sentences, and does so through the use of non-intrusive eye-tracking technology, reproducing natural subtitle reading conditions as closely as possible, in order to observe whether learning opportunities arise from such reading conditions. Moreover, since formal similarity is assessed not only through a recognition memory post-test but also an open questionnaire (see 3.15), the study has more potential to retrospectively gather an insight from the learners themselves about the higher-level processes (potentially leading to retention) they may have been engaging with while watching the subtitled video, such as making elaborative inferences or using previous knowledge whilst processing vocabulary and syntax. Having looked at the common characteristics between subtitle processing and more 'traditional' print information processing, the next section looks at how the two differ.

2.7 AVT Processing and Reception

2.7.1 Reading Print and Dynamic Texts

Reading interlingual subtitles is a more complex activity than monolingual, traditional reading for a number of reasons. First of all, subtitle processing is not self-paced, since the viewer does not have full control of *where* and *when* to move the eyes but has to follow the subtitles at the pace imposed by the video itself (Schotter and Rayner, 2012: 83). Secondly, unlike traditional reading, where the reader also has more overt control on *whether* to read, AVT research has demonstrated that subtitles are automatically processed (d'Ydewalle et al., 1987), and this is regardless of whether the viewer is familiar with subtitling, knows the foreign language and has the soundtrack available (d'Ydewalle and Poel, 1999; Verfaille and d'Ydewalle, 1987; d'Ydewalle et al., 1989; d'Ydewalle and Gielen, 1992). This automatic and compulsory processing was shown to occur already with children of grade 4 and 5 (d'Ydewalle and Van Rensbergen, 1989) and was deemed to result from the prevalence of visual over oral input (d'Ydewalle et al., 1991). Thirdly, during the consumption of subtitled video, *verbal* content is received through two modalities at once: the audio (dialogues, monologues, songs, etc.) and the printed text (subtitles as well as inserts, hoardings, etc.). Fourthly, whilst simultaneously processing the auditory input, viewers constantly move their gaze back and forth between visual and written input, thus engaging in constant, dynamic switching mechanisms between images and subtitles (d'Ydewalle et al., 1987; Schotter and Rayner, 2012: 83-84; Perego, 2008: 213), which changes the way audiovisual information is processed compared to a situation without subtitles, and creates what Perego calls "a particularly complex perceptual situation" (2008: 213).

Some have proposed that, because the eyes have to move not just along the words in the subtitle sentences, but also between subtitles and visuals, there are higher attentional demands as content needs to be extracted from both channels (Hefer, 2013: 25). A similar point is outlined by Guichon and McLornan, who explained that "reading subtitles adds an extra operation to the comprehension task and creates potential interference due to the constraints upon human information processing capacity" (2008: 3). These arguments link subtitle processing to the concept of cognitive load, whereby written information is thought to compete for the learner's attention rather than complement it. Although it is commonly accepted that attention is limited resource (2.8), said limit does not automatically imply that viewers will concentrate exclusively on the subtitles during AV processing, nor that the cognitive system will be overloaded. In fact, in a study on the distribution of attention during watching subtitled television, d'Ydewalle and Gielen found that the allocation of attention occurs effortlessly and does not necessarily overtax the cognitive capacity of the viewers, who seem to apply successful, adaptive strategies to process both subtitle and images. Reading the subtitles had a major place in the overall distribution of attention, but it did not prevent subjects from attending to the images (1992: 425). This point was taken up again in an eye-tracking study by Perego et al. (2010), who further clarified the issue through both subtitle recognition and scene recognition post-viewing tasks. If a trade-off between image and subtitle processing exists, there should be a negative correlation between the two: the higher the recognition rate for subtitle content, the lower the recognition for the corresponding scenes. The researchers found "good levels of performance both in subtitle recognition and scene recognition" and a "remarkable performance in the scene recognition task observed, even if less than 40% of the fixation time was devoted to visual scenes." (2010: 262). The results thus admit to a decreased overall visual processing, but do not support the proposition that the dynamic switching of attention between channels overloads the cognitive system. They provide evidence towards the view that visual processing and reading are highly efficient and automated cognitive activities (Hulstijn, 2001; Zhou, 2004) and confirm that the integration of two visual information sources appears to be more cognitively effective than it could be *prima facie* argued (Perego et al., 2010: 247). It must be noted that the study looked at one genre only (drama), so it is not representative of the whole audiovisual landscape, and it is quite possible that with different types of audiovisual texts a trade-off would be present. The cognitive effectiveness of subtitle processing is therefore not a given, but depends on genre, audiovisual text complexity, as well as the level of redundancy of information between the various sources. As Perego and colleagues themselves admit (2010: 265), in cases of dense, non-redundant input, for example, cognitive overload could indeed occur. Moreover, in FLL contexts, L2 learner proficiency is also bound to play a role. However, research has demonstrated that when information between the channels *is* redundant – at least with L2 bimodal input subtitles – receiving information through multiple channels has a facilitative effects on learner's auditory processing (Baltova, 1999), text comprehension (Guichon and McLornan, 2008) as well as memory retention, both implicit and explicit (Birds and Williams, 2002). Therefore, the statement that subtitles create cognitive overload and interference with processing cannot be accepted at face value but needs to be further qualified. This is all the more so when one considers that, in certain circumstances, not only do subtitles *not* interfere with processing but, in fact, they can enable learners to clarify instances of semantic and lexical ambiguity (Guichon McLornan, 2008: 3). This is in line with psycholinguistic research showing that translation can provide cognitive advantages during processing (2.4.2) and with the special cognitive status of translation equivalents in the learner's mind (2.5), which might be linked to these processing advantages.

A fifth reason why subtitles differ from traditional print text in terms of reading, is that while the latter usually occupies a relatively large portion of the support it appears on (e.g. a sheet of paper), the former occupy a particular spatial position on the screen, i.e. usually two lines maximum, centred at the bottom of the screen (Lavaur and Bairstow, 2011: 457). Finally, although certain parallels can certainly be drawn between the two reading modes, subtitle processing research has shown that fixation durations in the former trend to be shorter than the latter. D'Ydewalle et al. (1985) report mean fixation durations of 124ms and d'Ydewalle and De Bruycker (2007) of 178ms for L1 standard subtitles (185ms for L2 reverse), all appreciably lower values than the average 200-250ms found in standard silent reading (see 2.6). Given these findings, I also expect to obtain an average fixation duration somewhat shorter than 250ms, although the extremely high individual variability recorded in eye movements (Rayner, 1998: 376), coupled with the translation manipulation applied in this study – designed to have the potential to trigger noticing and attract more visual attention – may increase the average values.

2.7.2 Processing Reverse Subtitles

Scholars from the Belgian School were among the first to investigate subtitle processing through eye-tracking, and some of their studies included reverse subtitles. Pavakanun (1992; in d'Ydewalle and De Bruycker, 2007) recorded eye movements of adults watching movies with either standard or reversed subtitling, and found longer latencies (the time it takes a reader to make the first fixation on a subtitle since its appearance on screen) with reverse subtitles. Moreover, the standard condition resulted in more time spent looking at the subtitles compared to the reverse condition. However, she also found that the reversed condition resulted in at least 40% subtitle presentation time spent looking at the L2 subtitles, suggesting that attention was nevertheless allocated to these subtitles.

In d'Ydewalle and De Bruycker (2007), reading patterns of children and adults watching one- and two-line, standard (L1 Dutch) and reverse (L2 Swedish) subtitles were compared with respect to several fixation measures (fixation duration, word-fixation probability), saccade measures (saccade amplitude, percentage of regressions, shifts between image and subtitle) and other attention allocation variables (percentage of skipped subtitles, latencies, percentage time spent on the subtitles). Little overall difference in reading between children and adults was found, and reverse subtitles displayed an irregular pattern of reading, leading to fewer fixations and significantly fewer shifts, significantly less time spent in the subtitle area, significantly more subtitle skipping and significantly longer latencies, which, according to the authors, "suggested that the participants only occasionally grasped some keywords in two-line reversed subtitles, without really reading the sentences as a whole." (2007: 203).

Unfortunately, a total sample of twenty participants means only ten subjects per subtitling condition. Moreover, since twelve were adults and eight were children, an even smaller number of participants was considered in relation to the multiple variables analysed in the study (three variables: age, number of lines and subtitling condition, with two levels each, yielding a 2x2x2 design). If, as d'Ydewalle and De Bruycker did in their analyses, one is interested in looking at variables together, i.e. testing for interactions, then the numbers in the sub-groups considered in each analysis become very small. For example, although the total number of subjects in the study is twenty, comparing reading performance of children in the reverse and standard conditions results in comparing means of sub-groups of four subjects each, too small a sample to warrant any definitive conclusions. Moreover, there is no mention of whether two-line subtitles were segmented and distributed equally between subtitling conditions. Nor is it clear whether the cueing was left the same in the two subtitling conditions. The authors stated that they followed the six-second rule and that "shorter subtitles are timed proportionally" (2007: 198). This means that each of the two subtitles for the same dialogue line can still display differences in presentation time and in whether they cut across shot changes, so it is difficult to know if the attention allocation differences reported are genuinely due to subtitle condition or stem from asymmetries in subtitle cueing, e.g. differential required reading time.

Despite these methodological blemishes, however, this study remains one of the most thorough investigations of reverse subtitle processing to date, and managed to further demonstrate earlier findings on automatic reading behaviour. It also highlighted the flexibility of the human cognitive system, which was found to be fully capable of allocating attentional resources to multiple sources of information (visual, auditory and textual), thus providing support to the cognitive effectiveness of subtitles view (Perego at al., 2010).

Findings like increased skipping rates and reduced shifts, fixation number and time spent looking at reverse subtitles seem understandable, since we are comparing a reverse situation with L1 audio, where there is *no immediately urgent need* to look at the L2 subs for meaning, to a standard situation where the L1 subtitles are *the only way of accessing meaning* in the video. A crucial aspect of the study design is that the reverse subtitles were "basically meaningless" (d'Ydewalle and De Bruycker, 2007: 203) for the subjects, since none of the groups knew any Swedish or could identify with certainty which foreign language they encountered in the video. Yet, the subjects did spend time processing them (26% of the subtitle presentation time), and did so despite the availability of the native language in the audio. Since the content of the L2 subtitles was completely inaccessible for the subjects, a lack of motivation to read them seems justifiable. This would explain the decreased time dedicated to the reverse subtitles, an inclination to occasionally skip them altogether, and the differences in latencies and shifts recorded compared to the standard condition. In a situation where viewers are less motivated to read L2 text because the information they provide is meaningless, they will shift attention from the images to the subtitles less often, especially if they can perfectly understand video content through their L1 audio. Yet, the subtitles are eventually (with a slight delay, i.e. longer latencies) looked at due to the well-established automatic reading behaviour described above. A crucial question therefore arises: if the subjects knew Swedish, would this 'irregular' Swedish subtitle reading pattern remain? Would viewers really 'only occasionally grasp keywords' and not really read the L2 sentences as a whole? As we have seen in 2.3, and as the authors themselves underline, FLL tests on reverse subtitles in previous studies with language learners led to performance significantly above chance level, consistently demonstrating foreign language vocabulary gains stronger than standard subtitling (see also d'Ydewalle and Pavakanun, 1997), which is only possible if the foreign language is processed in a less superficial way than suggested by the results reported above. Therefore, it seems likely that the irregular pattern of reverse subtitle processing may be due at least in part to the substantial inaccessibility (meaninglessness) of the reverse condition in this particular study, where Dutch native speakers had no knowledge of the Swedish reverse subtitles presented. In fact, what this behavioural study – unlike previous ones by the same school – failed to clarify, is the relationship that reading and attentional patterns bear to performance measures of language acquisition (Kruger, 2016: 7). In the authors' words: "An experiment that directly links a detailed analysis of eye movements in reversed subtitles with foreign language acquisition is needed in order to make more conclusive inferences on the linguistic processing of foreign language subtitles in the presence of a native language soundtrack." (d'Ydewalle and De Bruycker, 2007: 204). My research project bridges this gap precisely by considering behavioural and reception measures together, investigating the correlation between physiological characteristics of attention allocation (eye movements) and performance in mnemonic retention of L2 vocabulary and syntax for English L1 learners of L2 Italian. In light of the above discussion, therefore, in the present study on *meaningful* reverse subtitles, some evidence of more regular reading in the number of fixations and their durations is expected, alongside potential vocabulary gains.

2.7.3 What Factors Influence Processing in AVT

In the consumption of an audiovisual text, both linguistic characteristics of the subtitles and dialogue-script interaction can affect perceptual behaviour (Ghia, 2012a). Guillory (1998), for example, looked at a common translation strategy adopted in subtitling, text reduction. Comprehension of L2 French video in a bimodal condition was tested with L2 full-text and key-word subtitles, where the former was a transcription of French while the latter contained only French words which were essential for comprehension. A no captions group served as control. Significant differences in comprehension scores were found, with participants in the full-text group outperforming those in the key-word group, and participants in the key-word group outperforming the no captions group. The overall pattern of results therefore showed that key-word subtitles helped viewers understand the gist of the text. De Linde and Kay (1999) examined word omission in bimodal input in a series of experiments and showed that reading time was higher in strongly reduced subtitles, suggesting that input-specific factors, and in particular lack of correspondence between aural and written L2, slow down reading in same-language subtitling. The authors also found that the flickering effect of shot changes across subtitles, as well as subtitles reporting off-screen speech, led to increased re-reading of the written input and more deflections, suggesting enhanced processing complexity. Subtitle segmentation, layout and position on screen can also affect perceptibility (Ghia, 2012a: 33). The script should "end at naturalistic breaks, that is, ideally, at clause or phrase boundaries" (Perego et al., 2010: 249), and this is the current trend in the creation of professional subtitles. Although Perego et al. (2010) did not find a significant difference between ill-segmented and well-segmented subtitles, they nevertheless report processing the former to be 12ms longer than the latter, which "does seem to indicate at least a slightly higher cognitive load when the subtitles are ill-segmented" (Kruger et al., 2015: online). Moreover, other studies confirm an effect, with unusual segmentation being liable to increase time spent on the subtitle area (d'Ydewalle et al., 1991), and scrolled subtitles requiring longer reading time than subtitles appearing in blocks (Rajendran et al., 2013). Rajendran and colleagues wanted to assess whether text chunking had an effect on the speed of subtitles processing and overall video comprehension. They compared four viewing modalities: no segmentation (i.e. the subtitle area is filled with as much text as possible), word-for-word scrolling (words appeared one by one), subtitles chunked by phrase (phrases appeared one by one, with one line of the subtitle area being filled at a time) and by sentence (sentences appeared one by one). Significant differences did not emerge in comprehension, but were registered for the eye movements. The eye-tracking data shows that word-by-word subtitle appearance in scrolling inflates the number of gaze points and produced significantly more image-subtitle shifts compared to subtitles chunked by phrase or by sentence. Chunking by phrase provided the best relative viewing situation, leading to a more natural and steadier viewing experience (2013: 15-16) as subtitles in this condition provoked the least number of shifts and the smallest percentage of gaze points. In the current study, subtitles appeared in blocks – as it is customary for pre-recorded subtitles and in line with virtually all investigations of subtitling in SLA – and were chunked by phrase. Further details on the segmentation rationale are provided in 3.7. The number of lines also seem to affect processing. D'Ydewalle and De Bruycker (2007) found significant main effects of this variable on subtitle skipping (15% of the subtitles were skipped with one-line, 10% with two-line presentation) and attentional shifts (participants made more shifts when reading two-liners compared to one-liners).

However, there are some contrasting results (see Kruger and Steyn, 2014), which make taking a clear-cut position difficult. In this study, the common practice solution used in the industry was chosen, i.e. the maximum line numbers was limited to two. Since this variable is not of primary interest for the purposes of the experiment, no difference will be made in the analysis between one-liners and two-liners, and more space will be given to other more relevant subtitle-specific variables, such as duration and length (4.3.3 and 4.3.4).

The interaction between soundtrack, video and subtitles also affects general viewing patterns: in a study on video genre viewers were found to process the subtitles in both films and news broadcasts, however reading times were longer and shifts from the video to the script faster with news broadcast (D'Ydewalle and Gielen, 1992). The trend was explained in terms of the higher content density of this genre, requiring greater cognitive resources to be allocated to input due to the complexity and non-redundancy between the visual and aural channels. Moran (2008, 2012) analysed word frequency and cohesion in terms of readability, and found that coherent subtitles containing well-linked sentences and frequently used terminology were more easily processed by the viewers, who took less time to read them even though they were sometimes longer than their low-frequency, low-cohesion counterpart, thus providing evidence that longer and more explicit translations appear to be more easily processed.

Subtitle translation is another factor that influences subtitle consumption (Ghia, 2012a; Perego and Ghia, 2011). In her analysis of translation simplification, Pavesi (2002) attributes translation issues to the L1-L2 mapping operated by the viewers as they access an audiovisual text. Translation is also addressed, although in general terms, in D'Ydewalle and Pavakanun (1997). They looked at how vocabulary acquisition is affected by similar vs. distant language pairs, but similarity was addressed in very broad terms, by looking at kinship between L1 (Dutch) as a whole and other languages, rather than tacking the similar vs. distant dichotomy in individual instances of translated input. Thus the researchers selected programmes in similar (German and South African) and dissimilar languages (Chinese and Russian) and broadcast them with Dutch subtitles (standard condition). Perhaps unsurprisingly, they found that learning rate was higher with similar languages, which suggests that it might be more difficult to learn words when they have to be detected and discriminated in a flow of unfamiliar sounds (Koolstra et al. 1999: 54). This may also be taken as support to the view that formal similarity between the written and spoken words (e.g. words with similar spelling or similar root, such as cognates) facilitates memorisation of the foreign input.

To explore the possibility of language maintenance (i.e. maintaining an already acquired level of proficiency in a given foreign language) through standard subtitles, De Bot et al. (1986) looked at if and how viewers 'use' the foreign language in the audio, i.e. if they focus on L1 subtitles only or L2 speech or both. Participants were shown a news bulletin with L1 subtitles which, in parts, contained information conflicting with the audio. Discrepant translations were thus included and the deviations were divided in (a) phonological, (b) grammatical, (c) lexical, (d) omissions. The researchers were interested in what channel the viewers relied on to answer the memory post-test, and if they noticed the discrepancy between subtitles and speech. Results indicate that speaker orientation occurred, i.e. that the L2 audio was processed and a discrepancy had been noticed. Interestingly, participants said that that they had been aware of the incongruence but could not ultimately tell on the basis of which input they answered the questions (1986: 80). In their paper, only one test item was given as an example of L1-L2 discrepancy, no precise information is given with regard to the process of creating different L2 versions and analysis of the different translation pairs created is not given. Deviation and correspondence between L1 and L2 are mentioned but only in passing, as translation is not the focus of the study. This is the case with most AVT research so far: translation is employed as a tool in the analysis of various aspects of processing but it is not systematically addressed as it does not occupy a central position in the investigation. The only exception to this, to my knowledge, is a recent study by Ghia (2012a), in which eye tracking was used to explore the differences in perception arising from the use of different translation strategies between L2 audio and L1 subtitles (standard mode). Ghia looked at how the ST was simultaneously processed, mapped and compared to the TT by the viewers. She divided input in a condition of literal and non-literal (diverging) translation, within which she identified quantitative (simplification and reduction) and qualitative (substitution of L2 words and syntactic patterns with formally discrepant L1 words and patterns) strategies (Perego and Ghia, 2011: 182). The main theoretical linchpin in this study was perceptual salience in the context of translation. Ghia theorised and described this as *translational salience*, namely "the prominence acquired by linguistic items when input is delivered" (2012a: 52). Thus, translational salience is a type of perceptual salience achieved through input enhancement, where a subject component, i.e. a receiver – in the case of audiovisuals, the viewer – interacts with such delivery (ibid. 53). Translational salience includes a gradable contrast (ibid.: 51) and a deviation from expectations (ibid.: 71). Ghia proceeded to test the assumption that "the greater the formal contrast which is created between ST and TT, the more likely the contrastive elements are to be perceived as perceptually salient" (Ghia, 2012a: 66). The term *input enhancement* therefore refers to the way in which such contrast is created by highlighting formal divergence of input through translation strategies such as reduction and substitution. To investigate the topic, Ghia deemed it essential to look both at a concrete manifestation of contrast (formal discrepancy in the input) and the perception of such contrast by viewers, i.e. their noticing (ibid.: 71). For the purpose of investigating noticing processes, retrospective recall, an explicit report task and eye-tracking methods were used. Ghia was expecting greater divergence to "correlate with different patterns of attention allocation and viewing behaviour, to be manifest in eye-movements, noticing of the discrepancy in translation and greater verbatim recall of vocabulary and syntax" (ibid.: 75). She did indeed find signs of increased visual activity in the non-literal condition, as deflections[15] were significantly higher with diverging translations, both qualitatively and quantitatively, and second-pass fixations were directed to diverging content words, usually involving re-reading of isolated strings rather than the full subtitle (Perego and Ghia, 2011: 184). Therefore, the intensified processing occurring in certain diverging subtitles proves that input was attended to and suggests that some form of discrepancy detection might have been in place, whether conscious or not. Cognitively speaking, it seems thus reasonable to postulate that there is a difference in source-target mapping in the two conditions. According to Tanenhaus (2007), increased mapping generally indicates a higher degree of attention to the verbal dimension of input. One might therefore assume that the extent of mapping is larger in the diverging condition, where visual attention and processing are higher. In Ghia's experiment, however, this difference did not lead to improved memory performance: overall recall scores are higher in the literal condition, and for both subgroups of analysis (verbatim recognition of vocabulary and syntactic patterns respectively). Interestingly, in the majority of recognition memory mistakes, Ghia discovered that learners erroneously picked the back-translation of the Italian L1 subtitles. Moreover, she found that regressions did not vary significantly between the two conditions. Finally, the correlation between presence or absence of explicit report (open-ended questionnaire) and mean deflections scores for the two conditions was also not significant (.34 coefficient). This correlation measure was

[15] What Ghia calls 'deflections' have elsewhere been called shifts (Rajendran et al., 2013; Perego et al., 2010; d'Ydewalle and De Bruycker, 2007).

chosen "to investigate whether any increase in the amount of deflections paralleled any conscious noticing of a translation discrepancy" (Ghia, 2012a: 84). Thus, although visual attention was higher with divergence, this was not necessarily reflected by an overt conscious report of said divergence. In terms of the relation between noticing and recall, some form of noticing might still have occurred during watching, but the memory trace could have been too weak for information to be fully stored in memory and thereafter recalled and reported, or information could have been stored in such a way that did not necessarily enable retrieval. This finding supports my hypothesis (see 2.4.2) that increased mapping may not be enough to trigger successful retrieval *per se*, that in order for multiple memory paths to result in more elaborate (and less prone to be forgotten) connections the input must be clearly and meaningfully linked in both verbal systems (aural L2 and written L1 here). Perhaps one of the two terms of the mapping connection in Ghia's study (presumably the aural L2) was not successfully encoded because not meaningful enough or lacking sense (as a result of not being heard properly), or perhaps because discrepancy caused perceptual confusion and prevented the memory trace from being stored solidly enough to allow for retrieval. As we mentioned earlier, when L2 input appears in a more memorable format (written instead of aural, i.e. reverse condition), the situation might differ: if the process of establishing connections during L1 audio-L2 script comparison results in a clearer, more meaningful mapping, this might be enough to yield correct item recognition during post-tests, or at least to leave a trace resulting in the viewers explicitly reporting discrepancy.

The reception study presented in the following chapters is designed in a similar way to Ghia's noticing experiment and explores translation in light of the above-mentioned issues using a combination of explicit report, recognition scores and eye-tracking in order to investigate the processing mechanisms at play during reversed subtitle consumption specifically. In line with the previous research in AVT and visual attention reviewed so far, we are generally expecting higher number of fixations on non-literal, diverging subtitles, and longer durations on particularly striking or 'difficult' items, since increased visual activity is tightly linked to how easy it is to access the meaning of a word (Cop et al., 2015: 2) and incorporate it in the subtitle reading flow. Despite Ghia's assumptions that input salience through translation divergence would produce higher mnemonic retention, recall mean scores in her study indicate that a more accurate response was produced in the literal transfer condition. Unfortunately, and somewhat surprisingly, Ghia does not carry out a test of mean differences for recall in the literal and non-literal conditions, so it is not possible to know whether the reported high success rate in recalling literal items is above chance. However, she presents the reader with a graphical representation of this difference, which, at least in the vocabulary sub-group, looks wide enough to be able to reach statistical significance (see Ghia, 2012a: 86-87). Based on these results, and on the discussion of how formal similarity affects processing and learning (2.5), I also postulate that item recognition will be higher in the literal transfer condition, i.e. when not only content but also form is similar between the two languages, with possible differences between syntactic and lexical items. The point is clearly not just that input is encoded twice, aurally and visually (Bird and Williams, 2002), as this is also the case with diverging translations. There might therefore be a difference in quality of the mapping resulting from the use of semantically *and* formally similar input. This is also in line with incidental learning in the d'Ydewalle and Pavakanun experiment (1997) being higher with languages similar to Dutch, the native language of the viewers. Deflections are "most likely related to the process of mapping between dialogue and subtitles" (Ghia, 2012a: 81). As Ghia's study demonstrates, even if literal translations are better recalled, deflections can still be higher with diverging subtitles and more second-pass fixations tend to fall on diverging words, because these might be salient enough to catch the viewer's attention during moment-by-moment processing and trigger noticing, regardless of whether they are subsequently recognised correctly in the post-test. However, the salience resulting from this formal divergence might just not be powerful enough to make the input consciously noticed (i.e. reported in the open-ended questionnaire), or the L2 discrepancy might be consciously noticed and reported, yet the specific form of the diverging L2 word or expression not remembered. This would be in line with De Bot et al. (1986) reporting that viewers in their experiment had noticed the incongruence between subtitles and speech, yet could not tell with certainty what was the basis for their answers. In the AVT literature, the word 'noticing' has so far been used in generic terms and has rarely been linked to the relevant debate in SLA. Ghia limits her study to a general notion of noticing, defined as "a cognitive process between the reception of input by learners and its internalisation" (2012a: 14). Noticing thus "refers to the registration of stimuli from input and is closely related to attention: learners register some linguistic stimuli from input and relate them to their prior knowledge of the L2 (Gass, 1997, 2003)" (ibid.). For the purposes of our experiment, I found it crucial to define noticing in more detail and provide at least a short overview of how this concept has been addressed in SLA literature.

2.8 Attention and Noticing

The constructs of attention and noticing can be considered key in cognitive approaches to SLA and visual processing. In the cognitive psychology and SLA literature, the common view is that attention is a limited resource system (Tomlin and Villa, 1994: 187; Truscott, 1998: 105), and therefore that it must be selective (Tomlin and Villa, 1994: 187; Schmidt, 2001: 13; Leow and Bowles, 2005: 180). In other conceptions, attention is linked to the idea of automatic vs. controlled processing (Tomlin and Villa, 1994; Truscott, 1998), where it "represents effortful processing" (Tomlin and Villa, 1994: 187). It is widespread belief that that attention facilitates L2 development (Leow and Bowles, 2005: 183), and some believe that attention is strictly necessary for learning, at least in terms of long-term memory storage of information (Schmidt, 2001: 16). Schmidt also provides examples from the literature showing the widespread belief that attention controls access to consciousness, to which it is believed to be strictly related (2001: 14-15). However, this view is not universally accepted and is currently object of a controversy, perhaps because describing and operationalising the concepts of attention and awareness has so far been rather challenging, due to both theoretical and methodological issues (see Leow and Bowles, 2005, as well as Ahn, 2014), *in primis* because the notion of attention and awareness themselves are very confused (Truscott, 1998: 105).

Noticing is also closely related to attention. In fact, it has been defined as its subjective correlate (Schmidt, 2001: 5). Several scholars have proposed different models to explain this construct (Gass, 1988; Schmidt, 1990, 1995, 2001; Robinson, 1995; Tomlin and Villa, 1994; Truscott and Sharwood Smith, 2011). Moreover, different methodologies have been put forward to measure it, which is amongst the reasons why there is still a lack of consensus among researchers on the nature of the relationship between noticing and SLA (Smith, 2012: 55). It becomes therefore paramount to attempt to define as precisely as possible what is referred to by 'noticing' and how it is going to be measured. For the purpose of this study, I start from the assumption that noticing is a necessary condition for FLL learning (Schmidt, 1990, 1995). Schmidt believes that conscious mental processes are central to learning and that SLA largely depends on what is attended to and noticed in L2 input (Schmidt, 2001: 4)[16]. In his Noticing Hypothesis, it is claimed that

---

[16] Some studies suggest that a degree of retention of novel forms without awareness is possible (e.g. Williams, 2005). While I do not deny the possibility of subconscious language learning, it remains to be seen whether the quality of input elaboration in such a condition can actually lead to long-term

input must be consciously noticed in order to become intake, where input can be defined as "the actual language used by people speaking to the learner", whereas intake is "that part of the input that the learner can and does process" (Andersen, 1983: 7). Therefore, intake occurs at the initial stages of the acquisitional cline and refers to that part of the language that has been processed by the learner but not necessarily yet integrated in their language system (Leow, 2013a: 270).

An alternative view to that of Schmidt is presented in a seminal paper by Tomlin and Villa (1994), who, despite agreeing on the important role that attention plays in FLL, argue that *conscious* mental processes are not strictly paramount, at least in the preliminary processing occurring during exposure to input. In their model, Tomlin and Villa distinguish between alertness, orientation and detection. Alertness is referred to as "an overall, general readiness to deal with incoming stimuli or data" (1994: 190). Orientation and detection are more relevant to the present study as they are concerned with "the outcome of specific allocation of attention resources" (ibid.). Orientation is defined as the act of "committing attentional resources to sensory stimuli" (ibid.), it does not necessarily involve awareness and it can facilitate detection. According to the authors, it can have a facilitative or inhibitory effect on further processing, and this is based on whether what follows falls within the subject's expectations or not (1994: 191). As such, orientation can be explored through the use of eye-tracking, a commonly employed method to investigate visual attention (Perego and Ghia, 2011: 177). Detection is defined as "the cognitive registration of sensory stimuli", which "selects, or engages, a particular and specific bit of information" (Tomlin and Villa, 1994: 192). According to Tomlin and Villa it does not necessarily require awareness, it makes further processing possible and it can be enhanced by alertness or orientation, although it can happen without them (ibid.: 197).

Another model that addresses the construct of noticing was introduced by Robinson, who defined noticing as "detection plus rehearsal in short-term memory, prior to encoding in long-term memory" (1995: 296). In his model, Robinson explicitly links the attention and memory systems, while also neatly bridging the gap between Schmidt's and Tomlin and Villa's views on awareness by placing detection at an earlier stage than

learning and have a stable impact on cognition (Baars, 2002). I agree with Godfroid et al.'s statement that "while the issue of whether there can be any learning on the basis of unconscious detection alone has not yet been settled, many researchers now agree that unconscious learning – if it exists at all – is negligible" (2010: 173). From these consideration stems my endorsement of Schmidt's assumption that noticing is necessary for learning.

noticing in the acquisition process (Leow, 2013a: 271). Therefore, "linguistic information may be detected and taken in by the learner, but if this information is not accompanied by awareness, then the chances of this information being further processed is relatively minimal" (ibid.). It is precisely this view of noticing that I adopt in this piece of work. Although awareness is not central to the present investigation, since it is such a highly debated topic and so closely related to noticing, any study dealing with the latter should be clear on its theoretical stance on the former. Like Schmidt and Robinson, I consider noticing to be a conscious process.

Two distinct aspects of noticing will be herein considered, namely (conscious) noticing and metalinguistic awareness or overt noticing. The latter refers to a higher level of abstraction, a conscious reflection of what has been attended to. It includes open accounts on aspects of utterances or the text as a whole as described by the learner and happens *after* the learning activity (in this case, viewing). The former, on the other hand, refers to the objective detection of surface structure elements in L2 input and happens *during* the learning activity. Therefore, noticing is hereby intended in a very concrete acceptation, at a very low level of abstraction. In the present approach, it is words and structures in the input that noticing is concerned with, rather than the abstract rules governing them. In this sense, the term 'noticing' as used by Schmidt can be recast as Tomlin and Villa's detection within selective attention (1994: 199) – where selective attention has been defined as "the sustained focus on relevant information" (Farmer et al., 2012: 358) – and as such it will be used throughout the thesis.

Both aspects of noticing addressed are assumed to involve awareness. What changes is the degree of awareness involved. On a theoretical level, noticing 'proper' (*during* viewing) involves attention and a lower level of awareness (Leow, 2013a; Robinson, 1995), whereas metalinguistic reflection (*after* viewing) naturally involves a higher level of awareness. It should be pointed out that this view does not categorically exclude subconscious learning, since endorsing Tomlin and Villa's three-way model means accepting that detection without awareness can indeed occur. In fact, some time after the publication of Tomlin and Villa's seminal paper, Schmidt (2001: 18) updated his theoretical framework of noticing by distinguishing between detection without awareness and detection within focal attention, that is, conscious noticing, *de facto* recognising a value in Tomlin and Villa's model. On a practical level, however, distinguishing and investigating the two is difficult, and beyond the scope of the present study. Teasing out such degrees of awareness (or lack thereof), if at all possible, would require a different design altogether, e.g. that of a tightly controlled priming experiment.

Part of the reason why investigating noticing has been so problematic is that it is an internal process, and as such it can only be observed and measured through inference and indirect methods. To date, a variety of measures have been used, both online or concurrent (think aloud protocols, verbal reports, underlining, eye-tracking, note-taking) and off-line or non-concurrent (post-task questionnaires, offline verbal reports, oral interviews, pretest-posttest-elicited recall), but researchers agree that neither are sufficient to fully assess the cognitive processes occurring in the mind of the learner (Ahn, 2014). The diffusion of eye-tracking has breathed new life into empirical explorations of this construct, since this tool differs from other concurrent measures such as, for example, underlining, in that the latter "indicates the locus of attention, but it never quantifies attention" (Godfroid et al., 2013: 488), whereas the former "not only performs well on the criteria of completeness and precision but also provides a measure of the *amount* of attention that participants allocate to target forms in the input." (ibid.). Since attention does not need to involve awareness (Tomlin and Villa, 1994), and eye-tracking assesses the location and amount of attention dedicated to input (Godfroid et al., 2013), in principle this non-intrusive technology may allow to account for both conscious and less conscious attentional processes. However, current eye-tracking designs employed in the study of noticing have not been fine-grained enough to disentangle these two aspects of attention on their own. Morevoer, Smith notes that both eye movements and stimulated recall *indirectly* reflect noticing and they may "tell us little about the nature of cognitive processing that ensues from such a [noticing] event" (2012: 62).

Therefore, noticing is best assessed through multiple measures (Godfroid et al., 2013; Leow, 2013b). Data triangulation may be particularly beneficial as it counterbalances the lack of clarity of each of these measures taken individually (Ahn, 2014; Godfroid and Schmidtke, 2013). In this study, noticing will be investigated through three measures: eye-tracking, a verbatim recognition post-test and an explicit report task. The attentional function of orientation to subtitled input will be measured through eye-tracking, while metalinguistic awareness through the explicit report provided in an open-ended questionnaire (see Appendix A.4). Triangulating eye-movement data with post-test scores and explicit reports from the viewers will allow us to gain some insight on whether detection of language features in the subtitles occurred.

If, as stated by Tomlin and Villa, detection "is the process through which particular exemplars are registered in memory and therefore can be made accessible to whatever the key processes are for learning" (1994: 193), then one can assume that specific instances of L2 input that has been attended to and correctly remembered in a memory post-test will have been successfully detected by the learner, though it would not be possible to establish whether with or without awareness (because, in principle, one could consciously or subconsciously detect an item and then be able to recognise its wording immediately after viewing). If, however, a learner attends to an item, correctly recognises it in the post-test *and* reports it as striking in the questionnaire, then one can infer that noticing has occurred, i.e. that some degree of further conscious elaboration must have ensued at the processing stage – e.g. as the learner realised that item does not match their expectations – thus allowing for the subsequent spontaneous production from memory. If that learner then ventured as far as explaining *why* they found that item striking, evidence for a higher degree of metalinguistic awareness would also be present. Therefore, although the role of awareness in SLA and cognitive psychology is not the main focus of this investigation, some insightful observations might be made in this sense. Moreover, should any L2 word or structure be described as previously unknown by learners in the explicit report task, this would provide some evidence that novel word learning can occur after a single exposure to reversely subtitled material. Thus, despite the grey area outlined above with regard to the subconscious aspects of processing and learning (which are not probed through the present experimental design), insights on how perception of surface elements interacts with short-term memory in an AV environment might still be gained by exploring the relation between orienting attention to a particular subtitle, performance in a memory recognition task for that subtitle and its potential spontaneous free-recall.

Finally, 'learner-initiated noticing' (Godfroid et al., 2010: 169) is addressed in this study, meaning that orientation to a subtitle is non-elicited: learners direct their attention to aspects of the input themselves, without being led to do so by a teacher or another form of prompt. Investigating this type of noticing (as opposed to externally-induced noticing) can be particularly beneficial in a FLT perspective, since the more spontaneous noticing occurs during classroom time, the less a teacher will have to draw learners' attention to formal L2 features, and the more free time the learners will have to develop their L2 skills through active practice, for example through communicative activities (ibid.: 170).

2.9 Memory Research

Memory plays an important role in real-world complex cognition (Shah and Miyake, 1999: 1), and language is a prime example of a naturally occurring, hierarchically ordered complex system (Simon, 1962). Early theoretical accounts (Atkinson and Schiffrin, 1968) divided the construct of memory into short-term memory (STM) and long-term memory (LTM), where the former was considered to be responsible for the maintenance and integration of information and have discrete stores through which this information would be passed on to a more permanent LTM store (Wood Bowden et al., 2005: 115-116). In this traditional view, attention allocation and rehearsal play a crucial role. It is thanks to the former that the input is selected to be moved to the short-term store (STS) and thanks to the latter that fast decay of input in STM is prevented and encoded information can be moved to LTM. It is generally assumed that memories leave a trace in the brain, where a trace is a form of physical and/or chemical change in the nervous system (McLeod, 2008). Traces are believed to decay fast and automatically, with STM being able to hold input active in the mind between 15 and 30 seconds, unless it is rehearsed (ibid.). After this threshold, the memory trace decays, i.e. quickly fades away. Rehearsal has been defined as "active processing that keeps information available in consciousness such that the information can be immediately and accurately recalled at any time during which it is being rehearsed." (Dark and Loftus, 1976: 480). Rehearsal can be divided into 'maintenance' and 'elaboration'. If, on the one hand, subjects only rehearse the memory trace as to maintain it in a simple way (for example through a phonemic representation such as repeating a phone number to oneself), the transfer of information from STM to LTM will not be facilitated, and such repetitions or a longer sojourn of the trace in STM will not necessarily result in learning and long-term retention (ibid.). If, on the other hand, the information is enriched and elaborated in some way, then subsequent retention will be enhanced (ibid.). Possible examples of elaborative rehearsal are "making associations between the new information and what one already knows, creating a mental image of the new information, recoding information in some way such as taking notes on a chapter while reading it, or creating some mnemonic device that helps memory of the information" (Thorne, 2003: online). An alternative to this modal view postulated working memory (WM) as a series of representations within LTM (Norman, 1968) where WM ability is essential not only for the maintenance but also for the active control of attention on a stimulus, especially if conflicting or interfering stimuli also occur. Shah and Miyake (1999: 1) define WM as "the theoretical construct that has come to be used in cognitive psychology to refer to the system or mechanism underlying the maintenance of task-relevant information during the performance of a cognitive task". Therefore, WM is considered to be a system not only dedicated to the *storage* of information, but also to its *processing*, whose resources are limited and "must be shared between the work and the memory, between the processing and storage demands of the task to which the working-memory system is being applied." (Daneman and Merikle 1996: 423).

One of the most influential representations of memory is Baddeley's working memory model (Baddeley, 1986; Baddeley and Hitch, 1974). Three components of WM are identified: the central executive or supervisory attentional system (SAS), responsible for attention allocation and the activation and inhibition of information, and two sub-systems, the phonological loop and the visuo-spatial sketchpad, which are related to short-term memory and manipulate temporarily held verbally- and visually-encoded information respectively. The phonological loop comprises a phonological store, containing phonological information subject to decay over time, and an articulatory rehearsal process whereby decay can be fought through inner speech, e.g. when we might mentally repeat a phone number to prevent ourselves from forgetting it. The visuo-spatial sketchpad deals with input coming from the visual and tri-dimensional environment, generating and managing mental images (N.C. Ellis, 2001). In this multi-component view, because attention is a limited capacity (Schmidt, 2001), the sub-systems also have limitations in terms of input that can be dealt with at the same time, and success in accomplishing a complex cognitive task such as reading or mentally calculating the amount of change one is due depends on the different demands posed during information processing and interference from other memory contents. So, for example, when explaining a route or giving directions, some people might find themselves closing their eyes, as to prevent other concurring visuo-spatial input from hindering the task at hand. More recently a fourth component, the episodic buffer, has been added to the model (Baddeley, 2000, 2003): whilst the SAS is responsible for attentional control, the episodic buffer deals with storage and integration of input into "single, multifaceted episodes" (Wood Bowden et al., 2005: 116).

Although the concept of WM is commonly used in the literature, there is often confusion as to what the term actually means (Shah and Miyake, 1999: 1). It is not yet fully clear what is the relationship between WM and STM (ibid.: 2), nor between WM and LTM (Wood Bowden et al., 2005: 116). However, most models still agree on the role played by attention as a limited resource, the existence of different modalities of storage, the distinction between the temporarily activated STM and a more permanent LTM, and the presence of specialist systems for perception and representation (N.C. Ellis, 2001: 35). Despite the fact that the processes, functions and subsystems of memory have clearly not been agreed upon yet (Wood Bowden, 2005: 117), it is widely accepted that memory and language, key elements in human cognition, are closely related, and language acquisition in particular is seen by some as "one of the best examples of the close collaboration of working-memory and long-term memory" (Szmalec at al., 2013: 76).

Several empirical studies used these theoretical constructs and measures of memory capacity to investigate different aspects of language acquisition and processing. As we have seen in the previous sections, in most of the experimental SLA and AVT research involving L2 perception, comprehension and processing (e.g. Lambert et al. 1981; Danan, 1992; Paivio and Lambert, 1981; Bird and Williams, 2002; Ghia, 2012a, just to name a few) memory for language was tested, with item recall and recognition being the most common memory skills utilised in this sense. In line with the SLA and AVT literature, scores on a recall post-test will be collected in this reception study, with the aim of shedding light on whether translation within the audiovisual medium, in the form of ready-made L2 subtitles attached to L1 audio (reverse condition) can be a psychologically effective tool for language retention, the first step towards learning. This study looks at processing effort and the amount of time spent reading literal and non-literal translation respectively to establish its relation with mnemonic accuracy as measured by verbatim recognition.

At this point, some relevant terminological clarifications need to be made explicit, first of all that between recall and recognition. These are two common ways to examine language knowledge in SLA and psycholinguistic research, yet measurement studies show that they start from separate mnemonic premises and require different processing tactics (Clariana and Lee, 2001; McDaniel and Mason, 1985) and neural imaging studies indicate that they have separate distinct neural correlates (Allan and Rugg, 1997). While in tests on recognition memory learners are typically presented with MCQs where they have to select the correct answer from a given set containing the target item and a number of foils, recall tests require learners to produce responses from memory. Recognition tests are generally deemed to be easier and strengthen existing memory traces, while recall tests are more difficult because learners have to actively retrieve the correct response from mental representations of the linguistic information they experienced during processing (Jones, 2004; Clariana and Lee, 2001). This difference is mirrored in the fact that, generally speaking, our recognition vocabulary is much larger than our production vocabulary (Lado 1957: 81). Throughout this thesis, 'recall' and 'retrieval' are both used as general terms referring to any memory process engaged in retrieving L2 form and/or L2 meaning. On the one hand, our recall post-test (see 3.12 and Appendix A.3) is *de facto* a verbatim recognition MCQ where subjects are given three possible L2 phrasings and have to *recognise the exact L2 form* they have read in the subtitles. On the other hand, our open questionnaire (see 3.15 and Appendix A.4) *de facto* investigates free-recall, i.e. whether the subjects spontaneously mention words or expressions that they found striking. Hence, subjects have to more actively *recall L2 form*, in full or in part. In this sense, the distinction made in this thesis between recognition and free-recall reflects the broad distinction made between receptive and productive knowledge respectively.[17] Finally, *recognition memory* as tested in the recall post-test should not be confused with (*word) recognition during reading*, the phenomenon reviewed by Koda (1996) and discussed in sections 2.5 and 2.6.3. While the former deals with retrieval, the latter deals with moment-by-moment processing. The reason why the immediate post-test chosen for this reception study involved verbatim item recognition rather than free-recall is related to noticing, and well explained by Schmidt:

> Failure to achieve above-chance performance in a forced-choice recognition test is a much better indication that the subjective threshold of perception has not been exceeded and noticing did not take place. If subjects (…) cannot identify which forms occurred in input when forced to choose between alternatives, that would be much stronger evidence for the absence of noticing than their inability to produce them. (2001:20)

Free-recall was nevertheless probed through the open questionnaire, thus allowing attention data triangulation, and providing a more thorough account of noticing (2.8).

Alongside measuring retention, the present investigation includes a measure of WM (see Appendix A.2). A number of tasks requiring subjects to store information as well as engage in other cognitively taxing activities have been used to test the influence of the WM system on several aspects of language processing and learning, such as the Reading

---

[17] For reasons of brevity and relevance to the present work, I do not delve deeper in the discussion of issues of productive/receptive vocabulary. However, I fully recognise that vocabulary knowledge goes beyond dichotomous distinctions of this type. For a thorough discussion of the active/passive, productive/receptive distinction, as well as of what it means to know a word, see Nation (2001).

Span Task (Daneman and Carpenter, 1980), a complex span task commonly used to assess verbal working memory specifically. Research using these paradigms has shown WM ability to correlate with native vocabulary knowledge and support the acquisition of new word forms, both in the L1 and L2 (Szmalec et al., 2013: 82; Gathercole, 2007: 761). The WM system has been found to influence language and reading comprehension (Daneman and Carpenter, 1980; Daneman and Merikle, 1996), as well as L2 comprehension specifically (Harrington and Sawyer, 1992). Moreover, although developing research paradigms to demonstrate the link between WM and higher-level text understanding is problematic, "it seems obvious that verbal working memory (…) is needed to retain the surface structure of a sentence until the proper syntactic interpretation has been made" (Szmalec et al., 2013: 83). And indeed, the relationship between syntactic complexity (King and Just, 1991) as well as online syntactic processing (sentence parsing) and verbal WM is now also attested (Farmer et al., 2012: 355; Gathercole, 2007; Friedman and Miyake, 2004). If WM is "a set of processes and mechanisms that can be used to support the temporary storage and manipulation of information in the course of complex cognitive activities" (Gathercole, 2007: 757), there is little doubt that a complex activity such as reading of L2 subtitles will be influenced by individual differences in WM ability. As Daneman and Merikle put it "(...) individuals with inefficient [WM] processes have a functionally smaller temporary storage capacity, because they must allocate more of the available resources to the processes themselves." (1996: 423). Therefore, in order to control for confounding effects of WM on recognition memory in the post-test, after watching the experimental video, participants sat a WM test, details of which are presented in 3.10.

2.10 Eye-tracking

Eye-tracking has been used in psychology research for over 100 years (Smith, 2012: 57). Since the 1960s, this technology has been exploited to explore human visual attention (Perego and Ghia, 2011: 177), and more recently has been employed in the field of Translation Studies (Göpferich et al., 2008), in particular to investigate processing mechanisms during the process of translating (O'Brien, 2006). Although reading research has also been investigating eye movements for over a century (Rayner, 1998: 372), it is with the tremendous evolution of eye-tracking systems witnessed from the 1970s onwards that reading research really proliferated, resulting in a flurry of psycholinguistic studies, particularly in the last 40 years (Keating, 2014). Since then, eye-tracking has become one of the prime means through which the perceptual, attentional and processing components of reading were explored. The ever-growing resolution and precision of this technology allows researchers to "rather precisely infer what is happening in the mind as we read" (Schotter and Rayner, 2012: 83). Before eye-tracking became widely available, other research paradigms were used to investigate attention allocation and processing during reading, such as lexical decision and naming tasks. However, many consider eye-tracking superior to these, because "reading processes in eye tracking are not confounded by task-related processes or strategies that other lab tasks (e.g. lexical decision or naming) entail. Hence, this method is considered to be the closest experimental parallel to the natural reading process." (Cop et al., 2015: 2).

Today, eye-tracking has gained considerable momentum in psycholinguistic investigations of subtitling (Kruger, 2016). Analysis of visual and text processing during exposure to subtitled input can shed light on issues related to usability, accessibility as well as SLA, and it has become evident that eye-tracking can prove a very versatile tool in several areas of investigation. As we have seen in 2.6.2, eye movements are indicative of moment-by-moment processing (Just and Carpenter 1980) and are considered "empirical correlates of processing complexity" (Smith, 2012: 57), therefore enabling us to make inferences on the perceptual processing of input. Eye-tracking is currently used in several fields of study to explore cognitive factors in relation to word recognition, frequency and familiarity, lexical and syntactic ambiguity, as well as syntactic, semantic and pragmatic factors, amongst others (for a review, see Rayner, 1998 and Clifton at al., 2007). A plethora of eye-tracking measures are currently available (Holmqvist et al., 2011). Some have traditionally been associated with certain research paradigms, however measures can also be borrowed between fields. So, for example, first fixation duration came from the field of reading research and was then borrowed in scene perception studies (Holmqvist, 2013). The most common eye-tracking measures are fixations and saccades. Eye fixations are deemed important because they allow readers to extract useful information from a text (Dussias, 2010). Therefore, measures of fixation duration, location and number can be instructive with regard to how people acquire information from a printed text (Clifton at al., 2007: 344-345). The first fixation, for example, is believed to be the locus where lexical activation starts and provide a measure of early word processing (Paterson et al., 2012). Although a fixation typically lasts for 200-250ms, individual differences are wide, and this variability has been identified as a function of cognitive complexity during text comprehension (Rayner, 1998). Task context is also crucial in measure analysis, as the same measure can be interpreted differently depending on the task. For example, higher fixation frequency on a particular area can be indicative of greater interest in the target, but it can also be interpreted as a sign of complexity, meaning that the target is in some way more difficult to encode (Jacob and Karn, 2003; Just and Carpenter, 1976). However, in a search task, these interpretations may be reversed: a higher number of individual fixations, or clusters of them, can be considered an index of greater uncertainty in recognising a target item (Jacob and Karn, 2003).

In the context of research on noticing, few studies have exploited eye-tracking to investigate this aspect during L2 processing (e.g. Godfroid et al., 2010; Smith, 2010, 2012; Kuhn, 2012), and virtually none have done so within the field of AVT (with the exception of Ghia, 2012a). The experiment described in this thesis aims at filling this gap and, in line with the above-mentioned studies, explores noticing through eye-tracking because this methodology has been found to be a suitable a tool for the investigation of attention allocation and detection processes during fruition of a text (in this case, an audiovisual text). Specifically, fixation measures will be taken as a proxy for the attentional function of orientation (Tomlin and Villa, 1994). As we have seen in 2.8, orienting attention is defined as "committing attentional resources to sensory stimuli" (ibid.: 190), which is precisely what happens when viewers fixate on words and sentences during subtitle reading. Eye movements *per se* do not necessarily imply conscious commitment of such attentional resources, however. Words and sentences can be read without immediate awareness, for example when the eyes of a reader continue moving across a text even though their mind wanders to think about something different and independent from the text, so that nothing in that portion of text is comprehended (Rayner and Fischer, 1996). This process is known as 'mindless reading' (Reichle et al., 2010) and shows that subliminal processing of language can occur while the reader is engaged in different, unrelated thoughts. Eye-movements can therefore be all the more appropriate a representation of orientation, since, as we have seen, this function can but does not necessarily involve awareness (Tomlin and Villa, 1994: 197). Moreover, in the Van de Poel and d'Ydewalle's study reviewed in 2.5, the authors state: "[o]ur study does not enable a decision regarding whether the facilitation effect is due to attentional factors (a similar language could be easier to attend to), to the acquisition process itself (linguistically closer structures are easier to retain) or to an interaction of both." (2001: 271), and this is because their study did not exploit eye-tracking. This piece of research, on the other hand, does, and can therefore substantially improve our knowledge of the relationship between CLI and attention in reverse subtitling. Assuming for a moment that a facilitation effect of formal similarity will be found, if it was due to attentional factors alone, then one should register a noticeable difference in attention allocation between subtitles that are formally close to and formally distant from the ST. More specifically, diverging translations should attract more looks, if, along the lines of Van de Poel and d'Ydewalle, one postulates that similar words and structures are easier to attend to (where ease of processing corresponds to less and shorter overall fixations, resulting in smoother reading). If no significant difference was found in attention allocation between formal similarity and formal discrepancy, but a significant difference was nevertheless found in recall, with literal renderings resulting in better mnemonic performance, one would be more confident that CLI is to some degree intrinsic to the acquisition process rather than being determined by attention alone, with linguistically closer structures being easier to retain *per se* compared to diverging ones (at least as far as verbatim recognition memory is concerned).

In a more general context of AVT, several studies have investigated perceptual aspects of subtitle consumption through eye-tracking. Some of the earliest were researchers from the Belgian school, who provided evidence of the automaticity hypothesis, whereby viewers read the subtitles regardless of both input-specific and viewer-contingent factors such as sex, age, hearing impairment and familiarity with subtitles (2.7.1). Furthermore, it is through the use of eye-tracking, for example, that Perego et al. (2010) provided evidence of the cognitive effectiveness of subtitles (2.7.1) and d'Ydewalle and De Bruycker (2007) analysed subtitle line distribution by monitoring children and adults watching films with one-line and two-line subtitles in the reverse and standard condition (2.7.2). Amongst other applications of eye-tracking to AVT, Caffrey (2008), Secară (2011) and Kruger and Steyn (2014) are also noteworthy. Caffrey analysed subtitle processing with Japanese fansubbed *animae,* a type of unconventional abusive crowd-sourced subtitling where extra input is added on screen (several lines containing notes and glosses, sometimes appearing simultaneously, along vertical and horizontal directions, at both the top and the bottom of the screen). Results found a taxing effect on viewers, as informed by saccade, fixation and pupillometric measures, regardless of their level of interest in the video. In an initial small sample pilot (n = 4), Secară compared conventional and creative (including txt lingo) subtitles. She analysed fixation duration, image-subtitle attention shifts and within-subtitle regressions, and found that txt lingo in subtitles did not disrupt the viewing process as they did not result in increased regressions compared to normal subtitles. Kruger and Steyn investigated the relationship between subtitle reading and academic performance across six English L2 interlingual subtitled videos (recorded psychology lectures where both the audio and the subtitles were in English). They compared reading patterns of a test group (videos with subtitles) and a control group (without subtitles) of native speakers of different indigenous languages from South Africa, who also completed a MCQ comprehension post-test about the content of the lectures. They created an original measure of subtitle visual processing, the RIDT (Reading Index for Dynamic Texts), which was found, through empirical validation, to capture "reading behaviour in dynamic texts in a time-efficient and robust manner" (2014: 114). Kruger and Steyn found a high positive correlation between the average RIDT participant score and the post-test comprehension score ($r = .56$): students who watched the subtitled lectures and also fully read the subtitles performed more accurately than those who watched the subtitled lectures but did not read the subtitles as fully (2014: 118), suggesting a positive role of subtitles in reading instruction and language learning in academic contexts.

As these examples show, AVT scholars have realised that "eye movements represent one of the best ways to study language comprehension processes" (Rayner and Pollatsek, 2006: 613), and eye-tracking is becoming a widespread technique to study subtitle reading patterns and appraise access to subtitled input (Ghia, 2012a: 73) for a diverse range of research questions. It is through this technique that the present study also investigates information processing and attention allocation during subtitle reading. The aims of the present experiment with regard to eye-tracking are threefold: first, examine if and how much learners attend to reverse subtitles during processing of rich and complex audiovisual input, where multiple channels are used to convey meaning, starting from the previous evidence that matching subtitles with video is a cognitively effective strategy, as images do not prevent captions from being read, nor captions impede images to be processed. Secondly, I seek to compare visual activity between the two translation strategies and establish to what extent eye movements correlate with recall performance for each condition. Thirdly, we will explore the three-way relationship between (a) form of input (translation condition), (b) the measures of noticing mentioned above, including orientation to the written input examined through fixation measures, and (c) the explicit report measure (questionnaire). As Hefer noted: "The importance of choosing the [fixation] duration and count parameters as dependent variables lies in the insight they offer: they are indicative of processing difficulty" (2013: 27). Therefore, fixation duration and fixation count data will be collected, both on individual subtitles and in total. I expect that, at the linguistic level, lexical items will be more thoroughly processed, based on the common trend in reading whereby content words are fixated more than function words (Field, 2004). This is also in line with theoretical perspectives on input processing (Van

Patten, 2007) according to which "learners know that there are differences between content lexical items (e.g. *cat, sleep*) and non-content lexical items (e.g. *the, is*) and will seek out content lexical items first" (2007: 117).

To conclude, eye-tracking can add a powerful dimension to the exploration of constructs of attention and noticing, and contribute to defining their role in SLA. It has been hypothesised that smoother, more streamlined reading might result in shallower and less attentive encoding (Perego and Ghia, 2011: 192) and that salient input might on the other hand produce increased visual activity, which has been equated to increased mapping, which in turn may or may not lead to more retention. In the present study, eye-tracking will be used to put these theories to the test, in order to assess whether the formal similarity (literal transfer) vs. formal discrepancy (non-literal transfer) distinction has a psychological validity and whether salience created through translation enhancement has an effect on noticing and memory within the reverse subtitling viewing modality.

2.11 Frequency

Frequency plays a role in several language-related areas, including sociolinguistic variation and language change (Ellis, 2001). In fact, it has been maintained that frequency information predicts human performance in several other non-language domains, "from the acquisition and representation of knowledge (…) to decision-making to sex-role development" (Hasher and Zacks, 1984: 1372).

As we have seen in 2.6.3, psycholinguistics presupposes a level of rule abstraction from encountered instances of language, whereby we form "an internal model of the world" (Ellis, 2001: 148) through mental representations. Frequency is deemed to play a key role in defining the structure of the mental lexicon and to exert its effect upon lexical access (Whitford and Titone, 2012; Rayner, 1998). Frequency is believed to guide this process of abstraction, so that "in the course of conversation we acquire frequencies of the elements of language and their mappings" (Ellis, 2001: 146). When frequently encountered, these elements will have a higher base of activation, meaning that their storage will allow easier accessibility, since they will need less additional activation to be retrieved (Morton, 1986; Schmauder et al., 2000). Indeed, frequency is a milestone in both word processing (reading) and word recognition research, as measured by many paradigms, including lexical decision RTs and eye movements (Whitford and Titone, 2012: 73). Frequent words are read faster than rare words (Adelman et al., 2006: 3; van

Heuven et al., 2014: 1776). More specifically, when word length is controlled for, frequent words will receive shorter fixation durations, while rare words will receive longer fixations (Rayner et al., 1989; Pollatsek et al., 2008; Moran, 2012), which is known as the Word Frequency Effect (Inhoff and Rayner, 1986; Rayner and Duffy, 1986). Moreover, when frequent words are short function words, they are often skipped (Rayner, 1998; Rayner and Fischer, 1996). Frequency can also affect neighbouring words. Recall from 2.6.2 that, thanks to the preview benefit and word identification span, parafoveal words to the right of the currently fixated word will undergo some form of parafoveal processing. If a such a word *n+1* is frequent and/or short enough to be fully identified before it is fixated, it may be skipped, while if such a word is longer (i.e. only its first few letters will fall within parafoveal view), the preview benefit will still facilitate foveal processing (Rayner, 1998: 381). Therefore, alongside word-level effects on the currently fixated items, combined frequency effects also affect overall measures such as reading speed (Moran, 2012: 190).

However, frequency effects are not universally accepted, their value having been recently reassessed in the literature. For instance, although frequency has been found to be an important predictor of both lexical decision and word naming tasks, some recent studies support the idea that frequency is often confounded with other variables, such as contextual diversity (CD), i.e. the number of contexts in which a word is experienced (Adelman et al., 2006: 4). Frequency is usually extracted from corpora, so it can be affected by peculiarities of certain text passages more than contextual diversity would, since a rare word could be found several times in a single passage, which would inflate frequency but leave contextual diversity unaffected (ibid.: 10). Through regression analysis, Adelman and colleagues found not only that CD predicts RTs independently of frequency but also that there was no evidence of a facilitatory effect of frequency independent of CD on lexical decision and word naming RTs for both young and older participants (ibid.: 5-6). These findings suggest that frequency effects in processing can sometimes be problematic to assess. Frequency effects in learning are also not always clear-cut. For example, Ellis and Schmidt (1998) demonstrated that frequency effects on morphology are moderated by L2 proficiency: when asked to produce past tense forms, lower-level learners show shorter latencies for frequent items in both regular and irregular past forms. As L2 proficiency increases, the frequency effect on regular items diminishes, whereas it remains for irregular items. Frequency effects are therefore not absolute and can be less evident in cases of high proficiency (at least as far as morphology is concerned). This is relevant to the present investigation of reverse subtitles, since the participants in the current study also had a relatively high level of proficiency (CEFR B2), which could moderate the influence of frequency.

Even in reading research, where frequency effects are considered rather robust, their influence is variable and needs to be further qualified. As we have seen in 2.6.4, most reading research is concerned with isolated, single-word processing. An exception is Cop et al. (2015), who analysed eye movements in unbalanced bilingual L1 and L2 reading compared to monolingual reading of a complete novel (56,000 words). They report few frequency effects at the text level, proposing that low frequency words can be easier to process with continuous text "because of the context it [the text] provides to identify such a word" (2015: 32). Moreover, frequency has a clearer impact on *early* measures of processing (e.g. single fixation duration and first fixation duration), not as much on other measures used by Cop and colleagues, e.g. sentence reading time including fixations and second-pass fixations (i.e. dwell time), fixation count, regressions, average fixation duration and average rightward saccade per sentence, amongst others. Thus, frequency has a smaller effect on natural reading than on processing of out-of-context words or target words embedded in isolated sentences. This could also be the case in a highly contextualised situation such as viewing subtitled video, a form of input typically considered 'rich' precisely for the multiple sources of information it provides, especially when authentic (see 3.7).

As far as frequency effects in subtitling are concerned, one study was recently published (Moran, 2012; also see 2.7.3), where the author varied word frequency as to obtain two clip versions, the first (C1) with half the subtitles containing low-frequency words, the second (C2) with half the subtitles containing high-frequency words. The other half of the subtitles did not change between C1 and C2, as to have 'constant subtitles' (where the words did not change between conditions) and 'variable subtitles' (where words and their frequency changed between conditions). English clips with English subtitles were produced, and English native speakers were tested (monolingual condition). Moran found differences in mean fixation duration in the subtitle area, with low-frequency subtitles attracting significantly longer durations compared to their high-frequency counterparts. Differences were also found for sum of gaze points on the image, with subjects in the low-frequency condition spending significantly less time looking at the images. While fixation durations indicate processing difficulty, gaze point data show the distribution of visual attention. The findings suggest that high-frequency words were more easily processed by the viewers, who had more time to redirect attention to the moving image and read the subtitles more quickly even though they were sometimes longer than their low-frequency counterparts.

Research has also shown that low-frequency items are better recalled than high-frequency items in recognition memory tasks (Gorman, 1961), suggesting that frequency effects may play a role in mnemonic processes as well. In fact, a more explicit integration of memory processes into models of reading has recently been proposed (Adelman et al., 2006). In line with Gorman (1961), the word frequency effect paradox states that lists of low-frequency words are better identified as targets and more accurately rejected as lures in word *recognition* memory, whereas lists of high-frequency words are better remembered in *free-recall* tests (Mandler et al., 1982; Lohnas and Kahana, 2013). These contrastive findings have traditionally been explained by low-frequency words attracting greater attention or being more differentiated from one another (resulting in greater recognition), and high-frequency words being more easily associated with each other, thus facilitating recall (ibid.: 3). As far as the present study on reverse subtitling is concerned, although frequency effects on memory do not seem to be as clear-cut as those on processing presented above, this variable could nevertheless bias recall results in relation to translation condition. If translation condition did not have a psychological reality and did not impact recall, but frequency did, one should find that low-frequency subtitles are recognised more accurately than high-frequency ones regardless of how formally similar or distant from the ST they are. If, on the other hand, an effect of translation condition on learner processing and performance is found despite subtitle frequency, this would provide additional support to the view that the concept of 'literalness' does exist in the mind of the learner, who is affected by its manipulation independently from frequency. Including frequency in this study will therefore also help establishing whether the literal vs. non-literal distinction is tangible and meaningful psychologically. Since frequency could play a role both in the processing and recall of reverse subtitles, it was included in the eye-movement analysis as well as the recognition memory analysis (see 4.3.7). Accounting for possible frequency effects in the present study also substantially improves on Ghia's (2012a) design, whose investigation of literal vs. non-literal translation did not include such analysis. Finally, the effects of frequency on language and cognition have been mostly investigated in monolingual domains such as L1 reading research (Hicks et al. 2005; Watkins et al., 2000), while L2 frequency effects remain relatively unexplored (Whiteford and Tritone, 2012: 73). The present investigation addresses this gap by devising a way of calculating subtitle frequency (3.14) and presenting initial results of the frequency analysis. To my knowledge, this is the first study to look at L2 frequency effects on L2 subtitle reading specifically.

2.12 Experiment Validity

Generally speaking, validity is the extent to which a measurement actually measures what the experimenter designed it to measure (Field and Hole, 2003: 44). In the field of psychology, research is considered valid if "it provides the understanding about behaviour that it is supposed to provide." (Goodwin, 2008: 177). As such, validity is central to the design and implementation of any such experimental study and therefore deserves to be discussed separately. Different types of validity exist, and they relate to different aspects of the experiment. A defining characteristic of the present study is that it seeks to strike a balance between ecological and internal validity. The term 'ecological validity' was coined by Brunswik (1947) in the context of perception research and originally had a different, much more specialised meaning than what is commonly accepted today. At present, there is still some debate in the literature as to what the term refers to, with some authors seeming to confuse it with external validity (e.g. Filed and Hole, 2003). I adopt Coleman's Oxford Dictionary of Psychology (2015) definition, which provides a more recent and up-to-date stance on the matter and clearly distinguishes between ecological validity and external validity. While the former is "[t]he confidence with which the conclusions of an empirical investigation can be generalized to naturally occurring situations in which the phenomenon under investigation occurs" (Coleman, 2015: online), the latter is "[t]he extent to which the conclusions of an empirical investigation remain true when different research methods and research participants or subjects are used." (ibid.). Therefore, both terms are related to the concept of replicability and representativeness of laboratory findings. However, while the latter considers replicability in *experimental* conditions different from those in which the initial experiment took place, the former considers replicability in *non-experimental* conditions, i.e. in natural contexts or field settings where the findings of the laboratory research are to be applied. This definition reflects how the term is used in psychology research today and is in line with previous definitions such as Orne's, for whom ecological validity is the "appropriate generalization from the laboratory to nonexperimental situations" (1962: 776). Internal validity, on the other hand, is "[t]he extent to which the conclusions of an empirical investigation are true within the limits of the research methods and subjects or participants used." (Coleman, 2015: online), and refers to how well an experiment is internally constructed, i.e. whether it is methodologically sound and avoids confounds. Confounds are extraneous variables that co-vary with the independent variable and can provide an alternative explanation for the phenomena under investigation (Goodwin, 2008: 167), whose lack of control flaws the experiment because these extraneous, confounding effects cannot be separated from those of the independent variable.

In the present study, internal validity was maximised in a number of ways. First of all, a working memory test (3.10) as well as a language proficiency pre-test (3.9) were introduced to control for potential confounds at the subject-level. Second, an age limit was applied during recruitment to avoid confounds deriving from, for example, a natural decline in cognitive abilities. Third, if the eye-tracking data recorded for a participant had a low sampling rate (3.17.3), that participant was not included in the experiment despite fitting the other criteria and having passed the pre-tests. By doing this, only complete and experimentally valid data was included in the analysis. Fourth, confounds at the item-level (in this case, the subtitles) were avoided by controlling for variables that are likely to affect memory performance and processing, such as subtitle length, segmentation and frequency. Finally, comparability between subtitles was ensured by maintaining the same subtitle duration between the two experimental conditions (formal similarity and discrepancy), so that the only element of change between two versions of the same audio string is the translation.

As Ellis has noted, while laboratory experimentation is required for fine-grained analyses, ecological validity is needed to bridge theory and educational practice (Ellis, 2002b: 314). Since this research purports to offer experimental results specifically in view of how they can inform teaching, thus providing a springboard for discussing precisely how to link FLT theory and practice, ecological validity is also one of its prime concerns. This construct is relevant to all experimental investigations, since "[f]or the behavioral sciences, (…) when humans or animals are the object of study, the act of observation may very well change the object of study" (Rosenthal, 1963: 268). For example, even Bird and Williams, in their very well controlled study on the effects of bimodal input subtitles on memory performance, admit that "repetition priming experiments discourage learning of new words because a large number of words and non-words are randomly presented in a brief period of time" (2002: 518). The authors are thus aware that laboratory experiments, however internally valid, do not necessarily capture what happens in the real world. In this case, they may not appropriately reflect what it means to learn words in a natural setting, because they create a prefabricated experimental condition where the words in questions are presented in lists and have to be learnt out of context. The goal of an ecologically valid study, on the other hand, is to obtain results that would also be found were the study carried out in a real-life setting – such as watching subtitled material in the classroom or at home – in order to make the results as representative and generalizable as possible.

In the present study, a number of methodological precautions were taken to maximise the confidence of obtaining ecologically valid results. First of all, a head-free tracking system was chosen as opposed to a chin-rest system. The latter forces subjects to keep their chin and forehead pressed against a special stand to avoid extra movement, which increases tracking precision but creates an unnatural viewing situation. The former instead compensates a decrease in precision with a less-intrusive tracking experience, where the eye-tracker hardware is inconspicuously positioned under the screen and subjects are allowed to move their heads (within limits), thus allowing the experiment to reproduce natural viewing conditions more closely. Second, authentic rather than prefabricated material was chosen. Creating a made-up video specifically designed for experimental purposes will more easily allow to produce homogeneous test items (for example, only nouns of a certain length and a certain frequency of occurrence in a corpus). However, in the plethora of video material available today, a satisfactory degree of control can also be achieved via careful selection of an authentic source. Finding an appropriate sequence within such source allows to consider whatever variables are under investigation, while also making the experiment more faithful to the reality outside the laboratory. Third, to prevent an artificial condition such as that outlined by Bird and Williams (2002), this experiment does not present lists of words in isolation but looks at the audiovisual text as a whole. The focus is on words as well as structures, both occurring in context, thus paralleling the way in which linguistic information is consumed in non-experimental audiovisual settings. Fourth, qualitative measures such as a questionnaire (3.15) were integrated in the analysis and discussion, which makes it possible to take into account participants more explicitly. Consider the following quote by Orne:

> (…) the experimental model has been so successful as employed in physics that there has been a tendency in the behavioral sciences to follow precisely a paradigm originated for the study of inanimate objects, i.e., one which proceeds by exposing the subject to various conditions and observing the differences in reaction of the subject under different conditions. However, the use of such a model with

> animal or human subjects leads to the problem that the subject of the experiment is assumed, at least implicitly, to be a passive responder to stimuli. (1962: 776)

By considering subjects as passive respondents to stimuli, there is a clear risk of making misguided inferences based on purely data-driven parameters. One may assume that participants engaged in one line of thought, when in fact they may have followed a completely different reasoning, albeit resulting in the same outcome (e.g. their choice of an incorrect item in the post-test). By including qualitative data such as an explicit report task with open answers (3.15), this study purports to give subjects a voice and recognise them as active participants, rather than passive respondents. Moreover, doing so makes it possible to take into account their explicit views on specific aspects of the research, which in turn may further inform the researcher about the true reasons why they may have picked an item in the post-test, or why they looked at a specific word several times. The last measure taken to increase ecological validity is the inclusion of an open question in the questionnaire (Q4 in Appendix A.4) to prompt subjects to express any other comments they may have. An effort was made to take into account from the piloting stages what the participants have to say, not just about specific aspects probed by ad-hoc questions, but also on their experience as a whole. If allowed, participants can give invaluable insights on whether the act of observation in the conditions created by the experimenter inadvertently happens to change the object of study (*de facto* potentially undermining the generalisability of results), so that the necessary pilot adjustments can be made before collecting the main body of data. Moreover, providing a space to express opinions in a less-structured fashion allows to establish more clearly how the experiment is received by the participants, whilst again considering them more overtly as active respondents to the stimuli presented. The next chapter will expand on all tests, methods and processes mentioned in this last section. In what follows, the methodological approach outlined herein will be addressed in depth, together with the research questions and hypotheses (3.4) as well as all practical aspects of the project implementation.

CHAPTER III
METHODOLOGY

3.1 Introduction

This chapter describes the experiment itself, including its rationale, preliminary stages, implementation and data processing phase. It starts with outlining the scope and objectives of the study (3.2.), followed by the research variables (3.3.), RQs and hypotheses (3.4). Then it introduces the experimental design (3.5), the eye-tracker used (3.6), the choice and preparation of the experimental stimulus (3.7) and the participant group selected (3.8). Two control tests were used in this study at subject-level: a language proficiency test and a working memory test. The rationale for their choice and implementation is provided in 3.9 and 3.10 respectively. After these, the operationalisation of the main predictor variable – translation condition – into literal and non-literal subtitles is explained in detail (3.11) and a description of the recall post-test (3.12) is presented. The study also addresses two subtitle-specific variables, linguistic category (3.13) and relative frequency (3.14). Following these, details about the open-ended questionnaire (3.15) are provided. After having covered the study preparation stages, a section on its execution (3.16) explains what the participants were required to do during the experimental session. Finally, the data collection and processing stages (3.17) are described.

3.2. Scope of Study and Research Objectives

This study explores the relationship between AVT and the language learner during the viewing experience. A specific subtitling mode is under investigation, reverse subtitling. Within this mode, translation is addressed specifically within a FLL context in order to explore its potential as a tool in second language learning. The effects of specific translational choices on viewers' perception and noticing of lexico-syntactic structures are investigated. Formal similarity (literal transfer) and formal discrepancy (non-literal transfer) will be compared.

In the field of AVT, it has been postulated that translation may involve deep processing in the learners as they are forced to continuously map and establish comparisons between ST and TT (Pavesi, 2002). This process might in turn facilitate retention of certain linguistic structures. Despite this potential, the role of translation in interlingual subtitling has not been explicitly addressed (Ghia, 2012a). As we have seen, in the AVT research concerned with SLA, translation is mostly referred to in passing rather than addressed specifically, and often as if only a one-to-one L1-L2 equivalence and one end-product was possible. In fact, a variety of techniques are used in the translation process, resulting in very different outputs depending on context, text genre, SL and TL, audience, and so on. Despite the fact that these differences might affect learning, no specific account of translation in SLA is given in the AVT literature regarding reverse subtitles. This project was therefore designed to fill this gap and analyse the effect of translation on subtitle perception. It addresses two types of translation, how they are perceived within the audiovisual medium, and their contribution to recall scores in a memory post-test. It develops along three major axes, translation, memory and attention allocation within the audiovisual medium as measured by eye-tracking, specifically in the case of reverse subtitling (L1 audio, L2 subtitles). The core aims of the research are to establish (a) if L2 learners notice translational discrepancies between ST and TT, and (b) if they recall L2 strings more successfully when they share some formal similarity (orthographic or structural) with the L1.

It has been proposed that "reversed subtitles appear best suited for beginners since they enable them to access the foreign code in the written medium and draw initial comparisons between such dialogue and its spoken counterpart in their L1." (Danan, 2004; Mariotti, 2002, in Ghia 2012a: 33). However, since in this modality L2 listening skills and video content comprehension are not called upon, one might assume that for more proficient students (e.g. upper intermediate or advanced level) cognitive effort and attentional resources can be even more (or at least equally) successfully oriented towards the L2 subtitles, making it easier to detect specific features of the written target and draw comparisons between ST-TT, which in turn could facilitate memorisation. That said, the present investigation looks at what type of translation output can better capture viewers' attention, potentially leaving a deeper memory trace and thus yielding higher recall.

Finally, the difficulty of achieving strict variable control in experimental research implies an in-built lack of confidence in the veracity of any one single study (Sheen, 1996). For this confidence to be granted, the findings need to be reproduced and validated by a number of similar studies. As we have seen, the only comparable study on learners' noticing of language features was carried out by Ghia (2012a) in the context of standard subtitles. Given the importance of replication in the social sciences (Makel and Plucker, 2014; Mackey and Gass, 2005; Polio and Gass, 1997), this research asks parallel questions about reverse subtitles, to establish whether a similar or diverging pattern of results for memory and attention allocation is registered for the reverse mode, thus providing a more complete picture of noticing and attention in AVT.

3.3 Research Variables

This study employs both quantitative and qualitative methods. All variables considered are outlined in the table below.

Variable	Data type (Unit of measurement)	Variable type
Translation Condition	Categorical, dichotomous (L – N)	Primary, independent
Recall Accuracy	Categorical, dichotomous (Y – N)	Primary, dependent
Linguistic Category	Categorical, dichotomous (Lex – Syn)	Primary, independent
Relative Frequency	Numerical (IPM)	Primary, independent
Fixation Duration	Numerical (ms)	Primary, independent
Fixation Count	Numerical (count)	Primary, independent
Language Proficiency	Numerical (test score)	Control, independent
Working Memory	Numerical (WM score)	Control, independent
Subtitle Duration	Numerical (frames)	Control, independent
Subtitle Length	Numerical (no. words)	Control, independent
Order of Presentation	Categorical, dichotomous (1 – 2)	Control, independent
Overt Noticing	Questionnaire (open questions)	Primary, independent (qualitative measure)

Table 1. Research variables.

The quantitative data analysed are: eye-tracking metrics, scores on verbatim recall, language proficiency and working memory tests, relative frequency, linguistic category and subtitle metrics. Qualitative data, i.e. questionnaire responses, will be used to measure overt noticing (metalinguistic awareness). The dependent variable (outcome) is accuracy in a verbatim recall test, measured as number of correct scores on the MCQ (Multiple Choice Questionnaire). Scores were marked either as Y (Yes, correct answer) or as N (No, incorrect answer). The main predictor variable is translation condition, operationalised as literal (L) and non-literal transfer (N). Detailed information on how translation was experimentally manipulated is provided in the Translation Protocol (3.11). Corpus frequency was deemed to be likely to play a role and was therefore included in the analysis. After creating the two subtitle translation versions, frequency of the content words and lexicalised structures in each version was computed through the large Italian web corpus itWac. The measure used was instances per million (IPM) rather than raw counts as this gives the relative frequency weighted against the size of the corpus and makes items comparable across corpora (see 3.14). Moreover, in line with Ghia's experiment (2012a and 2012b), both individual lexical items and syntactic structures were addressed in this study. The dichotomous variable linguistic category thus classifies subtitles based on whether they are lexical or syntactic in nature (see 3.13) and its inclusion in the analysis seeks to establish if there are differences in the processing and memorization of syntax compared to lexicon. Eye-tracking was used to triangulate the findings, in particular by assessing the relationship between attention allocation and recall scores in the post-tests. The Tobii TX120 eye-tracker was used to record eye-movements. Fixation duration and fixation count are the Tobii Studio eye-tracking measures analysed in order to gain insight into the viewers' experience. Fixation duration is the length of a fixation within an area of interest (AOI), while fixation count is the number of fixations made on that AOI (3.17.2).

The above independent variables are the main predictors considered in the analysis. On top of these, linguistic proficiency, order of presentation and subtitle measures were considered. These variables are not the focus of this study and were controlled for in order to maximise the internal validity of the experiment. These control variables and their relationships with the main independent variables and the outcome will be addressed in chapter 4. Their inclusion in the analysis minimizes the chances of introducing confounds, arising when other possible independent variables not accounted for experimentally are acting at the same time as the predictors of interest, making it more difficult to conclude that a given effect – if one is found – is really due to the predictors rather than these confounding variables. In practice, to control for an effect of age and linguistic competence, recruitment was limited to participants between 18-35 years and a language pre-test was introduced (3.9). In the piloting phase, working memory (WM) was identified as a variable likely to influence subject performance and a WM test was therefore introduced in the main study. Moreover, the counter balanced design allowed to account for order of presentation of the experimental stimulus, by identifying which half of the video each subtitle translation comes from and factoring it in the analysis. Finally, two subtitle-specific measures were also included to investigate the balance of the experimental manipulation. Subtitle duration was calculated in seconds and frames, and remained the same in both conditions, as to keep L and N renderings comparable. Subtitle length was calculated in number of words and had to be allowed to change between L and N to permit the experimental manipulation of translation condition. However, to keep the two renderings comparable, the difference between L and N did not exceed a maximum of 3 words, and in several cases it was possible to achieve two translation versions with the same length (see 3.7).

Finally, the last row in table 1 represents the qualitative data considered in the analysis, namely the open questionnaire (3.15) subjects completed immediately after the post-test phase. The list of open questions is included in Appendix A.4.

3.4 Research Questions and Hypotheses

Five main RQs are addressed in this study. These are presented below and summarised in table 2 at the end of this section.

RQ1. Do different subtitling choices lead to differences in recall rate? If so, which translation condition (formal similarity vs. formal discrepancy) yields a better recall rate and is this difference statistic ($p < 0.05$)? This point will be explored through t-tests in the by-subject and by-item analyses. Regarding this first and most important research question, two outcomes are possible: (a) participants will achieve better recall scores with items presented in the literal condition (formal similarity). Indeed, in the AVT literature it has been postulated that literal transfer could facilitate the matching of source and target (Pavesi and Perego, 2008a) in multimodal contexts through redundancy of information. This hypothesis is in line with research on orthographic and semantic similarity (2.5), where several studies revealed that formally equivalent input (such as L1-L2 cognate words) can have a facilitative effect on processing (Dijkstra et al., 1999; Talamas et al., 1999; Van Assche et al., 2009, amongst others), which in turn can foster mnemonic retention (Yu, 1996; Hamada and Koda, 2011; Ghia, 2012a) and facilitate acquisition (Odlin, 2003). Alternatively, (b) participants will achieve better recall scores with items presented in the non-literal condition, as audio-text discrepancy will make the input salient, hence triggering increased noticing, which in turn will increase ST-TT comparison and mapping (formal discrepancy). One might argue that smoother reading might result in shallower and less attentive encoding (Perego and Ghia, 2011: 192), and that input salience achieved through L1-L2 discrepancy, on the other hand, might involve more processing and produce increased mapping. But does formal discrepancy result in better memory retention than formal similarity? This was the basis of Ghia's (2012a) investigation into standard subtitling, where she posited a positive effect of salient input on memory. However, the statistical data emerging from her study favour the formal similarity hypothesis: verbatim recall of both syntactic constructions and lexical items was higher in the literal transfer than in the diverging translation condition (Ghia, 2012a: 86-87). Based on this body of results, in the present study I expect recall rate to be higher when formal equivalence in L1 audio and L2 subtitles is achieved, with possible differences due to linguistic category (syntax, lexicon) within each condition.

RQ2. Two RQs related to the difference between lexical and syntactic items are explored. Firstly, is there a difference in recall between lexical and syntactic items? Lexical items are usually more easily recalled than function words (Field, 2004: 319) and more salient than longer syntactic structures (Py, 2004: 127-128), perhaps because syntax, on the other hand, typically involves complex, higher-level structures whose meaning is more difficult to access, especially when the structures are delexicalised (Seilhamer, 2010: 10). Eye-tracking research also shows that content words are fixated more than function words, which are often skipped altogether (Just and Carpenter, 1980, 1984; Hogaboam, 1983; Paulson, 2000, 2002). It is therefore not surprising that major lexical units play a crucial role in the comprehension of audiovisual material (Guillory, 1998). Moreover, while several studies have evidenced incidental vocabulary acquisitions through reading (Paribakht and Wesche, 1999; Pitts et al., 1989; Laufer, 2003; Hulstijn, 1992; Dupuy and Krashen, 1993), incidental learning for grammar is much less well-documented (Kuppens, 2010), with studies showing that instructed contexts are generally superior to incidental ones as for the retention and learning of L2 rules (Robinson, 1996). For these reasons, we expect individual lexical items to have more chances of being noticed and being subsequently recalled in the verbatim recall test than syntactic ones.

Secondly, are recall scores affected by linguistic category and translation condition together? That is to say, is there an interaction between these two independent variables? In simple terms, statistical interaction can be defined as the simultaneous influence of two independent variables on a third, the outcome. If there is no interaction, the effect on the outcome of each independent variable is completely separate from the other. When there is an interaction, the effects are not additive, i.e. the effect of one independent variable on the outcome depends on the value of another independent variable. Different linguistic levels are addressed in the post-tests, e.g. Italian nouns, adjectives, adverbs and syntactic structures. In her study about translational salience and noticing in standard subtitles (L2 audio – L1 subtitles), Ghia (2012a: 84-88) found a difference in recall rates between lexicon and syntactic patterns. Both lexical items and syntactic patterns were better recalled in the literal transfer condition, but recall of syntax was higher than lexicon in the non-literal condition, suggesting that diverging translation hinders the recall of vocabulary more than that of syntax. This may or may not hold true in the case of reversed subtitling, where the L2 is presented in writing rather than aurally, meaning that post-tests addressed input appearing in the subtitles rather than spoken. In Ghia's experiment, due to the small sample size (n = 13) and skewed distribution of the data, it was not possible to apply inferential statistics or explicitly test for interaction effect between linguistic category and translation condition on recall accuracy. An improvement in this sense is achieved by this study, where both t-tests and regression analysis between linguistic category and memory scores will be used to address this RQ.

RQ3. Are formal similarity and formal discrepancy registered differently by viewers? That is to say, do participants concentrate more on formally equivalent or discrepant, contrastive input? This question will use eye-tracking measures to analyse if the amount attention allocated to the subtitles changes according to translation condition, namely if more and/or longer fixations occur in the literal or non-literal condition. The distribution of looks will be analysed for the test items (22 subtitles) as well as for the whole experimental stimulus (110 subtitles). Moreover, how does visual attention relate to recall scores? In other words, if more or longer fixations occur on a subtitle, would its wording be better remembered? In psycholinguistics, fixation duration is understood to reflect language processing and be a proxy for the ease with which word meaning is accessed and integrated in the current sentence (Cop et al., 2015: 2). If a non-literal translation is perceived as more striking than a literal one, one would expect differences in attention allocation to be registered in the two conditions. Like Ghia (2012a) we assume that perceptual salience will arise from input translated non-literally and may in turn attract more and longer fixations. However, based on Ghia's post-test findings, we do not assume that perceptual salience and an increased attention allocation to a subtitle automatically result in a mnemonic superiority of the formal discrepancy condition. In fact, a subtitle could attract more and overall longer fixations in at least two situations: (a) if it contains a word or expression that is novel or interesting for the viewer or (b) if it contains a word or expression that is difficult to process, thus reflecting the above-mentioned ease (or lack thereof) of access and integration in the flow of information. Therefore, a non-literal rendering could attract more and longer looks and then be incorrectly recalled in the post-test, for instance if the meaning of the subtitle input is not clear to the viewer. Correlations and t-tests will be used to assess the relationship between recall scores and visual attention.

RQ4. What is the role played by relative frequency (3.14) in this study? This RQ will be explored by looking at a number of sub-questions. Firstly, the relationship between relative frequency of the subtitles and translation condition will be explored. Because translation condition is the main variable of interest, the experimental manipulation was carried out first and foremost following the principles outlined in the translation protocol (3.11). Consequently, frequencies of the two translation versions are necessarily calculated after the versions are created. For example, when a satisfactory literal Italian translation of the English source existed, it had to be chosen and kept regardless of the frequency of its words. While this is inevitable given the current experiment design, it could introduce a confound if frequency turned out to be not uniformly distributed within the subtitles. That is to say, for example, if all non-literal subtitles contained infrequent words, and these subtitles were found to be remembered significantly better in the recall test, one could not be sure of whether the effect was due to formal discrepancy or to the low frequency of the items. To ensure that any possible effects detected are really attributable to translation condition, relative frequency of the 22 test items was calculated and a t-test will be used to compare means in the literal and non-literal conditions. Secondly, frequency will be explored in relation to recall, to see if the latter is affected by the former. Research has shown that low-frequency items are better recalled than high-frequency items in word recognition tasks (Gorman, 1961) like the one that was used in our recall post-test. If translation condition did not have an impact on recall, but frequency did, one should find that low-frequency items are recalled more accurately than high-frequency items regardless of which translation condition they are in. Subjects should therefore perform accurately in the post-test with low-frequency items and vice versa, accuracy should be lower with high-frequency items. Thirdly, the relationship between frequency and eye movements will be analysed. Rayner et al. (1989) showed that both word frequency and length affect how long people fixate on that word during silent reading. When word length is kept constant, infrequent items are fixated for longer than frequent ones (Pollatsek et al., 2008). Given the clear frequency effect found in the literature on silent reading, (see review article by Rayner, 1998), we expect that eye-movements will vary depending on the relative frequency of the words in the subtitles, namely that subtitles containing rare, infrequent words will be fixated for longer. Both correlations and t-tests will be computed in the by-item analysis in order to assess to what extent relative frequency does affect the variables discussed herein.

RQ5. How is translation in this AVT mode perceived by the viewers themselves? Will they spontaneously acknowledge that the AV text is not uniformly translated, that there are some translation discrepancies? And if they do, will they be able to give examples? The explicit report task (questionnaire) will be analysed here. Viewers' metalinguistic awareness (overt noticing) of translation discrepancies is important to assess the validity of the literal vs. non-literal translation as a psychological construct and to investigate noticing. If viewers did explicitly report their noticing of a general L1-L2 contrast, this would suggest that viewing condition does matter, that translation is perceived differently depending on whether literal or non-literal outputs are presented. In his critical review of noticing research, Truscott (1998) distinguishes between awareness of input in a global sense and awareness of specific grammatical details within such input. If viewers overtly reported noticing L1-L2 discrepancies, empirical support to this distinction would be given, providing evidence that, even if a viewer might not recall specific linguistic instances correctly after watching, global awareness of input can be achieved in just one exposure. Moreover, if viewers did free recall some specific instances of L1-L2 discrepancy, this would suggest that for these items not only noticing definitely occurred, but also that it was "sufficient for the storage of information in memory" (Ghia, 2012a: 89).

For ease of reference, the RQs addressed above are also summarised below together with the corresponding variables.

	Research question	Variables analysed
1.	1. Are literal items (formal similarity condition) recalled more or less accurately than non-literal ones (formal discrepancy)?	➢ Translation Condition ➢ Recall Accuracy
2.	2a. Is there a difference in verbatim recall between lexical and syntactic items?	➢ Linguistic Category ➢ Recall Accuracy
	2b. Is there an interaction between lexical and syntactic items and translation condition?	➢ Translation Condition ➢ Linguistic Category ➢ Recall Accuracy
3.	3a. Are literal and non-literal translations registered differently by viewers?	➢ Translation Condition ➢ Fixation Duration ➢ Fixation Count
	3b. Is there a relationship between visual attention and recall rate?	➢ Fixation Duration ➢ Fixation Count ➢ Recall Accuracy
4.	4a. Is there a significant difference in frequency between literal and non-literal items?	➢ Translation Condition ➢ Relative Frequency
	4b. Is there a relationship between the frequency of an item and its recall rate?	➢ Relative Frequency ➢ Recall Accuracy
	4c. Is there a relationship between frequency and eye-movements?	➢ Fixation Duration ➢ Fixation Count ➢ Relative Frequency
5.	5. Will viewers report noticing translational differences between the source (audio) and the target (subtitles)?	➢ Open-ended questionnaire

Table 2. Research questions.

3.5 Experiment Design

This study adopts an experimental design where two viewing conditions are compared: literal translation (L) and non-literal or diverging translation (N). In line with Ghia (2012a), a 2x2 counterbalanced design was chosen. To test for the effects of translation condition, one could have two clip versions, one fully translated literally and the other non-literally, and assign participants to either of these conditions. However, doing so would mean that some participants do not experience literal transfer, while others do not experience non-literal transfer. This would not allow to see if there are changes within each participant depending on the experimental treatment, e.g. if literal or non-literal items are recalled more accurately, or looked at more and for longer, by the same participant. Moreover, characteristics of the video clip might affect performance, such as speech density or plot development. To make sure that any effects found are not specific to a part of the clip or to subject variables, the counterbalanced design was chosen, where each stimulus acts as its own control (Godfroid et al., 2010: 178). This method ensures that the behaviour of each subject is explored under both viewing conditions and that both video halves are subtitled in both conditions. To do this, the following standard 2x2 Latin square design was used:

L – N (Group G1L, Version 1)
Part 1: literal transfer – Part 2: non-literal transfer

Reverse subtitles

N – L (Group G2N, Version 2)
Part 1: non-literal transfer – Part 2: literal transfer

Participants in G1L watched the first half of the video (Part 1) in L (literal condition, i.e. formal similarity) and the second half (Part 2) in N (discrepancy condition, i.e. formal divergence), whereas participants in G2N received condition N first and then L. In this design, the two conditions L and N have repeated measures on each participant and the order is completely counterbalanced between subjects. Making multiple measurements (here, collecting data on several subtitles) for each subject "holds out the prospect of factoring out each [subject's] idiosyncratic (…) proclivities as part of the analysis, thereby improving the signal-to-noise ratio" (Barr et al., 2013: 256). Moreover, this design also allows to avoid systematic confounding effects of fatigue, practice or habituation that might arise from presenting either condition first. As Baayen and colleagues note: "By means of counterbalancing, adverse effects of learning and fatigue can be neutralized, in the sense that the risk of confounding these effects with critical predictors is reduced" (Baayen et al., 2008: 399).

The audiovisual text chosen was an excerpt from the Wachowski Brothers' film *The Matrix* (1999). The total length of the clip was 10 min 40 sec, during which a total of 133 Italian subtitles were presented on screen. Conventionally, the initial and final subtitles were kept the same for adaptation purposes (Ghia, 2012a), so a total of 110 subtitles were considered in the analysis, 55 in the first and 55 in the second half (see 3.7 for details). The recognition post-test was designed to capture potential differences in short-term verbatim memory stemming from discrepancies between the L1 audio and the L2 subtitles. It takes the form of an MCQ with questions on 22 subtitles (test items) in total, 11 appearing in the first and 11 in the second half of the clip (see 3.14). An open questionnaire was also administered after watching, to measure metalinguistic awareness of the above-mentioned discrepancies created through translation manipulation. Data were collected from 26 participants who were all English (L1) native speakers learning Italian (L2) at an upper intermediate level. They were randomly assigned to G1L or G2N at the beginning of the session, for a total of 13 participants per group. For a detailed description of what the participants were required to do during the experimental session, see section 3.16.

3.5.1 Pilot Experiment

As is customary in many psycholinguistic experiments, a pilot study was run in the initial phases of the research to gauge the appropriateness of the experimental design and stimuli. Any issues spotted in this initial stage make it possible to refine the procedure, thus improving the methodology of the full-scale study. A total of four participants took part in the pilot. They sat the proficiency test (3.9) first, and then watched the experimental subtitled video while their eye movements were being tracked. After watching, they answered the recognition post-test (3.12), followed by the open questionnaire (3.15).

After running this initial experiment, a pilot assessment phase determined a number of adjustments to the research design. First of all, for the reasons presented in 2.9, memory capacity was identified as another confound as crucial as language proficiency, so it was decided to add a WM test (3.10) to the main experiment. Secondly, during the pilot it was found that, with some participants, even if calibration was done correctly, eye-tracking during watching was not optimal and the scan path was slightly offset. This could obviously only be discovered when replaying the clip with scan path overlay, so *after* the participants watched the experimental video. If students could watch a short unrelated dummy clip after calibration but before the main experiment, this would help the researcher know if calibration settings (distance from the monitor, seatback tilt, sitting position and the like) for that individual person need to be adjusted, and more than one trial could be done with the same dummy clip until the setting producing the best gaze path accuracy is found. Thus, in the main study, a short (2 min) excerpt from the German-Austrian film *The Edukators* (2003) was introduced. This adjustment stemming from the pilot phase was very successful in minimising scan path offset and maximising the quality of eye-tracking data collected in the main study. Thirdly, after looking at the subjects' responses to the pilot questionnaire, it was decided to amend the wording of one of the questions before running the main study, namely question 3 (see Appendix A.4). Initially, it read as (a). In the main study, it was amended to (b).

(a) Were there any *new* Italian words or expressions in the subtitles *that you had not heard before*? If so, do you remember any of them?

(b) Did you find any Italian words or expressions in the subtitles somewhat striking, *either because you hadn't encountered them before or for any other reason*? If so, do you remember any of them?

The main reason for this correction (marked in italics above) is to provide a more comprehensive account of salience: participants should be prompted to list not only striking *new* words, but *any* striking words, even if already encountered before; it could be the case that, for example, a participant already had some level of knowledge of a word, yet it stuck in their minds because they thought it was not a particularly accurate translation of the source (noticing of a discrepancy). Moreover, the phrasing in (a) might dissuade participants from reporting words that they found striking and had not seen before but they did not consider fully unknown, because they are formally similar to the source (e.g. *obsoleto – obsolete*) and therefore semantically transparent.

The pilot also highlighted a difficulty in extracting eye-tracking data from the Tobii software. Due to the large size of the files generated, even the relatively 'small' export for the four pilot participants caused some issues, in that it took a considerable amount of time to open in Excel, often causing the programme to crash and making it inconvenient to export all data in one file. Thanks to data extraction trials in the pilot experiment, exporting individual files for each participant was identified as the most effective technique to be used in the main study (see 3.17.4 for details). Lastly, after considering by-subject variables (Italian language proficiency levels, WM and questionnaire answers), the pilot assessment phase also revealed that by-item variable (i.e. subtitle) confounds could be further reduced by controlling for frequency effects (2.11). This led to the calculation of relative frequency of the subtitles (3.14) and the inclusion of this variable in the main study analysis. Only descriptive analyses were run on pilot data, as no meaningful inferences can be made using inferential statistics on such a limited sample (2 participants per group). These proved nevertheless extremely useful for the researcher to practice data interrogation and graphical interpretation of the plotted variables. For obvious reasons of incompatibility, pilot data was not included in the main study. For a full account of what subjects were required to do in the main study following the pilot adjustments herein described, see 3.16.

3.6 Eye-tracker Specifications

Participants were recorded via a Tobii TX120 eye-tracker (Tobii Technology AB, Stockholm). This device has a sampling frequency of 120Hz, meaning it can collect samples from the viewers' eyes up to a maximum of 120 times per second (once every 8.3 ms). It uses the traditional Pupil Centre Corneal Reflection (PCCR) methods for remote gaze detection (Tobii, 2010: 4). It is a video-based, remote eye-tracker which allows relative freedom in terms of head movement, employs binocular eye-tracking (two camera sensors are used to film the eye), allows both dark and bright pupil tracking and has an accuracy of 0.5°. In all experimental sessions, it was placed in a central position just under the screen. The data collection software used is Tobii Studio (version 2.3.2 to 3.1.6). Data were collected for both eyes in all participants and the average was used to calculate gaze position. The event detection algorithm used to identify fixations was the standard Tobii Fixation Filter, which measures mean changes in the eye velocity to discriminate fixations from saccades. Minimum fixation duration was set to 100ms and fixation radius to 35 pixels (this indicates the maximum distance between raw gaze points to be classified as belonging to the same fixation). The velocity threshold for this particular set up with a 120Hz eye-tracker is 4.21 pixels/ms (this means that eye velocities below this value are grouped into fixations while those above it are considered saccades). The Tobii firmware automatically produces validity codes for each eye and every gaze data point. Validity codes are a way of assessing the system's level of confidence that it has recorded the correct data. They range from 0 to 4, where 0 corresponds to certainty that eye data were correctly recorded and 4 to gaze data missing or definitely incorrect. The Tobii Fixation Filter uses these validity codes to filter out invalid data automatically (i.e. any raw gaze point with a validity code of 2 or higher), which presents an advantage in terms of data post-processing and ensures that only eye data that the eye-tracker is confident about is included in the dataset for later analysis.

3.7 Selection Criteria for Video Material and Subtitle Preparation

A core principle followed in the choice of clip was that the excerpt should be selected from authentic sources. Authentic specifically refers to "materials that are originally produced in a given language for a native-speaking audience of that language, and not for learners of the language as a foreign language" (Garza, 1991: 241), as it is exactly these materials that the more proficient learner "often cites as the most desirable – yet most difficult – to comprehend" (ibid). These include all TV programmes, films, shorts, documentaries, news, commercials and so on. Authentic material was chosen as it provides "rich input environments" (Tschirner, 2001: 306), a complete communicative situation (Lonergan, 1989) and a window into the foreign culture (Sherman, 2003: 12), thus allowing L2 learning to occur in context. The filmic category was preferred first of all as filmic dialogues are the closest to everyday conversation (Pavesi and Perego, 2008b), and also because dialogue scripts and official subtitles are more readily available for this type of video material, which can help greatly during clip preparation. Moreover, films are motivating for learners, they use real language and they provide a variety of linguistic and paralinguistic features of oral conversation, thus fostering pragmatic competence (Lopriore and Ceruti, 2015: 301). The artificiality typical of this type of experimental studies was reduced as much as possible by using non-intrusive eye-tracking, arranging a comfortable set up for the viewer and choosing an example of real language-in-use rather than modified or adapted video input. *The Matrix* (1999) was selected for a number of reasons. First of all, the plot of this film is gripping, yet quite clear and straightforward (a computer hacker learns from a group of rebels the truth about the world he lives in, and joins the rebels in the war against the machines that control human society), meaning it is generally not too difficult to follow. Secondly, this action film also contains several lexically rich dialogue scenes, ensuring a balance between syntactic and lexical items, so that RQ2 (see 3.4) could be addressed appropriately. Thirdly, *The Matrix* contains much authentic, sought-after, everyday colloquial language, which increases participants' engagement with the video, while also diversifying the L2 input many of them received during their formal instruction, which usually mainly focuses on textbook-adapted language or literary and journalistic texts. The selection of the particular clip extracted (from 01h:07m:12s to 01h:17m:52s, for a total duration of 10m:40s) was based on situation and grammar appropriateness: the clip was self-contained (there was a clear start and end to the scene), the language used fitted well within the situation illustrated and contained a mix of grammatical and lexical features whose complexity was deemed appropriate for the level of the learners (B2).

A transcript of the English dialogues was translated using the professional software Softel Swift 6.0 to create and cue the subtitles (see Appendix B.1). The cueing was left the same, so that the in-time and out-time of each subtitle was kept constant for both clips regardless of the translation condition (in- and out-times are also reported in Appendix B.1). This ensured that allocation of visual attention to the screen was not affected by differences in subtitle permanence on screen, position relative to shot changes and flickering effects. The only element that changed between Version 1 and 2 (see 3.5) of the clip was the Italian translation contained in the captions. The minimum duration of a subtitle was 20fr (just under a second), while the maximum was 5sec 21fr (just under six seconds). During clip preparation, a number of minor issues arose. Compatibility between subtitling and eye-tracking software, subtitle position during the burning of the two sets of captions onto the video, and diacritics corruption (Italian letters with a graphical accent) were dealt with during this initial phase. To control for confounds relative to learners' reading speed and reading time, all subtitles were between 1-13 words long, ranging between 2-70 characters maximum, and characters per line did not exceed 37. Although in the subtitling industry the reading speed can go up to 190-200 WPM, for example for the DVD market, in this experiment the maximum reading speed settings were kept slightly lower, at 180 WPM, which roughly corresponds to 15 CPS (Hefer, 2013). This is because L2 reading is typically slower than L1 reading and also because the clip was played only once to the participants, who therefore only had one chance to read the L2 reverse subtitles. Subtitles appeared in a centre-aligned position at the bottom of the screen in both translation conditions. The video clip had a standard film aspect ratio of 21:9. Played on the eye-tracker computer monitor (Size: 38 x 30.5cm) with a resolution of 1024 x 768 pixels, the film therefore presented a black letterbox at the top and bottom of the screen, which measured 7cm at both ends (see fig. 1).

Figure 1. Example of the video stimulus with letterbox, as seen by participants during the recording.

The subtitles were placed in this letterbox area, which makes the identification of relevant eye-movements to target even clearer, as the text is clearly positioned at the bottom of the screen on a black background, not overlapping the video image. All subtitles were presented in blocks (rather than scrolled), comprised between one and two lines maximum, and were segmented in the same way between literal and non-literal renderings. That is to say, text chunking, or "the grouping of a block of text into coherent segments" (Rajendran et al., 2011: 6) was achieved in two-liners by means of placing the line breaks at grammatical and logical points in the sentences and using punctuation based segmentation (ibid.). Noun, adjective, prepositional and verb phrases were kept together as to produce these homogenous chunks wherever possible, following recommendations in Perego et al. (2010: 252), so that segmentation coincided with the highest syntactic node possible (Karamitroglu, 1998). 24% of the subtitles were two-liners (n = 26) and 76% were one-liners (n = 84). To keep the two translation conditions comparable, the two renderings differed by one (34.5%), two (12%) or three (4.5%) words maximum. In several cases (48% of the total), it was possible to manipulate the translation while keeping the same number of words in both literal and non-literal renderings (see section 4.3.3.2 for further details). The full transcript with the corresponding literal and non-literal subtitle translations is included in Appendix B.1.

3.7.1 Copyright

Choosing *The Matrix* as experimental stimulus bears an intrinsic question of how potential copyright issues are addressed. To answer this question, different pieces of legislation were considered, namely the Berne Convention, the Directive 2001/29/EC of the European Parliament and of the Council of May 22nd, 2001, as well as the 1988 Copyright Designs and Patents Act (UK). The present project was designed and implemented in the UK, which, at the time of research implementation and dissemination, is a full member of the EU.

Where copyright law varies between European states, the Berne Convention (signed by EU member states, including the UK) establishes a common framework in respect to intellectual property rights. For reference, a full text of the Berne Convention is provided by the World Intellectual Property Organization (WIPO, 2017). The Berne Convention allows certain exceptions on economic rights of copyrighted material, i.e. cases in which protected material may be used without the authorization of the copyright holder, and without compensation. On these exceptions, Art. 9(2) states that:

> *It shall be a matter for legislation in the countries of the Union to permit the reproduction of such works in certain special cases, provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author.*
> Berne Convention, Art. 9(2) (reproduction in certain special cases)

Directive 2001/29/EC is also relevant to the present work, as it provides an attempt of the EU to harmonise certain aspects of copyrights in today's information society for its member states. Art. 5(3) states that member states can allow some exceptions to the copyrights outlined in previous articles, for example in case of:

> *(a) use for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved.*
> Directive 2001/29/EC, Art. 5(3), clause (a)

Given what is stated in Art. 9(2) of the Berne convention, UK legislation needs to be taken into account. The UK implemented EU relevant Directives through the United Kingdom Copyright, Designs and Patents Act (1988), as amended on December 31st,

2003. In copyright law, what is outlined in Directive 2001/29/EC corresponds to the concept of fair use or fair dealing, which describes some activities that are allowed without infringing the rights of copyright holders. The use of *The Matrix* in this piece of work is fair dealing as defined under the 1988 Copyright Designs and Patents Act (UK). In this legislation, fair dealing covers the limited use of copyrighted material for various purposes, such as (i) research and private study, (ii) instruction or examination, (iii) criticism or review and (iv) news reporting, amongst others (UK Copyright Service, 2017). The present piece of work is undertaken for the sole purposes of research and therefore falls under (i). More specifically, copying parts of a copyrighted work in such cases is allowed if:

- ✓ The copy is made for the purposes of research or private study.
- ✓ The copy is made for non-commercial purposes.
- ✓ The source of the material is acknowledged.
- ✓ The person making the copy does not make copies of the material available for a number of people.
(ibid.)

In this piece of research, a short excerpt of the film was extracted purely for the purposes of creating the two translation versions to be tested experimentally (see 3.5), and the source of such excerpt has been fully acknowledged. No monetary gains were made as a result of using this short clip in the laboratory, the clip is not to be distributed commercially, and no copies of this material were made available for any number of people after viewing. The way in which this research project uses copyrighted video therefore meets all the criteria specified above and is to be considered fair dealing as defined in the 1988 Copyright Designs and Patents Act (from Section 28 onwards).

3.8 Participants and Recruitment

All the participants recruited were native English speakers, aged 20-35 (M = 23.8; SD = 3.9) with an upper-intermediate level of Italian (CEFR B2) and normal or corrected-to-normal vision. The gender ratio was 18:8 (69% female, 31% male), which reflects the bias towards female gender typical of foreign language learning demographics. Initially the age range was set to 18-25 years old, as the idea was to recruit only BA students of Italian in their last year of study (module ITAL 3010). However, in practice this was not possible as after a first round of advertising no students volunteered. By widening the baseline to 18-35 years and introducing a language proficiency test, it was possible to recruit a total of 26 participants over the course of two academic years. An age limit was introduced for a number of reasons. First of all, to avoid the well-established cognitive decline naturally occurring in older subjects, whose reduced short-term and working memory (Williams and Kemper, 2010) may easily affect recall scores as well as attention allocation. Secondly, because eye-movements strategies are different in older adults, who adopt special reading strategies to compensate for slower lexical access (Rayner et al., 2006).

The study was advertised via dissemination through various mailing lists, direct class contact and putting up notices in the relevant departments and throughout the university. Lecturers for the relevant courses also encouraged their students to take part. Face-to-face interaction proved to be the most effective recruitment method. Unfortunately it was not possible to pay participants a monetary reward, which may have contributed to the initial lower recruitment levels. However, changing the content and format of the advertisement helped towards recruitment: more and clearer information about why taking part would be useful was included, the session was presented as an occasion to practice Italian and participants were given Italian biscuits and sweets as a sign of gratitude, rather than printer credits. Individual feedback was also offered on the results of the proficiency test, if desired. A copy of the call for participants containing all this information was given to each of the prospective participants, and email addresses were collected *in situ* (e.g. in class). These procedures enhanced participation rate amongst university students, meaning that the vast majority of subjects in the study were either students, recent graduates or more mature individuals still in higher education. A total of 33 participants were recruited, but seven subjects were excluded from the experiment because they did not pass the language proficiency test, possessed too high a working memory, had an eye condition, turned out to be bilingual from birth and finally because the eye-tracker did not satisfactorily record their eye movements. Of the 26 participants who fit the recruitment criteria, some were BA students of Italian in their final-year, others level 2A (year two, advanced Italian), some were MA students and some young professionals who had graduated less than five years before the experiment took place and used Italian on a weekly basis. All participants had been studying Italian for at least four years and passed the proficiency test (3.9) before taking part in the study. The relative variety of these demographics means the sample analysed is not limited to university students ages 20-23 but is more representative of young adults in general.

3.9. Language Pre-test

Together with age, prior language experience is one of the main factors bound to affect acquisition (Wood Bowden et al., 2005: 105). Alongside controlling for age (see 3.8) individual variations in language proficiency were controlled for through a written pre-test (see Appendix A.1). CEFR B2 was identified as the level of Italian necessary to understand the non-specialised colloquial language used in the video excerpt. The pre-test comprised four exercises taken from past B2 upper-intermediate exam papers from CILS and from the University of Verona's Language Centre (Centro Linguistico di Ateneo, CLA). It was decided to extract the exercises from official tests rather than devise a new one, as the former have previously been tested for validity, reliability and appropriateness through item analysis, inter- and intra-raters reliability (Barni et al., 2009: 19).

The Certification of Italian as a Foreign Language (Certificazione di Italiano come Lingua Straniera, CILS) is the only language qualification for foreign speakers of Italian officially recognized by the Italian Government. It was initially developed by the University for Foreigners of Siena and is now a requirement in most Italian universities. The various levels of language proficiency identified within CILS have been standardised and successfully achieved CEFR-equivalence. Moreover, CILS is part of EALTA (European Association for Language Testing and Assessment). Test items were picked from past B2 level exams. Italian B2 CILS exams are 4 hours in total and encompass five core language areas (reading comprehension, written production, oral production, listening, grammatical competence). Not all areas were relevant to this study (e.g. listening comprehension and speaking skills are not applicable given the AVT mode chosen and the nature of the study), therefore only reading comprehension and grammatical competence were considered, and only the relevant exercises were extracted and adapted. Vocabulary, grammar structures and written comprehension were tested in the pre-test. In B2 CILS exams each of the 5 areas of ability is tested on a 0-20 scale for a total of 100 scores. In the CILS guidelines, the minimum pass score in each ability is set to 11/20, i.e. 55%. Subjects need to reach this threshold in each ability to pass the exam. In CLA exams, the minimum pass score for each ability is 60/100. For consistency with the two institutions' scoring methods, a 60% threshold was established in this study to discriminate between a pass and a fail. Therefore, a 60% mark in each of the two parts (part 1 extracted from CILS, 2 exercises; part 2 extracted from CLA, 2 exercises) had to be achieved for the subject to pass the test. The two exercises taken from CLA B2 upper-intermediate level were analysed against the specific and very detailed B2 guidelines issued by CILS (Barni et al., 2009) to make sure that the Verona's B2 level was consistent with the official CILS protocol. Once assembled, the pre-test was also sent to the Italian Department and double-checked by Italian native lecturers to make sure the language was representative of the genre (the CILS exercises came from magazine articles on a variety of contemporary topics) and appropriate for the level of their students. The test was self-paced and typically took between 20-30 minutes to complete. All 26 participants passed the test before experimental clip viewing (mean percentage correct scores = 83%).

3.10 Working Memory Test

As we have seen in section 2.9, an aptitudinal factor which is strongly related to language is working memory (WM). In this study, WM could be a confound as variance in post-test scores could be determined not only by translation condition but also by individual differences in this functional mnemonic ability. Therefore, after an initial piloting phase, a WM control test was introduced (see Appendix A.2). In the published literature, a variety of tests have been adopted as an operationalisation of WM ability. In our experiment, the reading span task (Daneman and Carpenter, 1980) was chosen, since it is taps both the storage and processing functions of WM (van den Noort et al., 2008: 35). The Reading Span Task (RST) is one of the most commonly used measure to assess verbal working memory (Farmer et al., 2012: 356) and it has proven to be both a reliable and valid measure of WM ability (Conway et al., 2005: 769). Moreover, the RST was chosen because it has been repetitively found to be a better predictor of language processing and reading comprehension than other span measures that only tap into memory storage abilities, such as the simple word or digit span (Daneman and Merikle, 1996). During this task, participants read aloud sets of sentences while simultaneously having to retain the final word of each sentence. At the end of each set, they try to recall these sentence-final words. The sets start from a minimum of two sentences and grow in size, so that the final set can contain up to five or six sentences. The number of final words subjects have to retain thus grows, increasing the processing effort required. Many variants have been used in the literature after Daneman and Carpenter's first version. Typically, later variants also included some form of comprehension check (see Daneman and Merikle, 1996) for example a true/false, a sensibility or a grammaticality judgement to be completed immediately after reading each sentence, to ensure that subjects are processing the sentences for meaning.

To be consistent with the published literature on WM, rather than creating new stimulus sentences, the RST adopted here used the same 42 test sentences used by Harrington and Sawyer (1992), who were among the first to investigate the relationship between WM and L2 reading specifically. Each of the 42 sentences had a relatively simple structure, an active voice, and was 11-13 words long (see Appendix A.2.2). Set size increased from two to five sentences. Each set size was administered three times (as to obtain the following 'set size x times' combinations: 2x3, 3x3, 4x3, 5x3 = 42). Some training sets (see Appendix A.2.1) were practiced before the WM experiment proper to allow the subjects to familiarise themselves with the task. Immediately after reading each sentence in each set, a grammaticality judgement task was undertaken, where participants had to say whether the sentence they had just read was a legal statement in the language (Turner and Engle, 1989). Half of the sentences were grammatical (i.e. they made sense both semantically and syntactically), half ungrammatical. The latter were obtained by reshuffling words before the final word in a sentence, as to obtain an illegal word order. Below is an example of a grammatical sentence (a) and its ungrammatical counterpart (b).

(a) The woman screamed and slapped the old man in the face.
(b) The woman screamed and the old slapped in the man face.

Each single sentence was printed and presented on a laminated index card. Subjects took the test individually and were asked to read aloud each sentence at their normal pace, without backtracking. After reading each sentence, they wrote on a test sheet whether the sentence was grammatical or ungrammatical (see Appendix A.2.4). The end of each set was indicated by a blank index card, at the sight of which subjects could turn the sheet and write at the back as many sentence-final words as they remembered, in whichever order they preferred. Sentence-final words were concrete and not semantically related. The experimenter timed the subjects while reading all sets to make sure that they had comparable reading speeds (see Appendix A.2.5). For the WM test to be considered valid, at least 85% grammaticality answers had to be correct (see Conway et al., 2005: 775), to ensure that the subjects were paying attention to the content of the sentences they were reading. In practice, this meant six misjudgements maximum on 42 sentences were allowed. The WM score for each subject was given by the total number of sentence-final words recalled. All subjects had to pass the WM test to be included in the main study (for the complete set of test instructions given to the subjects, see Appendix A.2.3).

The RST was carried out in the participants' L1 (English) rather than L2 (Italian). The reasons for this choice are manifold. First of all, in previous studies where both L2 and

L1 WM were tested, no difference was found in performance between the two (Keijzer, 2013; Harrington and Sawyer, 1992; Osaka and Osaka, 1992), suggesting that WM ability may be language independent. Secondly, subjects' WM performance in Italian would have been confounded by their L2 proficiency level, and this applies even to intermediate and advanced L2 users (Harrington and Sawyer, 1992: 28). If, for example, the sentences contained Italian words unknown to the subjects, this is likely to affect processing, so that one could not be sure of whether differences in reading performance are genuinely due to a lower WM span or to their insufficient level of Italian. Thirdly, grammaticality performance would be naturally lower in the foreign language, since L2 learners cannot be assumed to possess a native-like ability to recognise L2 sentences as acceptable. Given that the vast majority of the sentences (typically 85%) has to be assessed correctly to ensure that a subject is processing the sentences for meaning, lower performance in this part of the test could mean having to discard the entire dataset for some subjects not based on their WM but purely because they cannot assess Italian grammaticality like a native speaker.

As for the test results assessment, we used partial-credit scoring instead of all-or-nothing scoring, as the latter has decreased sensitivity and empirical results found the former to be superior (for a thorough explanation with scoring examples, see Conway et al., 2005). In all-or-nothing scoring, only the maximum number of remembered words per set is counted, and the count stops when a subject recalls correctly less than two thirds of a given set size. In partial-credit scoring, on the other hand, all remembered words across sets are taken into account, which gives more precise information about the performance of a subject across the whole RST (van den Noort et al., 2008: 38). Thus, in this study a composite score was calculated based on the proportion of correctly recalled words in each set. If all words in a set are recalled correctly, that set is awarded a score of 1. If, for example, three words out of four are correctly recalled in a set of size four, the score awarded for that set is 0.75. The possible scores range between 0 and 12, since there are 4 x 3 sets in total. The RST scores recorded in this study ranged between 8.24 and 10.95. One participant had to be excluded from the analysis because she had an unusually high WM, recalling all sentence-final words but two in the RST.

3.11 Translation Protocol

This section describes the rationale for the operationalisation of the independent variable. To wit, the criteria used in creating the formal similarity and formal discrepancy conditions will be outlined. Operationalisation is hereby defined as an attempt to render the variable under investigation quantifiable and measurable along a scale, in order to look at its effects on perception and memory under the two different conditions of exposure. Challenges to this process arise from a number of considerations. First of all, there is not a one-to-one relationship between form and meaning across languages (Baker, 1992), as one meaning could be expressed by one single word in one language but several orthographic forms in another, as in the two examples below (adapted from Baker, 2011: 10):

type (EN) - *battere a macchina* (IT)

if it is cheap (EN) - *yasukattara* (JA)

Moreover, words have vague boundaries and fuzzy edges (Anderman and Rodgers, 1996: 17), they do not so readily fit in boxes because meanings are usually negotiable and their realisation depends on the specific context (Baker, 1992: 15). Therefore, complete absolute equivalence does not exist, or at the very least is extremely rare, and this holds true not only at word-level, but at the broader grammatical, textual and pragmatic levels too, as well as for the ideas behind the words themselves. Concepts given in a source culture can differ radically from those of the target if the contexts in which words occur are not shared between the languages. As Culler (1976: 21-22) put it: "Each language articulates the world differently. Languages do not simply name existing categories, they articulate their own." Therefore, translation is always a process of interpretation involving a degree of choice and decision (Hervey et al., 2000: 41). Before producing any translation, the source text was explored in detail according to the following textual variables (Harvey et al. 2000):

- Phonic-graphic
- Prosodic
- Grammatical
- Sentential
- Discourse
- Inter-textual

Each string was analysed individually for each of these six discrete levels, and then globally in terms of interaction between levels. This helped both clarifying what parts of the source (and features within them) were most complex and identifying the locus of potential translation loss in the target. The semiotically rich environment typical of this multimodal medium was also considered, where para-linguistic features in the source audio (e.g. intonation, stress, pitch, and the like) can contribute to conveying the overall meaning of an utterance, potentially implying the need for compensation in the subtitles. Take, for example, the following audio segment, uttered by Morpheus to Neo outside the front door of the oracle's place:

"I told you I can only show you the door. *You* have to walk through it."

Morpheus stresses the word 'you' purely through intonation. This is a reference to a previous point in the film (the combat scene in which the captain and Neo engage in a fight in a simulated training environment) where Morpheus said almost exactly the same thing. If one were to disregard the prosodic level and translate just the words, they might lose the nuance contained in the sentence. A possible way of retaining it in the Italian translation could be to invert word order through a cleft sentence:

a. *Sei tu che la devi attraversare* [It is you who has to walk through it] rather than

b. *Tu la devi attraversare* [You have to walk through it]

Moreover, the purpose of the ST was considered. If translation can be defined as "an act of communication which attempts to relay, across cultural and linguistic boundaries, another act of communication" (Hatim and Mason, 1997: 1), what is the purpose of this communication? What is the intended meaning of the utterance for the SL audience? Answering these questions for each ST segment helped identifying possible ways of rendering them in the subtitles. The purpose of the TT was also of crucial importance and affected linguistic choices. The fact that translation needed to be operationalised, and that literal and non-literal renderings needed to be clearly distinguished, meant that the rationale behind the translational choices at points had to differ from what a translator (including the author herself) would have produced in a work environment, at least as far as the literal transfer condition was concerned. In other words, the purpose of the study posed some constraints to the degree of translational freedom allowed, because the goal was not to produce a single optimal native-sounding translation, as one would do in a professional environment, but to create two versions, one of which should be as close to the original as possible, both in terms of form and meaning.

First and foremost, the analysis process concentrated on identifying the textual levels of formal and semantic equivalence (or non-equivalence) between source and target for a particular string of text. To do this, a contrastive stylistic analysis of SL and TL was carried out. The ST presented a variety of features and parts of speech pertaining to both morphosyntax and lexicon. For each SL segment two subtitles were created, and a series of translation techniques were exploited, including transposition, modulation, adaptation and compensation (for definitions and examples see Vinay and Darbelnet, 1995 and Harvey at al., 2000). Unsurprisingly, the most problematic part of this process was often to create a translation that was as literal a rendering as possible without violating Italian grammatical rules. This is *in primis* because, as we have seen, complete identity of surface form does not exist, as there will always be phonic and graphic loss (different spelling, different numbers of letters hence different word length, different pronunciation of the same letter clusters, and so forth). Even cognates (e.g. EN: *impossible*/IT: *impossibile*), typically similar in form, are not exactly the same graphically and phonically. An exception to this are borrowings, where surface form is identical in L1 and L2 because one of the two languages borrowed the word from the other. It is worth pointing out, however, that in the case of borrowings other issues arise, e.g. often imported words can change meaning in the target culture, acquire a different connotation or refer to something different altogether. Consider the phrase *al fresco*, which can be used in English pubs or cafés to indicate the presence of an outdoor area such as a courtyard, roof terrace or beer garden. The expression is taken from Italian, where, however, this connotation is absent. Apart from the general meaning "to be in a cool, chilly place", the phrase is more often used as a colloquialism referring to being in jail (Italian prisons cells are, allegedly, chilly). So *essere al fresco* roughly corresponds to the English expression *doing time*. In no Italian bar, café or restaurant one will find a sign on the wall referring to an outdoor area which uses this expression.

Owing to typological differences between Italian and English, literal transfer, in particular of syntactic structures, is not always possible (Ghia, 2012a: 108). The analysis of the ST made it therefore clear that literal transfer (Gottlieb, 1992) in the formal similarity condition would have not meant total identity of both form *and* meaning. Given that absolute identity of form does not exist, the term 'equivalence' has been used in Translation Theory and will be used throughout the present discussion. It is beyond the scope of this work to address or review the theoretical argumentations regarding the use of this term in the literature, however, for an introduction to the discussion in Translation Studies, see Snell-Hornby (1988), Koller (1995) and Nida (1964). Equivalence in the present study does not refer to any specific translation theory but is used as intended by Harvey et al., namely "in its everyday sense of counterpart – something different, but with points of resemblance in relevant aspects" (2000: 19). In particular, equivalence of form and meaning were addressed. Having two expressions where both equivalence of surface form *and* meaning occur is rare, yet possible. For example, *impossibile* cannot be said to be identical to *impossible*, but can be said to be formally equivalent, and the Italian *sì* can be seen as functionally equivalent to the English *yes*, despite the lack of phonic and graphic identity. However, context needs to be borne in mind here, as a certain L2 word, graphically very similar to its L1 counterpart (e.g. cognates), may happen to have the same meaning as in L1 only when used in certain occasions, or may even have a completely different meaning from its L1 counterpart (this is the case with false friends, such as *factory – fattoria* [farm], *rumour – rumore* [noise], *canteen – cantina* [cellar] and the like). Most of the times, however, especially if two languages are not too far apart in terms of culture, there will be two structures (L1 source and L2 target) which are different in form but equivalent in function, i.e. they convey the same message. For example, question tags do not exist as such in Italian. However, there are a number of ways to reinforce a positive or negative statement in the same way as question tags do (e.g. *no?*, *non trovi?* [don't you think?], *vero?* [right?]). These translational routines do not share surface form but fulfil the same role and are used to create exactly the same effect in the interlocutor, hence they can be considered functionally equivalent to question tags.

3.11.1 What Counts as Literal and What Counts as Non-literal?

For the reasons outlined above, the concept of 'literalness' itself, if considered in absolute terms, does not provide any meaningful addition to the discussion. It was found to be much more useful to treat it as a scale (see Newmark, 1981: 39) ranging from verbatim to free translation, where distance from complete formal equivalence is a matter of degrees. Literalness was considered both in terms of surface form and semantic content for each ST utterance translated.

3.11.2 Literal Transfer

Generally speaking, a word-for-word translation was chosen whenever possible, that is to say wherever an expression equivalent in surface form as well as in meaning existed in Italian. This is the case with subtitle 18, 21 and 15:

Subtitle 18	
Yes.	L[18]: Sì.

Subtitle 21	
What?	L: Cosa?

Subtitle 15	
But an oracle can?	L: Ma un oracolo può?

Whenever a word-for-word translation in the TT was possible, even if the term is not as common as the ST, this was chosen as the literal counterpart in order to maximise clarity in the literal vs. non-literal distinction. This is the case with subtitle 16:

Subtitle 16	
That's different.	L: È differente. [It's different.]

Differente is a perfectly acceptable translation for *different*, although *diverso* also has the same meaning, occurs more often in spoken and written standard Italian, and would probably qualify as a more idiomatic and common rendering than *differente*, which has a higher degree of source bias. It is exactly this SL influence, however, that the present study addresses, in order to establish whether the closer the level of formal literalness between two terms simultaneously conveyed through different channels (source audio dialogue and target written subtitle) the more (or the less) memorable the target term is. *Differente* is more literal a rendering than *diverso* because the former is both formally and semantically equivalent to the English *different*, whereas the latter only shares equivalence of meaning. Therefore, *differente* was assigned to the literal translation condition, whilst *diverso* to the non-literal.

18 In examples taken from the subtitles, L will be used to indicate that the translation belongs to the literal transfer condition, N to the non-literal transfer condition. For the full transcript with aligned English source, L and N Italian translations, see Appendix B.1.

Sometimes full structural identity was not possible, but some aspects of word order could be retained. In these cases the author always prioritised faithfulness to the ST structure and tried to mimic the syntactic pattern of the original as much as the language allowed, such as in subtitle 40.

Subtitle 40	
You're right on time	L: Sei in perfetto orario. [You are in perfect time.[19]]

Italian simply does not have such a construction with the adverb *right*. However, *to be on time* translates literally as *essere in orario*, so this structure could be kept. The qualifier had to be moved from adverbial into adjectival position and was rendered with a collocate of *in orario*. This translation is fully functionally equivalent and much closer to the original in terms of syntactic structure than other alternatives like *sei davvero puntuale* [you're really punctual].

Wherever more than one word existed in Italian, none of which was transparent (i.e. clearly coming from the same root or sharing surface structure) the most commonly occurring was chosen. Consider the word *enough* in the following:

Subtitle 31	
She would say she knows enough.	L: Lei direbbe che ne sa abbastanza. [She would say that (she) knows enough about this.]

In this translation, this word could be rendered as: *abbastanza*, *a sufficienza* and *quanto basta,* all of which would fit the context, express the same communicative meaning, and be grammatically correct. *Abbastanza* was chosen because it is the first and most common

[19] In the present discussions, interlinear back-translations (Hervey et al., 2000: 15) into English will be provided in brackets whenever mirroring Italian grammatical units as closely as possible is needed in order to lay out the specific linguistic differences between these two languages. Such translations might not fully respect SL grammar but they will help explaining the diverging behaviours (syntactic in particular) between English and Italian addressed in each occasion. In all other cases, back-translations will provide a literal translation of the Italian but retain English grammaticality.

translation of *enough* (Sansoni, 2016) and because it allows the concept to be expressed in one word like in the source rather than two. Another instance is given below:

Subtitle 92	
'But' what?	L: 'Ma' cosa? ['But' what?]

In subtitle 92, both *ma* or *però* could have been chosen, as they both mean *but* and none of them is formally more similar to the source than the other. The two Italian terms differ in use and sometimes position in the sentence, but here both could equally fit. However, *ma* is by far the most frequently used in common language, both orally and in writing. Learners of Italian are likely to know both but will have come across *ma* much more often. Moreover, *ma* is a short and monosyllabic word just like *but*, which allows to keep a prosodic resemblance with the original.

Often, translation strategies had to be applied to produce a literal translation. The most common was grammatical transposition, defined as "the replacement or reinforcements of given parts of speech in the ST with other parts of speech in the target text" (Hervey et al. 2000: 15) without changing the meaning of the message. Due to the in-built differences in the way languages convey meanings and articulate concepts, transposition is often a compulsory procedure in translation if grammaticality in the TL is to be retained. An example is provided below.

Subtitle 57	
Smell good, don't they?	L: Hanno un buon profumo, no? [(They) have a good smell, no?]

In this case the English verb+adverb construction *to smell good* cannot be transferred literally or it would result in a syntactic violation. In Italian the same message is conveyed through the noun+adjective construction *avere un buon profumo* [to have a good smell], which therefore constitutes the closest Italian equivalent to the English source. The question tag has been rendered as *no?* rather than the longer *non è vero?* because of space constraints in the subtitle. It should be also noted that, in this very case, formal equivalence is intensified as both versions omit the subject. This is made possible by the nature of the scene, where the oracle uses low register words and colloquialisms (e.g.

kiddo, fate crap, balls to bones), idiomatic expressions (through and through, to bake one's noodle, to feel right as rain), and omits the subject in four occasions (subtitle 55, 56, 57, 80). Normally such omission does not occur, as English requires the subject to be overtly stated, whereas Italian, being a pro-drop language, systematically omits the subject unless it is needed to avoid ambiguity, provide emphasis, or for contrastive purposes. Another instance of grammatical transposition is given in subtitle 47:

Subtitle 47	
There is no spoon.	L: Non c'è nessun cucchiaio. [There isn't no spoon]

Here the literal rendering is as close as it can be to the source without violating Italian structural norms. Whilst Italian, like French or Russian, happily accepts double negation, English does not, so the above utterance could be rendered by either *there is no spoon* or *there isn't any spoon*. Constructions such as **there isn't no spoon* are ungrammatical in English, but they need to be used in Italian. The other only translation allowing to keep this very structure would have been (i) *non c'è alcun cucchiaio* [there isn't any spoon], however this option was not chosen because *alcun* is much less commonly used than *nessuno* in every-day language. Had the ST been *there isn't any spoon* rather than *there is no spoon*, then of course (i) would have been chosen as word-for-word counterpart.

3.11.3 Non-literal Transfer

Once the closest formal equivalent was identified, formal discrepancy was achieved through a second, freer translation (see Appendix B.1). Different strategies were used depending on whether lexicon or syntax was concerned. An example of lexical divergence occurs in subtitle 25 and 29:

Subtitle 25	
…the prophecy?	L: la profezia? [the prophecy]
	N: il vaticinio? [the vaticination]

Many Italian options could have been chosen as the non-literal counterpart: *divinizzazione* [divination], *oracolo* [oracle], *presagio* [presage], *previsione* [prevision], *pronostico* [prognostication] and so on. The focus was on picking a word that would not necessarily meet the viewer's expectations, causing them to be almost surprised and consciously notice the discrepancy between the subtitle and the aural input *prophecy* (for which they might even know, at this level of proficiency, that the first and most common Italian translation is the transparent term *profezia*). Before choosing *vaticinio*, the other options were weighed: *oracolo* cannot be used because it appears in the same sentence and it is used in another connotation (the person that prophesises rather than the prophecy or the place of divination) throughout the film; *divinizzazione* was too long to fit the limited space available in the subtitle; *previsione* would have sounded awkward used alone at the end of the sentence, as it can sometimes be used as a synonym of *profezia*, but it is normally used in different contexts, (e.g. *previsione metereologica*, weather forecast) has different collocates and a different range of literal meanings[20].

Subtitle 29	
Of the resistance.	L: Della resistenza. [of the resistance]
	N: Della guerriglia. [of the guerrilla]

Similarly, in subtitle 29, formal equivalence in the literal transfer condition is achieved with the pair *resistance* – *resistenza*, and it is the non-literal condition that presents a challenge. In this and the previous case, there simply is not a second word which maintains sematico-functional equivalence whilst diverging in surface form, so whichever word one chooses as the non-literal counterpart of that source string, there will be differences in meaning between the two. This does not detract from the translation (in fact, it does enhance the probability that the viewer will consciously notice the item and possibly remember it at a later stage) if the two terms maintain a degree of semantic overlapping and the non-literal target fits in the context of the utterance. *Guerriglia* is of course not in full synonymic relation with *resistenza*, but fits this situation: Morpheus uses *resistance* in the sense of a movement of opposition against the ruling power, referring to the small group of humans he is leading in the struggle against the overwhelming power of the machines, who took over the earth and enslaved humanity. In this sense, it is quite close to the meaning of guerrilla: "An irregular war carried on by small bodies of men acting independently" (OED, 2013).

[20] For an introduction to literal meanings and semantic overlapping, see Hervey et al., 2000: 85-91.

Subtitle 34	
She is a guide, Neo.	L: Lei è una guida, Neo. [She is a guide, Neo.]
	N: Ti farà da guida, Neo. [She will be your guide, Neo.]

Syntactic divergence was achieved in subtitle 34, where *fare da guida* [act/serve as a guide] allows to keep the central focus of the utterance – the noun *guida* – in both versions, maintain a similar length and rhythm, and has the same number of words of its literal counterpart.

Modulation was also employed in the translation. Vinay and Darbelnet (1995: 36) describe it as "a variation in the form of the message, obtained by a change in the point of view". Two examples are given. The modulation occurs in N in both cases.

Subtitle 19	
What did she tell you?	L: Cosa ti ha detto? [What did she tell you?]
	N: Cosa vi siete dette? [What did you tell each other?]

Subtitle 56	
Almost done.	L: Sono quasi pronti. [(They) are almost ready.]
	N: Ci siamo quasi. [We're almost there.]

In example 19, the subject changes from *lei* (she) to *voi* (you plural) whilst keeping the main verb *dire* and maintaining the same syntactic structure with *cosa* at the beginning. In example 56, the point of view is changed completely. In this scene, the oracle is talking to Neo while pottering in the kitchen. Although she omits the subject, it is clear from the image that she is referring to what is being baked (cookies), hence the subject of the English sentence is *they*. In the non-literal translation, a more idiomatic sentence with the verb *esserci* is used, causing the subject to change from *they* to *we* in Italian.

Another translation strategy adopted was compensation, a mitigation technique to reduce translation loss. In Hervey et al.'s words: "where any conventional translation (however literal or free) would entail an unacceptable translation loss, this loss is mitigated by deliberately introducing a less unacceptable one, important ST effects being approximated in the TT through means others than those used in the ST" (2000: 212). Let us consider subtitle 50.

Subtitle 50	
The oracle will see you now.	L: L'oracolo è pronto a vederti. [The oracle is ready to see you.]

Here, the word-by-word translation **L'oracolo ti vedrà ora* would sound awkward due to the simple future immediately followed by the adverb *ora*, as well as the use of the verb *to see* in this particular context ("to receive someone as a visitor", rather than "to perceive with the eyes"), which can be used in Italian (e.g. *il dottore deve vedere un paziente* [the doctor has to see a patient]), but not in the case of this particular verbatim rendering. The English future as used here has a volitive nuance, in that it indicates the intention on the part of the subject to carry out an action. In Italian, some uses of the future are increasingly being replaced by other tenses such as *presente* and *passato prossimo* (Berruto, 1998: 70). The temporal, modal and aspectual values of the future tend to be realised through means other than the simple future itself, for example through periphrastic structures (De Roberto, 2010). In particular, the so-called *presente pro futuro* (the use of the present tense instead of the simple future to indicate a future action) is often employed when there already is a temporal indication in the sentence (e.g. *domani vado al mare* [tomorrow I go to the seaside]), when reference is made to a planned situation or when an idea of intentionality is expressed (ibid.). The use of the simple present in Italian in this case would therefore be appropriate. Given that a verbatim translation is not possible in this case, the author wanted at least to keep the verb *to see*, because it conveys the core message in the utterance and because, as we have seen, it is possible to use such verb in this particular meaning in Italian, provided that some structural change is made to the sentence. The translation *L'oracolo è pronto a vederti* allows the verb *to see* to be kept, *l'oracolo* to remain in subject position and a similar utterance length to be maintained. The adverb *now* was eliminated, but such elision was compensated by *è pronto*, which already contains the idea of immediateness.

With compensation, "translators can choose amongst several solutions, which may lead to a different distribution of the units of translation without impinging on the global effect of the message" (Vinay and Darbelnet, 1995: 204). Thus, with this translation strategy the degree of formal literalness is by necessity somewhat reduced. For this reason, subtitles where compensation was used in the literal transfer condition were not selected as test items for the MCQ.

3.11.4 Culture-bound Terms

Realia and culture-specific expressions also occurred in the subtitles. The concept of dynamic equivalence (Nida, 1964) was used to deal with such instances. Nida defines a dynamic equivalent translation as "the closest natural equivalent to the source-language message" (1964: 166), where the key terms *closest* refers to the higher degree of approximation between ST and TT. Nida identifies two main areas where the translator will have to make adaptations, grammar and lexicon. At the lexical level, the translator is faced with three classes of words:

(1) terms for which equivalents exist, (e.g. hand, tree, flower)

(2) terms which identify culturally different objects but with similar functions (e.g. book and papyrus)

(3) terms which identify cultural specialities (e.g. igloo, synagogue)

Two instances of (2) occur in subtitle 10 and 129.

Subtitle 10	
Really good noodles.	L: Ottime tagliatelle. [Excellent tagliatelle.]

Subtitle 129	
Take a cookie.	L: Prendi un biscotto. [Take a biscuit.]

At the time of writing, *noodles* and *cookie* are two terms slowly seeping into Italian culture, but not established enough to be dictionary entries yet, unlike other words such as *computer*, *stop*, *coach*, and the like. The term *cookie* appears in the online Sabatini-Coletti Italian Language Dictionary, the Hoepli Italian Language Dictionary as well as in the Treccani Encyclopaedia, but only the internet-related meaning is given. Both terms refer to culture-specific concrete objects, for which there are other concrete objects fulfilling a similar function in the Italian culture: just consider the virtually endless varieties of string-like pasta (*tagliatelle, spaghetti, fettuccine, tagliolini*, and so on) or the several regional types of dry biscuits (*gallette, baci di dama, cantucci*, just to name a few). If one considers using a borrowing inappropriate as it would introduce an unnecessary foreignisation, an alternative strategy needs to be adopted. The source was culturally transposed here in both occasions: *noodles* was rendered as *tagliatelle* (a type of Italian pasta), hence with a domestication, while *cookie* as *biscotto* (biscuit), hence with a domestication and a generalisation at the same time, biscuit having a wider, less specific literal meaning than cookie (hyperonymy). This process of reformulation does entail an unavoidable level of translation loss, and does not allow a clear distinction between literal and non-literal rendering, as neither formal nor semantic equivalence are achieved, and functional correspondence between ST and TT is only partial. Even if the subtitler takes maximum care to produce a translation formally, semantically and functionally as close as possible to the source, in some cases the lexico-grammatical structure or the concepts (or both) of SL and TL simply do not concede this: some subtitles will inevitably render the literal-non-literal translation distinction less successfully than others. Due to this unconformity of translational outcomes, only the most suitable items were included in the post-test, while others (including culture-bound terms and idioms) were carefully excluded.

3.12 Post-test and Distracter Choice

The recall post-test consisted of a MCQ (Multiple Choice Questionnaire) with 22 items (See Appendix A.3), administered right after viewing. Three options were presented to the subjects: literal, non-literal and distracter. Distracters were introduced to limit guessing, enabling to make more stringent conclusions about the MCQ answers. By having only two options, subjects have a 50% chance of choosing randomly and still score correctly on an item. By introducing a distracter, the odds of guessing are reduced and confidence in the obtained results is increased. Test items were chosen based on the relevance and on the clarity of the literal vs. non-literal translation distinction. Table 3 summarises the options:

MCQ choices	**Description**
Literal (L)	Formal equivalence + semantico-functional equivalence
Non-literal (N)	Formal discrepancy + semantico-functional equivalence
Distracter (D)	Formal discrepancy + semantico-functional equivalence where possible, otherwise slight difference in meaning.

Table 3. Post-test options.

Most distracters are therefore non-literal renderings, especially when syntactic structures were translated. Distracter choice depended on a number of factors, *in primis* whether there was a third way of saying the same thing in a different form – and this was often not the case. In some occasions finding a third way of expressing the same idea without distorting the message was challenging, while on others there were several possible non-literal renderings. Distracters were chosen on a case-by-case basis, depending on source-target culture considerations, context, level of overlapping of semantic spheres, availability of synonyms or near-synonyms in Italian, sentence structure, discourse, register, word frequency, as well as other non-strictly grammatical considerations such as confounds derived from question lay out. Sometimes the distracter was just another possible non-literal rendering, and, as such, D could have been selected as N. Consider subtitle 27 again:

L: *Lei direbbe che ne sa abbastanza*.

N: *Lei direbbe che ne sa a sufficienza*.

D: *Lei direbbe che ne sa quanto basta*.

Both *a sufficienza* and *quanto basta* mean enough, but the former has a different root (*suffice*) which makes it less formally related to *abbastanza*. If one imagines these three expressions in spacial terms, where *abbastanza* is the central point, *quanto basta* and *a sufficienza* in this case can be seen as being at the same distance from the centre. Hence D and N could have been swapped here. However, *abbastanza* and *quanto basta* are fairly close as both contain the root form *basta*, so there is less of a discrepancy between the two. Because the main interest of this study lies in whether a stronger memory trace is left by redundancy (equivalence of meaning *and* form) or by discrepancy (equivalence of meaning, *not* form), *a sufficienza* was considered a more suitable choice as N because it allows for further discrepancy. Hence the choice of *quanto basta* as distracter. In other cases, N and D are not equidistant from the centre (the literal rendering). Take for example subtitle 44:

Subtitle 44	
That's impossible.	L: È impossibile. [(It) is impossibile.]
	N: Non è possibile. [(It) is not possibile.]
	D: È impensabile. [(It) is unthinkable.]

N is created through a modulation (the point of view changes from affirming impossibility to denying possibility). There is formal discrepancy between L and N but the semantic meaning is maintained. In D, the adjective *impensabile* comes from *pensare* (to think), and refers to something that cannot be conceived by the mind or explained through reason. N and L are in an exclusive relation with each other, such that this impossible/not possible dualism makes any other non-literal rendering sit on a different level. The distracter here was selected because it allowed to keep a degree of consistency between the three options (they all have the copula *è*, and *impensabile* has some formal similarity with *impossibile*). This was done to reduce possible bias in the learners towards a specific answer stemming from question layout.

In her noticing study on standard subtitling, Ghia (2012a: 89) mentions how the time gap between viewing and post-tests might have somewhat marred the results, as noticing could have occurred but might have not left enough of a trace in memory to allow for recall. To control for this, post-tests were created in Tobii Studio itself and administered immediately after watching, appearing as the next stimulus on the timeline, which shortened the time gap by avoiding having to switch to an online or paper questionnaire format. Moreover, the order of the items in the MCQ was arranged so that questions about Part 2 (subtitles 122-67) appeared first, while questions about Part 1 (subtitles 11-66) last. The time between appearance on screen of the first question and appearance of the subtitle said question refers to was therefore very brief, just over one minute. Questions on the last test items appeared first to see if, across participants, presenting an item very soon after watching made a difference in terms of response accuracy, i.e. if by decreasing the time between subtitle presentation and question presentation the number of correct answers increased.

3.13 Linguistic Category

Test items covered a range of linguistic constructions, as per table 4:

Linguistic Level	No. of items
Syntax	10
Noun	4
Adjective	2
Exclamation	1
Verb choice	3
Adverb	2

Table 4. Test items balance.

In accordance with Ghia (2012a and 2012b), both lexical items and syntactic constructions will be considered in the analysis. This distinction has been made not only in the SLA and psycholinguistic literature but also in AVT and Translation Studies where, as we have seen, Nida (1964) addresses grammar and lexicon separately as they pose different specific challenges to the translator. The present study addresses the question of whether these linguistic categories can affect the viewer as far as noticing and memorisation are concerned.

Syntax can be defined as the study of "the rules and patterns that can be used in a dynamic way to create and understand new English sentences" (Baker, 2001: 265). It involves "taking the finite building blocks made available by a given language and putting them together into an infinite number of representations, which in turn can express an infinite number of thoughts" (ibid.: 266). For the purposes of the present study, 'syntactic' is operationalised as 'referring to linguistic structures' and is contrasted with individual lexical items. Test items on syntactic structures therefore involved several parts of speech and their different arrangement in the subtitles, whereas test items on specific nouns, adjectives, adverbs, exclamations and past participles involved subtitles where only one part of speech in the original dialogue was changed between L and N. For instance, subtitle 114 (AOI.99) was considered syntactic, since the whole source verb phrase 'make a choice' was targeted rather than the individual word 'choice'. Subtitle 111 (AOI.96), on the other hand, was classified as lexical, since only the ST word 'blindly' was targeted. For the purpose of statistical analysis, the linguistic levels in table 4 will be clustered together so that the totality of test items will be divided into two groups, lexical and syntactic, the former with 12 items, the latter with 10. The complete list of post-test items is reported in table 5 below. Translating syntactic items involved a change in word order, government or rephrasing. Wherever possible, translating lexical items (where only one element in the ST was targeted) involved maintaining the same number of words in both N and L, e.g. in subtitle 38 (AOI.26), where only the past participle of the verb 'find' was translated differently in the two versions (*trovato* e *individuato* respectively). When maintaining the exact same number of words in both L and N lexical versions was not possible, changes were kept to the minimum and remained lexical in nature, like in subtitle 31, where the ST word addressed ('enough') was translated with *abbastanza* (literal) and *a sufficienza* (non-literal). Although the adverb phrase *a sufficienza* is made of two constituents, word order and syntactic make-up of the sentence remain exactly the same. The same applies to subtitle 109, where the source word 'believes' is the only element that changes in the two versions, being translated with *crede* (literal) and *ha fiducia* (non-literal). *Ha fiducia* has one more word than *crede* but fits in exactly the same sentence position and has the same government (both verbs require the preposition *in* afterwards). In table 5 below, AOI (Area of Interest) number is a subtitle identifier created in the eye-tracking software, and is presented alongside subtitle number for ease of cross-reference with chapters 4 and 5, where AOI numbers will be used. While the subtitle number counts all subtitles appearing in the clip (n = 133), the AOI number counts only the subtitles that were considered in the statistical analysis (n = 110), i.e. excludes the beginning and ending adaptation segments (see 3.5 and 3.17.1). The Q numbers refer to the order of the questions in the post-test. The bold highlights indicate what parts of the ST were modified during literal and non-literal transfer.

Subtitle no.	ST	TT - Literal	TT - Non-literal	Linguistic Category	Part of Speech Targeted in ST
120 (AOI.105)	Q1 - You have a good **soul**	Hai un buon animo [You have a good soul]	Hai un buon cuore [You have a good heart]	Lexicon	noun
114 (AOI.99)	Q2 - You're going to have to **make a choice**	Dovrai fare una scelta [You will have to make a choice]	Dovrai scegliere [You will have to choose]	Syntax	verb phrase
111 (AOI.96)	Q3 - He believes it so **blindly**	Ci crede così ciecamente [He believes it so blindly]	Ci crede così intensamente [He believes it so intensely]	Lexicon	adverb
109 (AOI.94)	Q4 - Morpheus **believes** in you, Neo	Morpheus crede in te, Neo [Morpheus believes in you, Neo]	Morpheus ha fiducia in te, Neo [Morpheus has faith in you, Neo]	Lexicon	verb choice (present tense)
106 (AOI.91)	Q5 - **Without him, we're lost**	Senza di lui, siamo perduti [Without him, we're lost]	Se non ci fosse, saremmo perduti [If he wasn't there, we would be lost]	Syntax	sentence structure
103 (AOI.89)	Q6 - He almost had me **convinced**	Mi aveva quasi convinto [He almost convinced me]	Mi aveva quasi persuaso [He almost persuaded me]	Lexicon	verb choice (past participle)
98 (AOI.85)	Q7 - **Your next** life, **maybe**	La tua prossima vita, forse [Your next life, maybe]	Una vita futura, magari [A future life, perhaps]	Syntax	possessive + adjective + adverb
82 (AOI.70)	Q8 - Being the One is just like **being in love**	Essere l'Eletto è come essere innamorato [Being the One is like being in love]	Essere l'Eletto è come amare qualcuno [Being the One is like loving someone]	Syntax	verb phrase
77 (AOI.65)	Q9 - **Honestly, I don't know**	Onestamente non lo so [Honestly, I don't know]	Non ne ho idea, davvero [I have no idea, really]	Syntax	sentence structure
72 (AOI.60)	Q10 - **Not too bright, though**	Ma non sei tanto sveglio [But you are not very bright]	Però ti manca l'intuito [But you lack intuition]	Syntax	sentence structure
68 (AOI.56)	Q11 - If I hadn't **said anything**	Se io non avessi detto niente [If I hadn't said anything]	Se io non avessi aperto bocca [If I hadn't opened my mouth]	Syntax	verb phrase

16 (AOI.5)	Q12 - That's **different**	È differente [It's different]	È diverso [It's different]	Lexicon	adjective
25 (AOI.13)	Q 13 - The **prophecy**	La profezia [The prophecy]	Il vaticinio [The vaticination]	Lexicon	noun
29 (AOI.17)	Q 14 - Of the **resistance**	Della resistenza [Of the resistance]	Della guerriglia [Of the guerrilla]	Lexicon	noun
31 (AOI.19)	Q 15 - She would say she knows **enough**	Lei direbbe che ne sa abbastanza [She would say she knows enough]	Lei direbbe che ne sa a sufficienza [She would say she knows a sufficient amount]	Lexicon	adverb
38 (AOI.26)	Q 16 - That I would **find** the one	Che avrei trovato l'Eletto [That I would find the one]	Che avrei individuato l'Eletto [That I would identify the one]	Lexicon	verb choice (past participle)
44 (AOI.32)	Q 17 - It's **impossible**	È impossibile [It's impossible]	Non è possibile [It's not possible]	Lexicon	adjective
51 (AOI.39)	Q 18 - **I know you're** Neo	So che sei Neo [I know you're Neo]	Tu devi essere Neo [You must be Neo]	Syntax	sentence structure
54 (AOI.42)	Q 19 - **Bingo!**	Bingo! [Bingo!]	Indovinato! [Well guessed!]	Lexicon	exclamation
57 (AOI.45)	Q 20 - **Smell good**, **don't they?**	Hanno un buon profumo, no? [They have a good smell, don't they?]	Senti che buon profumo! [What a good smell!]	Syntax	sentence structure
63 (AOI.51)	Q 21 - **I'm sorry**	Mi dispiace [I'm sorry]	Sono desolato [I'm terribly sorry]	Syntax	sentence structure
65 (AOI.53)	Q 22 - I'll get one of my **kids** to fix it	Lo farò aggiustare da uno dei miei ragazzi [I'll have it fixed by one of my kids]	Lo farò aggiustare da uno dei miei studenti [I'll have it fixed by one of my students]	Lexicon	noun

Table 5. Post-test items summary table with linguistic category details.

3.14 Relative Frequency

All frequencies for this study were extracted from the itWac corpus, a general-purpose corpus containing just under two billion words (1,909,826,324 tokens), one of the largest resources for Italian to date (Baroni et al., 2009: 209). The corpus was built by web-crawling, which has the benefits of more solid statistics and a larger number of usage examples compared to other types of corpora (ibid.: 219). The data is crawled from the *.it* domain (i.e. excluding sources such as Facebook or Wikipedia) and provides a comprehensive snapshot of the Italian web, including sources such as *La Repubblica* and *Il Corriere della Sera*, two major Italian newspapers. ItWac has a wide range of text types and topics. Blogs, bulletin boards, as well as academic and journalistic texts are featured; politics, sports, culture, lifestyle, chronicles, opinion, current events, technology and travelling are amongst the topics covered. This makes it a varied corpus in terms of content and genres. Since it contains information from blogs and personal pages as well, itWac is not restricted to the formal register typical of written communication but includes examples of more informal and colloquial language, such as the one used in film dialogues. Moreover, it is annotated with POS (part-of-speech) tagging and has a format specifically tailored to linguists and language researchers (ibid.: 224).

Queries on itWac were made via the IntelliText interface (University of Leeds), which can produce both raw and relative frequencies. Raw frequencies are the arithmetic counts of the instances of a linguistic feature (a word or structure) in a corpus. These depend on the size of the corpus and can therefore be difficult to interpret and potentially misleading. Relative frequencies, on the other hand, are corpus frequency values normalised to a common base, e.g. per million tokens. Relative frequencies do not just tell us how many times a noun, verb or verb construction occur in a corpus, but relate these numbers to the total number of tokens in that corpus, allowing meaningful comparisons of frequencies across corpora. The unit of measurement is usually IPM (Instances Per Million). In the present study, relative frequency was calculated for the 22 test items. Two relative frequencies were extracted, one for the literal (IPM_L) and one for the non-literal (IPM_N) wording of each subtitle, for a total of 44 distinct frequencies.

In most cases, corpus queries were straightforward. This is true in particular for lexical items, where only one part of speech (a content word in the ST) was manipulated in the translation. With nouns, adjectives, adverbs and exclamations, the exact wording was queried (e.g. AOI.13 *profezia – vaticinio*)[21]. As far as verbs are concerned, the lexical test items include two past participles and a present tense. For the present tense (AOI. 94 *crede – ha fiducia*), the query was as follows:

AOI.94	L: [lemma="credere"]
	N: [lemma="avere"] [word="fiducia"]

By searching the lemma of the verb, the syntax of this query ensures that other representative instances are returned as well as the 3rd person singular of the present tense. That is to say, in L, sentences containing *credo* [I believe], *credevano* [they believed, impf. tense], *credimi* [believe me] and similar are included as well as *crede* [s/he believes]. In N, sentences containing *avere fiducia* [to have faith], *abbiano fiducia* [that they have faith, subj. tense], *abbi fiducia* [have faith, imper. tense] are included as well as *ha fiducia* [s/he has faith]. For the other two cases (AOI.26 *trovato – individuato* and AOI.89 *convinto – persuaso*), since the past participle behaves differently from other examples that would be returned if the lemma was queried (see, for example, the use of the past participle as an adjective), the past participles themselves were queried rather than the infinitives.

As we have seen in 3.13, syntactic items addressed several parts of speech and their different arrangement in the subtitles. To ensure maximal representativeness of the frequency produced, concordance lines were also checked for each query. Take, for example, AOI.70 (*essere innamorato – amare qualcuno*). The final query was as follows:

AOI.70	L: [lemma="essere"] [word="innamorato"]
	N: [lemma="amare"] [word="qualcuno"]

Initially, for N, only the lemma *amare* was searched. However, concordance checks revealed that this query also produced non-relevant occurrences such as *amare + infinitive*, e.g. *io amo raccogliere i fiori* [I love picking flowers], in the sense that an action (in this case

[21] Due to space constraints, shortened references like this will be made to each AOI (subtitle) in this section. The first item in the pair is always L, the second N. For the full list of subtitles used as post-test items, classified by translation condition and complete with back-translations into English, please refer to table 5 earlier in this chapter.

picking flowers) is enjoyed by its agent. This is not the meaning of the verb used in the context of the video (amorous love) and therefore counting these instances artificially inflates frequency. By including the word *qualcuno* [someone] in the query, on the other hand, concordance checks revealed that only the relevant examples were included in the frequency count.

With three syntactic items (AOI.56, AOI.65 and AOI.85), two queries were possible within each subtitle. Let us take AOI.85 (*La tua prossima vita, forse – Una vita futura, magari*) as an example.

AOI.85	Query 1	L: [word="prossima"] [word="vita"]
		N: [word="vita"] [word="futura"]
	Query 2	L: [word="forse"]
		N: [word="magari"]

In this test item, both the adverb (*forse, magari*) and the noun phrase (*prossima vita, vita futura*) were targeted in the translation. Therefore, two separate queries were made, resulting in two IPM values per translation condition. In these instances, the two frequencies for each translation condition were averaged, in order to obtain one single value for IPM_L and one for IPM_N.

In one case, i.e. the syntactic item AOI.60 (*Ma non sei tanto sveglio – Però ti manca l'intuito*), it was not possible to produce a meaningful frequency. In Italian, the adjective *sveglio* can refer to someone who is (a) awake, up, wide-awake; (b) alert, quick, quick-witted, smart; (c) shrewd, cunning, crafty (Garzanti Linguistica, 2016). In the context of the video, the word is used as (b). By querying the verb phrase *essere sveglio*, however, all three meanings were added to the frequency count, and in the vast majority of examples the word was used as (a). Unfortunately, the difference between (a) and (b) is purely semantic (it cannot be identified based on formal properties of the queried string) and corpora do not yet allow to differentiate between meanings in a query purely based on semantics. Including all results would skew the perception of the real frequency of (b). Moreover, only the frequencies of the conjunctions *ma – però* could be obtained for both translation conditions, yet using only the frequency of the adverbs ignores the other content words in the sentences (*sveglio, mancare, intuito*) and addresses too small a portion of the subtitles to be representative. For these reasons, AOI.60 was excluded from the frequency analysis (see 4.3.7). Relative frequency for the 44 items in this dataset ranged between 0.01-333 IPM (M = 38.69, SD = 64.74).

3.15 Open-ended Questionnaire

After the post-test, a questionnaire was completed by the viewers (see Appendix A.4). This explicit report method was chosen because it provides a direct indication of viewers' noticing of a discrepancy in itself (Ghia, 2012a: 89). Moreover, this choice was in line with empirical research on noticing and attention, which typically uses retrospective verbal accounts and post-exposure questionnaires (Curran and Keele, 1993; Robinson, 1996, 1997a, 1997b; Godfroid et al., 2010). This type of offline report has also been used in the AVT literature in the context of FLL (Talaván, 2010; Talaván and Rodríguez-Arancón, 2014; Čepon, 2011; amongst others).

The explicit report task contains five questions with open answers, which draw from Ghia's questionnaire. The latter contains only two general questions, the first requesting "overall comments on the subtitles" and the second asking if "any difficulties [were] encountered during viewing" (Ghia, 2012a: 78). Ghia maintains that keeping the questions so general was done to avoid influencing the participants' answers (ibid.). However, her questionnaire analysis reveals that relatively little interesting data emerges with such design, other than the fact that, in answering the second question, six out of 13 respondents (46%) reported noticing a discrepancy between ST and TT (ibid.: 86). In the present study, therefore, a more explicit approach was taken to elicit more nuanced responses. Asking subjects to generically report any difficulty in following *the film* (as Ghia does in her study on standard subtitles) may result in even more under-informative responses in the case of reverse subtitles, since the soundtrack of the film is in the viewer's L1, which makes comprehension effortless and may dissuade subjects from commenting on noticed discrepancies in subtitle translations because these, in principle, do not affect the ability to follow the film plot. For this reason, the 'difficulty' question was made more specific by enquiring whether participants had any problems in following *the subtitles* (Q2). Before Q2, participants were asked more generally if they had noticed anything about the subtitles (Q1), and after Q2 they were asked more specifically to enter any L2 expressions that they found striking (Q3), thus producing an ordered crescendo from the general to the particular as far as the subtitles are concerned. Then, a loosely-structured question probed the participants' impressions about their experience of the study as a whole (Q4).

Providing a space for interviewees to express any other thought they may have not only helps establishing how the experiment is received more broadly, but also recognises them more overtly as active respondents to the stimuli presented. The last item (Q5) was introduced to collect data on whether the film had been seen before (yes/no dichotomous response). The questions were presented in digital format in Internet Explorer through the Bristol Online Survey (BOS), an easy-to-use tool developed by the University of Bristol to facilitate the development, deployment, and analysis of surveys through the web (BOS, 2016). The answers were completely open, the only constraint being that participants had to provide an answer, however short, to move to the next question. Thus, participants were encouraged to elaborate on what they had just seen and write their impressions about the reverse subtitles. This type of qualitative data allows both the participants to express what is most important to them and the researcher to gauge their opinion on the experience as a whole. The analysis of their open answers was used to triangulate the findings on the construct of noticing, while providing a measure of the participants' metalinguistic awareness, i.e. their explicit report of what aspects of the L2 they found striking and why. Finally, the questionnaire was chosen because "by permitting greater freedom of expression, open-format items can provide a far greater 'richness' than fully quantitative data. The open responses can offer graphic examples, illustrative quotes, and can also lead us to identify issues not previously anticipated" (Dornyei, 2003: 36). By allowing participants to openly express their opinion, this explicit report might lead to unexpected relevant data, especially on the conscious noticing of translation discrepancies.

3.16 Procedure: The Experimental Session

All sessions took place in the Human Communications lab at the University of Leeds between May 2013 and June 2015. Participants were tested individually. They were welcomed to the lab and briefed about the experiment. They were given only general information, i.e. that the study was looking at language and film and that there would be subtitles, but they were not informed about the language in which these would be presented to avoid drawing attention to the reverse mode and prime them to pay more attention to the TT. A short description of the eye-tracking technology used was given. Participants were also instructed not to wear make-up and bring their glasses if they usually wore lenses, since make-up can sometimes cause issues in the tracking process and so can lenses, due to micro-air bubbles that can form between the lens and the surface of the eye. These can be misjudged as the glint, i.e. the pupil-corneal reflection, used by the eye-tracking hardware to locate the position of the eye (Holmqvist et al., 2011). Hair pins and bubbles were provided by the researcher in case hair was covering the eye area. All participants signed an ethical approval statement on arrival (see Appendix B.2) and were randomly assigned to G1L or G2N (see 3.5). First, participants sat the language pre-test (around 20min) in written form, at a desk without dictionaries or access to the web. Then they moved to the eye-tracker station for the central part of the experiment (around 40min), i.e. viewing and post-tests. The researcher made sure they sat comfortably on a static chair (to limit participant movement) at around 60cm distance from the screen. After a 9-point calibration of the remote eye-tracker, the dummy clip was started (see 3.5.1). The gaze paths were then inspected by the researcher to ensure minimal offset (see 3.17.1) and calibration was repeated, if necessary, until an optimal set-up producing the best gaze path accuracy was found. After the dummy clip, participants watched the subtitled clip. Before watching, they were instructed to behave normally, as if they were watching the video at home. They were told that there would be some comprehension questions at the end of the viewing session, but no further information was given: they were unaware that their verbatim memory was under investigation. Immediately after viewing, they were presented with the surprise recall post-test, which was the next stimulus in the Tobii Studio experiment timeline. The questions appeared one by one, and the participants could take as long as they wanted to answer them. An MCQ option had to be chosen to be able to move to the following question. Up until this point, no verbal interaction occurred between subjects and the researcher since the start of the experimental video. After the MCQ, the researcher intervened to open and start the online questionnaire. Again, participants could take as long as they wanted to complete the five open questions, and they were encouraged to answer as thoroughly as possible and to give their opinion on the experience. When they finished the online questionnaire, they were told they could relax, offered sweets and chocolate and given a chance to stretch their legs if they so wished. The third and last part of the experiment consisted of the WM test (around 20min), which took part in an adjacent lab room. The participants were briefed about the RST (see 3.10) and a short training test was practiced first. The WM test was always presented last, to avoid drawing attention to the memory aspect of the experiment before watching the video. The sessions lasted between 1h15min-1h45min, depending on the eye-tracker calibration time, and how long the participants took to complete the self-paced language test and the post-tests.

3.17 Data Collection and Processing

This section addresses the process of collecting, extracting and preparing data for the statistical analysis. The main procedures adopted within the eye-tracking software (Tobii Studio), spreadsheet programme (Excel) and statistical package (R) used are described hereafter.

3.17.1 AOIs

After recording eye-tracking data, a number of manual operations had to be carried out in Tobii Studio. First of all, individual Areas Of Interest (AOIs) had to be drawn manually around each subtitle. A total of 133 subtitles appeared in the video; AOIs were drawn on 110 of these, so that they could be analysed. Of the remaining 23, the first ten and the last eight adaptation subtitles contained the same text in both groups and were excluded from the analysis, in line with the previous literature (e.g. Ghia, 2012a: 78). Five subtitles (no. 13, 22, 89, 102, 105) were not translated differently in the two versions and therefore also had to be excluded from the analysis. For practical reasons, in chapter 4 and 5, subtitles will be referred to by their AOI number. Each of the 110 AOIs were drawn around the whole subtitle. In some cases, participants presented a gaze path offset. An example can be seen in fig. 2 below.

Figure 2. Example of a gaze path offset, where the fixations do not fall exactly on the words.

Although it is obvious from fig. 2 that the participant is looking at the subtitle, some fixation information would be lost if the AOIs were drawn *exactly* around the subtitle words. Therefore, the AOI borders were set to 1cm left and right of the first and last word in the longest line of the subtitle. The bottom border was always set to the bottom of the letterbox, while the top was set to 5mm just above the letterbox border. This was done to prevent the loss of relevant subtitle fixations which fall just outside (a few millimetres above or below) the actual words. This operation allowed all eye movements to the target text to be captured. A screenshot of a typical AOI can be seen in fig. 3.

Figure 3. Example of an Area Of Interest (AOI) around a subtitle.

One of the limitations of Tobii Studio (up to version 3.1.6 at the time of writing) is that it does not automatically create AOIs around any text stimuli. This makes initial eye-tracking data processing a largely manual operation where the researcher draws single AOIs around the target text, which then can be copied over to all participant recordings in that same test only. Because of our experimental design, this operation had to be done twice, once for G1L and once for G2N. One of the major problems with researching AVT through eye-tracking is that the subtitles still need to be burnt onto the image before being uploaded to the eye-tracking software (in our case, Tobii Studio). This effectively makes subtitles an integral part of the image, so they cannot be automatically identified even when the eye-tracking software allows for text detection and automatic AOI drawing (Kruger et al., 2015). The day eye-tracking technology will allow automatic subtitle text recognition, a tangible methodological improvement in AVT eye-tracking research will ensue. For the time being, the procedure described herein is the standard adopted in AVT studies of this kind.

3.17.2 Eye Metrics

The eye metrics analysed in this study are the following: fixation count (number of fixations on a given AOI), fixation duration (length of individual fixations in seconds or milliseconds within a given AOI), total fixation count (the total number of fixations on all the AOIs considered), total fixation duration (the sum of the length of all fixations in in seconds or milliseconds for all the AOIs considered), average fixation count (the sum of all fixation counts on an AOI made by all participants, divided by the number of participants) and average fixation duration (the sum of length of all fixations made by a subject on an AOI divided by the number of fixations they made on that AOI).

In the by-subject analysis, eye movement information was extracted through the Tobii Studio Statistics tab (see 3.17.4 below), a tool that pre-processes the eye-tracking information to produce specific data tables. In these Tobii tables, fixation duration data is aggregated by participant and presented in seconds. In the by-item analysis, on the other hand, the eye-tracking data was not pre-processed through the Tobii Studio statistics tool. To enable more advanced analysis in a dedicated statistical tool, the full dataset was exported directly into Excel. In this much more detailed dataset, eye-tracking data is not aggregated and is therefore presented in milliseconds.

Initially, I considered analysing regressive movements as well. However, another major limitation of the Tobii system is that it does not include a regression metric. If interested in analysing such eye movements, a researcher has to count them manually for every subtitle and participant in all conditions. In this study, a total of 2860 items (110 subtitles x 26 participants) would have to be processed manually, a considerably lengthy process which has to be weighed against the relevance of the eye metric for the experiment. Moreover, in the parallel study carried out by Ghia (2012a), regressions were analysed and did not reveal any significant difference between the two translation conditions. Finally, regressions were not crucial to the RQs (3.4). Therefore, it was decided not to include them in the analysis. On this point, also see the discussion in 5.4.

3.17.3 Sampling Rate and Sampling Frequency

Sampling rate (of a recording) is a rather coarse measure of the confidence with which an eye-tracker identifies eye position and movement throughout an experiment trial. This is summarised in Tobii Studio's Replay tab, where a sample percentage is assigned to each participant recording. This sample rating roughly estimates the tracking quality and is based on the validity codes provided for each gaze data point collected for each eye, which are then summed and normalized resulting in percentage values between 0 and 100. The Tobii manufacturers recommend checking the validity codes rather than relying solely on sampling rate, which "is only meant to give a general feeling for the number of valid gaze points in a recording and cannot be used to gauge the accuracy of the data" (Tobii Pro Global, 2015). Since no information was found on this topic in the relevant published literature, and given the manufacturers recommendations, a value of 50% was chosen as minimum threshold: all recordings with sampling rates lower than these were discarded. This meant that 5 subjects had to be excluded from the analysis. The sampling rate in the main study varied between 53-97%, with most recordings being above 80% (M = 83%, SD = 13.85%). For the few (n = 3) recordings with a sampling rate of 50-70%, the data file was manually inspected and validity codes were checked to ensure that a satisfactory number of high-confidence samples were present.

Sampling frequency (of the eye-tracker), on the other hand, is an extremely important property of an eye-tracker, because it affects the type of research that can be carried out with it. For example, small saccades or tremors are so sensitive that studies looking at these movements typically use high sampling frequencies to estimate them accurately (Andersson et al., 2010: 1). Sampling frequency is measured in Hertz (Hz), which refers to the number of data samples per second. In this experiment, data was recorded with a sampling frequency of 60Hz, meaning eye position samples are taken every 16.67ms. As Andersson et al. put it, it is crucial "to know what speeds we need for our particular research questions and also to know when it yields a net improvement to sacrifice speed in order to capture the behaviour in a more naturalistic setting, e.g. using the less intruding remote filming for slower eye movements" (2010: 2). What frequency it is best to adopt is therefore ultimately determined by the research questions, as well as considerations of ecological validity versus variable control. This study was specifically designed to maximally approximate 'normal', more relaxed viewing conditions by using authentic material in a naturalistic setting. Most importantly, our research questions address the realm of fixations, typically slower eye-movements compared to saccades or microsaccades. Low-level visual cognition studies, e.g. those involving gaze-contingent paradigms, typically work with really high frequencies (1000-2000Hz), whereas research addressing higher-level cognition in naturalistic settings prefer remote tracking systems allowing greater head movement, whose speed typically range between 25-250Hz (ibid.).

Moreover, 50Hz is a standard in much of the published eye-tracking research involving subtitles (Ghia, 2012a; Perego et al., 2010; Moran, 2012, just to name a few), so at the time of writing 60Hz is fully acceptable in terms of field standards. Finally, recording data at 60Hz corresponds to data files of a relatively small size (half the size of data recorded at the maximum available 120Hz frequency), making it a great deal easier to handle them in Excel, which was used for data export and part of data tidying. This may at first seem like an unworthy reason, but the sheer size of the data collected through eye-tracking makes it relevant to the discussion, especially if processing limitations in the available laboratory desktop computers make it impossible to manipulate extremely large files. The next section will clarify and expand on this last point.

3.17.4 Data Export

Data in Tobii Studio can be accessed through either the Statistics or the Data Export tabs. In the Statistics tab, the software processes the recorded data internally, already aggregating fixation and saccade information as to produce descriptives and tables with the available eye-metrics. The tables can be copied out of Tobii and used as they are or opened in Excel for further manipulation. As mentioned, this procedure was used to obtain the by-subject dataset. In the Data Export tab, the actual raw data are exported with the chosen fixation filter applied but *not* already pre-processed into aggregated eye-metrics. This procedure was used to obtain the by-item dataset. In this export, data relative to all individual fixations (and saccades) made by each participant is recorded in what is called the long format, meaning that each row of data is a measurement occasion, i.e. an eye sample. These data files are extremely large. Tobii Studio exports data in either .tsv or .xlsx formats, individually (one file per participant) or as a group (e.g. all participants' data in one file). In the Data Export tab, one can choose exactly what to export: what AOIs (subtitles), number of participants, and data types to include. Upon completion of the main study, several attempts at data export were made. Initially, data were exported including all data types, for all subtitles and all participants together. This operation took over a day and was not successful: the researcher discovered that such a file cannot be exported as it is too large for Excel itself to handle (over one million rows). It became clear that keeping all the non-aggregated data in one file would not be possible at export stage. It was therefore decided to export in batches, for example data for G1L and G2N separately, but these files were still too large to handle. Due to the large size of the file, even an export as 'small' as only four participants took a considerable amount of time to open in Excel, and often caused the programme to crash. Tobii allows several data types to be exported: general, media, segment and scene, timestamp, recording event, gaze event and AOI activity information, gaze tracking, eye-tracking and validity codes. Not all of these are used in our study, for example data relative to events, mouse clicks or external cameras, so these can easily be excluded from export. Under each of the above data type groups, only the relevant items were selected, as to eliminate unnecessary columns and lighten the resulting file. Exporting files individually was found to be the best option in terms of length of the process and further file maneuverability in Excel. Exporting one file per participant also allowed the researcher to try different aggregation and other manipulation techniques on smaller, more manageable datasets, before merging them into the final by-item dataset for analysis.

3.17.5 Data Tidying[22]

Data tidying is necessary to produce a suitable dataset for analysis in R, the package used in the statistical analysis (R Development Core Team, 2015). This process is not straightforward; in fact, it constitutes one the most complex, unintuitive and frustrating parts of data analysis (Wickham, 2005 and 2007), not least because data often has multiple levels of grouping and is investigated on multiple levels (ibid.). In R, this manipulation involves many functions: `tapply`, `by`, `aggregate`, `xtabs`, `apply`, `summarise`, `subset`, as well as `cast` and `melt` in Wickham's reshape2 R package (2007). Understanding which final data format is more desirable for the type of analysis envisaged is part of the challenge. This is crucial especially given that, although the wide format is typically used when entering data for many experiments (Field, 2012: 96), some statistical analyses such as mixed or multilevel models require data to be in the long format (Grace-Martin, 2015), so knowledge of how to reshape data to achieve both is extremely useful. Carrying out these reshaping steps in R and Excel on the relatively small data sub-sets described in 3.17.4 proved invaluable to achieve a workable dataset.

In line with the standard in psycholinguistics (Locker et al., 2007) in this study by-subject and by-item analysis were carried out separately, and corresponded to two main separate data files. As mentioned, Tobii tables were used to create a usable dataset for the by-subject analysis. The tables were exported to Excel where data were tidied and other

[22] By data tidying, I refer to any operation that helps restructuring the dataset to then facilitate analysis (Wickham, 2007). In this sense, data manipulation, post-processing, tidying, restructuring and reshaping have all been used interchangeably in the literature.

subject-specific variables such as age, gender, recall accuracy, language proficiency and WM scores were added before importing in R. In the by-subject analysis, only the 22 test items were considered (i.e. those items for which there is also recall accuracy data available). Two Tobii tables were exported, one for G1L and one for G2N. The data file exported had participants as rows and individual subtitles (AOIs) as columns. These 22 items were further pooled together under each participant to create totals (sum function in Excel) for both fixation count and duration of the subtitles, as well as the breakdown per translation condition (L or N). Then these data were combined in one Excel file with all participants from both groups (n = 26) in the rows and totals for fixation metrics in the columns (i.e. total fixation count on all items, L items, and N items, and same for fixation duration). Descriptive statistics, t-tests and correlations were produced on this dataset in R (see 4.2).

In the by-item analysis, on the other hand, both data on the test items (22 subtitles) and the whole experimental stimulus (110 subtitles) was considered. Data tidying operations were carried out in Excel and R. R is a very powerful command-line environment for statistical computing and graphics based on a language from the S family (Torf and Brauer, 2014). R typically has a steep learning curve (Verzani, 2002: 1) and can be initially very time consuming, but allows much more freedom and flexibility than Excel and other dedicated statistical packages such as SPSS. Wherever possible, data tidying was therefore carried out directly in R. For example, the exported individual files contained irrelevant fixation data at the beginning and at the end of each set of participant observations: this is because the eye-tracker starts collecting eye data for each person immediately as a recording is initiated, yet it takes slightly longer for the video stimulus to start playing. This is visible in the replay tab function in Tobii Studio, where eye movements on a black screen are recorded just before the video starts playing. The same but in reverse happens at the end: the media stimulus stops but the eye-tracker keeps collecting data for slightly longer after the end of the video. These portions of the file contain superfluous data that should therefore be eliminated. In this dataset, elimination could easily be achieved in R by the following subsetting command:

```
mynewdata <- subset(mydata, MediaName == "matrix_LNL-PFR(2)_mpeg4.avi")
```

This and other tidying operations were carried to produce the final dataset on which the statistical analysis is carried out. For example, irrelevant information such as saccadic movements and all fixation data outside the AOIs was eliminated. The exported dataset had to be transposed, i.e. all AOI information, initially contained in (over a hundred) columns had to be transferred into rows. If-statements in Excel and the functions `cast` and `melt` in Wickham's reshape2 package (2007) proved very useful to create the backbone dataset, i.e. a semi-final dataset containing all relevant eye movement data for all participants and subtitles. To carry out the complete by-item analysis for this study and answer all RQs in 3.4, a number of other variables was then added to the dataset, namely subtitle duration and length (number of words), linguistic category, WM and language proficiency scores, order of presentation, frequency and recall accuracy. Adding these variables can also be done in R through the `merge` function. Once the by-item dataset was finalized, graphics and descriptives, t-tests, correlations and a logistic regression were carried out in R as part of the data analysis, which is the subject of the next chapter.

CHAPTER IV
DATA ANALYSIS

4.1 Introduction

After collection, the data were analysed with a variety of methods in R (R Development Core Team, 2015). R is a free, open-source data analysis package but also a computer programming language based on the S-family, with a powerful syntax, excellent graphics and many more in-built statistical functions compared to proprietary counterparts such as SPSS. It also allows the user to define new functions, making it a really flexible and efficient tool for both statistical analysis and data handling. R has no commercial support but a good built-in help system and an outstanding support network via online forums and mailing lists, which many would consider just as good if not better than commercial support (Verzani, 2002: 1). One major disadvantage for users not already accustomed to programming is the already mentioned steep learning curve involved with effectively using a command line to communicate with the software, as opposed to a typical point-and-click user interface. In my personal experience, however, if one is willing to accept the challenge, the choice quickly pays back. R is a collaborative project with many contributors, so help with inevitable troubleshooting can be sought online, where hundreds of statisticians and researchers from the most disparate fields, who are using and continuously improving R, can provide effective and timely assistance. And all of this is for free, since R "comes with absolutely no warranty" (R version 3.2.0), which makes it a truly unique tool, and one that is worth supporting.

Both descriptive and inferential statistics are used in R to explore the relationships between the variables of interest. By-subject and by-item analyses are considered in turn, as it is customary in psycholinguistic experiments (Locker et al., 2007). In the by-subject analysis, data from 26 participants on the 22 test items were considered. In the by-item analysis, data both on the 22 test items and on the whole experimental stimulus (110 subtitles) were considered. Throughout the chapter, L and N will stand for the literal and non-literal conditions respectively. The test items, i.e. the subtitles that are part of the recognition memory MCQ (the recall accuracy post-test), will be referenced by their AOI name. For additional information on the test items, the reader is invited to refer to table 5 in the methodology chapter.

By-subject control variables such as Working Memory (4.2.2), Italian language proficiency (4.2.3) and experimental group (4.2.4) will be addressed in the by-subject analysis. By-item variables like subtitle duration (4.3.3), subtitle length (4.3.4), linguistic category (4.3.5) and frequency (4.3.6) will be addressed in the by-item analysis. The relationship between recall accuracy and translation condition will be addressed in both by-subject and by-item analyses (4.2.1 and 4.3.1) and so will the fixation measures (4.2.4 and 4.3.2). A conclusive section (4.3.8) considers all variables together through generalised mixed-effect modelling (GLMEM), in particular to establish if they affect variance (individually or combined in interactions) in post-test accuracy scores. All research variables (see table 1) for which a *quantitative* analysis can be carried out are addressed in this chapter. This corresponds to RQ1-4 (see table 2). The last research variable in table 1 (overt noticing) and its corresponding RQ5 is explored through a *qualitative* measure (open-ended questionnaire) on which no inferential statistics was applied, and is therefore addressed directly in the discussion chapter (5.3.5).

The first section of the by-subject analysis (4.2.1), which addresses the core question of this study (RQ1 in table 2), will explain in detail how the analysis was carried out and provide all the descriptives, tables and graphs that were produced. Describing data both numerically and graphically is a crucial pre-requisite of any analysis, and the assumption that the data sample collected in a study is normally distributed holds for every parametric test (Larson-Hall, 2010: 62). Yet, such descriptive analyses are seldom reported in published journal research (ibid.: 63), despite the fact that they can present issues that need to be tackled before being able to move on to inferential analyses (e.g. lack of normality or homogeneity of variance). I purposefully chose to include detailed numerical and visual summaries of the data in section 4.2.1 even when such data satisfied the necessary assumptions, in order to give the reader a clear idea, at least once, of what guiding principles and steps were followed not only for *inferential* statistics, but also for the necessary *descriptive* statistics that have to be carried out first. The same guiding principles are then applied throughout the chapter, where data in all analyses were tested, plotted and explored comprehensively. Yet, for obvious issues of space, not *all* graphical and numerical descriptions can be reported in full for all analyses in this document. After 4.2.1, therefore, graphical representation and numerical description will be reported only when relevant to the discussion, e.g. if the data departs from normality and deserves further consideration. When descriptives are not present, it is because the data satisfied all the relevant assumptions.

4.2 By-subject Analysis

4.2.1 Recall Accuracy and Translation Condition (by-subject)

Irrespectively of presentation order (variable 'Group' in table 6 below), the majority of participants (65%) registered a higher performance with literally translated items (for the recall test, see Appendix A.3). Consider the following data:

Subj	**Group**	**Total recall (x/22)**	**Total recall %**	**Recall_L (x/11)**	**Recall_L %**	**Recall_N (x/11)**	**Recall_N %**	**Difference in %**
1	G1L	15	68	7	64%	8	73%	-9%
2	G1L	18	82	9	82%	9	82%	0%
3	G2N	18	82	10	91%	8	73%	18%
4	G1L	21	95	11	100%	10	91%	9%
5	G1L	19	86	10	91%	9	82%	9%
6	G2N	10	45	7	64%	3	27%	36%
7	G1L	15	68	7	64%	8	73%	-9%
8	G2N	17	77	11	100%	6	55%	45%
9	G1L	16	73	7	64%	9	82%	-18%
10	G1L	13	59	8	73%	5	45%	27%
11	G2N	21	95	11	100%	10	91%	9%
12	G2N	16	73	8	73%	8	73%	0%
13	G2N	17	77	9	82%	8	73%	9%
14	G2N	16	73	11	100%	5	45%	55%
15	G2N	16	73	11	100%	5	45%	55%
16	G1L	16	73	6	55%	10	91%	-36%
17	G1L	14	64	7	64%	7	64%	0%
18	G1L	13	59	9	82%	4	36%	45%
19	G2N	15	68	9	82%	6	55%	27%
20	G1L	17	77	10	91%	7	64%	27%
21	G2N	18	82	10	91%	8	73%	18%
22	G2N	17	77	10	91%	7	64%	27%
23	G1L	16	73	8	73%	8	73%	0%
24	G1L	11	50	8	73%	3	27%	45%
25	G2N	12	55	6	55%	6	55%	0%
26	G2N	17	77	10	91%	7	64%	27%

Table 6. By-subject recall performance table by translation condition (across items).

The cells in orange indicate the highest recall rate for each subject (row) across all items. Overall, 17 subjects out of 26 recalled items more accurately when presented in the literal condition (L). Only four subjects recalled non-literal items (N) more accurately. The cells highlighted in blue refer to those subjects who obtained equal recall scores in both translation conditions (n = 5). All recall scores are given as percentages as well; for example, out of the 22 questions in the post-test, S_1 correctly identified 73% of the 11 non-literal (formally divergent) items and 64% of the 11 literal (formally similar) ones, whereas S_8 identified all (100%) the 11 literal items and 55% of the non-literal ones. The last column of the table reports the difference in percentage between the scores. The negative signs indicate the four exceptional cases of superior performance with divergent items. For the most part, however, performance superiority in this small subgroup is slight: only S_{16} provided 36% more accurate answers with divergent items (10/11 correct answers in N versus 6/11 in L). S_9 provided 18% more accurate answers with divergent items and S_1 and S_7 only 9% (8/11 correct answers in N versus 7/11 in L). In most of these cases, therefore, the proportion of accurate answers in N is only slightly higher than in L. These findings strongly suggest that, after one single exposure to subtitled input, literal translations are remembered more precisely. The recall accuracy data were further explored both graphically and numerically to gather information regarding their shape and a number of assumptions that need to be met for inferential statistics to be applied.

Description	Recall accuracy (L)	Recall accuracy (N)
Mean	8.846	7.076
Median	9	7.5
Mode	10	8
Variance	2.695	4.073
Standard deviation	1.641	2.018
Sum of Squares	67.384	101.846
Standard Error	0.321	0.395
Min value	6	3
Max value	11	10

Table 7. Descriptives for subject recall data with literal and non-literal items.

All the relevant descriptive statistics are presented in the table above. Literal recall data have a SD of 1.6 and a variance of 2.7, while non-literal recall data have a SD of 2 and a variance of 4. Scatterplots, boxplots, histograms, quantile-quantile (Q-Q) plots, and barplots were produced for each condition and are reported below.

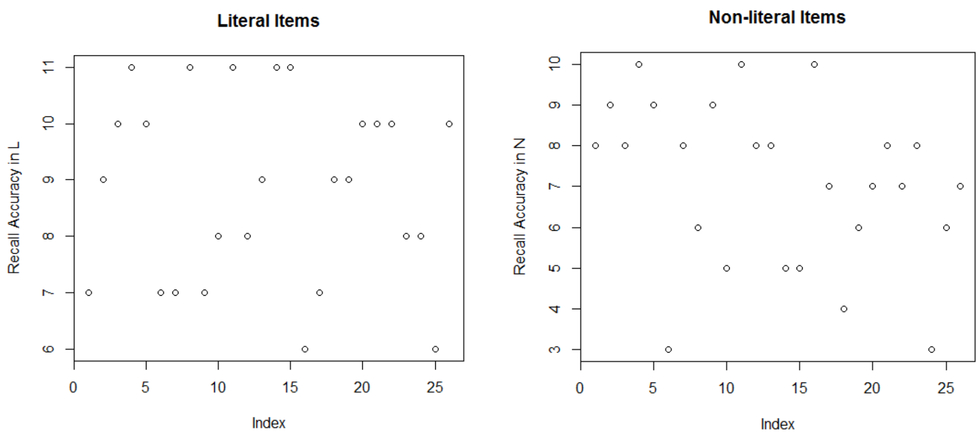


Figure 4. Scatterplots of literal and non-literal recall respectively. Participants (n = 26) are reported on the x-axis, while recall accuracy for items in the two translation conditions (n = 11 per group) is reported on the y-axis.

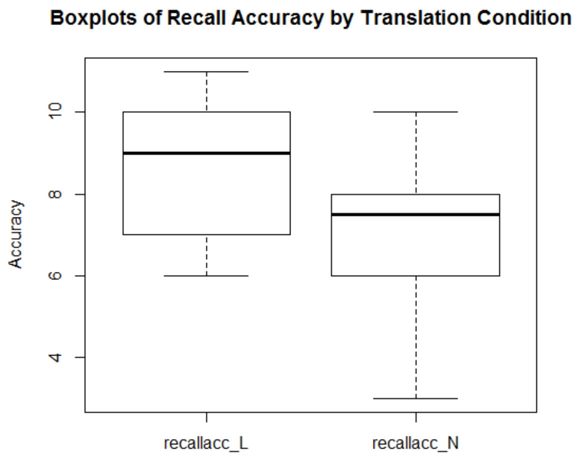


Figure 5. Boxplots for recall accuracy for literal and non-literal items.

The data points in fig. 4 are randomly scattered throughout the graphs, indicating the absence of a clearly non-normal pattern. Fig. 5 presents accuracy data by translation condition in a box-and-whisker plot. These plots present a lot of information at once, for example about the location and spread of the data, their distribution around the median, skewness and extreme outliers. The thick line inside each box represents the median, i.e. the cut-off point at which 50% of the data sit above it and 50% sit below. The boxplots also confirm the information presented in the descriptive tables, for example that the data have different min and max values. That is to say, when recalling literal items, subjects scored no less than 6 correct answers out of 11 in the post-test, whereas when recalling non-literal items their minimum accuracy score possible was 3 correct answers out of 11. The same goes for maximum values: whereas recall of literal items did in some subjects reach 100% accuracy (11/11 correct answers in the post-test), recall of non-literal items reached a maximum of 91% (10/11 correct answers). Moreover, one can get a sense of the potential skewness of the data from the whiskers. These are dashed vertical lines extending out of the boxes, which end in a horizontal stroke. If these have roughly the same length, data are not skewed. Data with a positive skew would have a longer whisker in the positive direction than in the negative direction. Here, both plots look fairly symmetric, especially literal data. The lower whiskers in the non-literal data are slightly longer than the upper ones, indicating a mild negative skew. Thanks to the boxplots one can also safely conclude that there are no extreme outliers (points plotted individually above or below the whiskers) in the recall data.

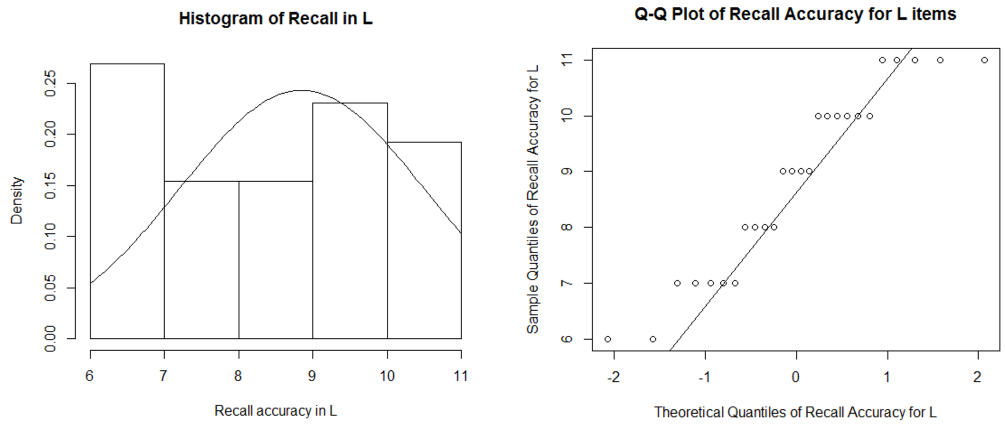


Figure 6. Histogram and Q-Q plot of recall accuracy for literally translated items.

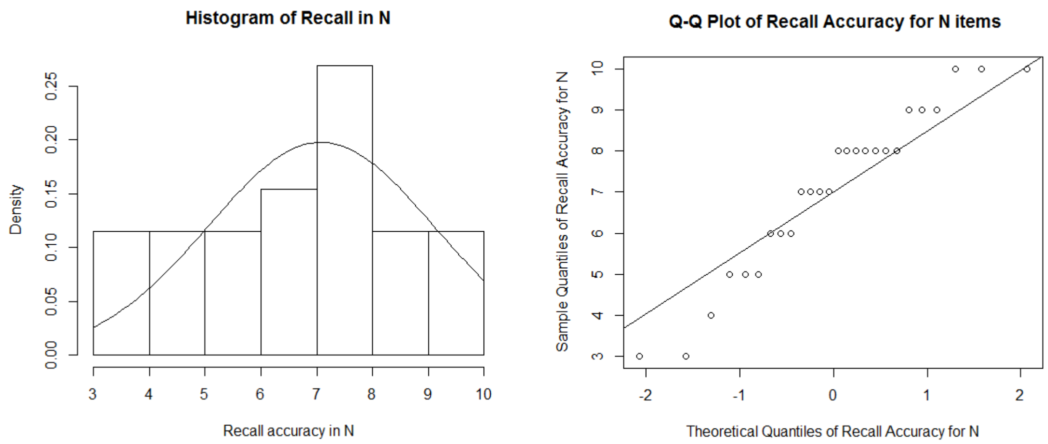


Figure 7. Histogram and Q-Q plot of recall accuracy for non-literally translated items.

Histograms are another way of assessing the shape and distribution of the data. The graphs were overlaid with a normality curve to assess whether data conform to a normal distribution, which seems to be the case in both conditions. However, histograms do not provide an exact measure of normality and should therefore be used in conjunction with numerical measures and other visual data representations (Larson-Hall, 2010: 77). The Shapiro-Wilk goodness-of-fit test is one of the most common tests of normality. According to Ricci (2005), it is the most powerful normality test for studies with small sample sizes (under 50). If the p-value is less than 0.05 the null hypothesis is rejected and the alternative hypothesis that the data are not normal must be accepted. Results show that non-literal recall data is normal ($W = 0.93$, $p > 0.05$) whereas literal recall data presents some departure from normality ($W = 0.91$, $p < 0.05$). However, because these formal tests often have low power, "we cannot be any surer that a distribution is exactly normal by looking at the numerical results of these tests than we can of looking at a histogram. For this reason, it is important not just to rely on a formal test of normality but to examine graphics as well" (Larson-Hall, 2010: 84-85). In a perfectly normal Gaussian distribution, mean and median are exactly the same. With real-life data, because of inherent variability, a *perfectly* normal distribution is extremely rare, and most distributions labelled as 'normal' are *de facto* an approximation of absolute normality. Here, mean and median are very similar in both cases (Med = 9.00, M = 8.84 in L; Med = 7.07, M = 7.50 in N), which suggests an approximation of normality. Skewness and kurtosis were also investigated. The skewness test reports a very mild left-skew to the recall data in N (-0.43), and an even smaller skew in L (-0.18). Usually, a negative skew means that the mean is smaller than the median, which is confirmed in these data. However, with skewness levels under 1 normality is not violated (Bulmer, 1979; Porte, 2002; Brown, 2015). The kurtosis levels registered are also quite low (-0.75 for N and -1.36 in L) and present no alarming cause for concern. The curves appear very mildly platykurtic, meaning that the central peak of the curve is slightly lower and broader, and its tails are shorter and thinner, which is confirmed visually by the two histograms. Recall data for both literal and non-literal items was also plotted in Q-Q plots (fig. 6 and 7) to provide further information on normality. If the sampling distribution is similar to the normal distribution, the points should fall along a straight line. However, only with large sample sizes the points will conform closely to the reference line; with small samples, even if they are drawn from a known normal distribution, there can be enough variation so that the data points will show some departure from the reference line (Larson-Hall, 2010: 82). In both the L and N recall conditions, there is some departure from the line, but the Q-Q plots reveal a roughly linear pattern. After a thorough exploration of these two variables, there is very little evidence of critical departure from normality, therefore the data in question can be considered roughly normal. A barplot for the recall data was also obtained (fig. 8 below). Barplots are useful to visually show the difference in mean score for different groups (8.84 in literal and 7.07 in non-literal recall).

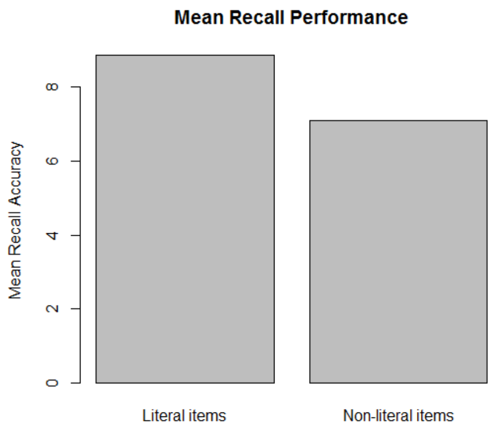


Figure 8. Barplot of recall accuracy by translation condition.

Next, a t-test was carried out on these mean differences between literal and non-literal recall scores. The paired t-test is used when the samples or sets of data used to produce the difference scores are linked in the population through repeated measurement, natural association, or matching (McDonald, 2014). The paired t-test was chosen in this analysis because each participant contributed a score in both literal and non-literal recall, and because this is a by-subject analysis. The paired t-test assumes that the *differences* between pairs are normally distributed; to validate this assumption, the variable recall accuracy difference was produced and plotted:

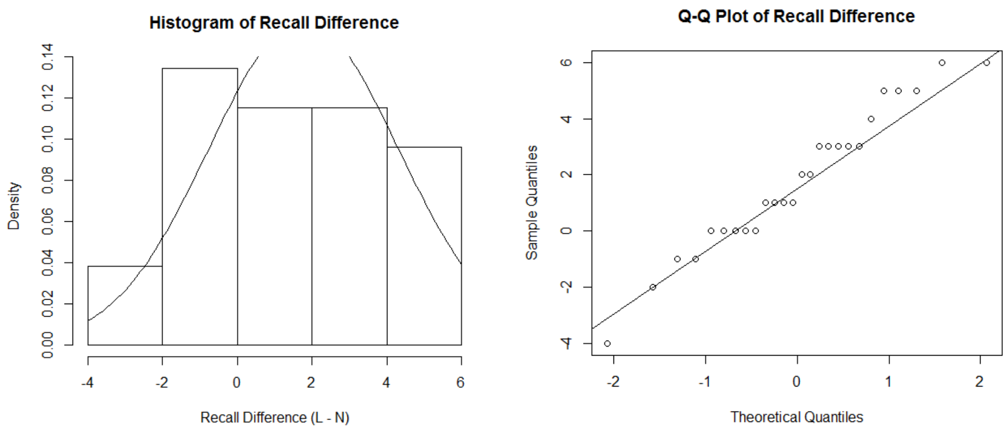


Figure 9. Histogram and Q-Q plot of recall accuracy difference between L and N.

From visual inspection of fig. 9, the data look roughly normal: the shape of the histogram suggests a bell curve and the values in the Q-Q plot do not excessively diverge from the straight normality line. Moreover, a test of both skewness and kurtosis reveal no cause for concern: the shape of the curve is almost entirely normal, with a negligible skew (-

0.10) and a small level of kurtosis (-0.68). The difference in recall between conditions (L – N) is confirmed to be normally distributed. The t-test shows a highly significant difference (95% CI: 0.74, 2.79; $t = 3.5$, $df = 25$, $p = 0.001$) between mean recall accuracy for literal ($M = 8.84$, $SD = 1.64$) and non-literal ($M = 7.07$, $SD = 2.01$) items. Effect size for this test was also calculated. The effect size is increasingly considered the main finding of a quantitative study (Sullivan and Fein, 2012: 279). It allows researchers to quantify the magnitude of mean differences and is therefore essential to ascertain the true importance of an effect. Whenever possible, effect sizes should be presented alongside p-values (Coe, 2002); while the latter indicate that a certain treatment affects people, they do not give any indication about *how much* people are affected by that treatment. In other words, p-values indicate whether an effect exists, but do not reveal the size of that effect. This t-test has a large effect size (Cohen's $\delta = 0.96$[23]). The power of the t-test was 0.88.

4.2.2 Working Memory

Working Memory (WM) is expected to relate to recall accuracy, as subjects who maintain and manipulate information active in memory more successfully may also later recall this information from memory more successfully (see 2.9). WM may therefore have an influence on the number of subtitles the subjects correctly identified in the recall tests. The WM test is presented in full in Appendix A.2, while the recall test is presented in Appendix A.3. First, the two variables were described and summarised, then their relationship was analysed.

[23] Given the mean and standard deviation for two samples of equal size, different measures of effect size can be calculated for a two-tailed t-test, such as Cohen's δ, Pearson's r or R^2. Pearson's r cannot be used with matched t-test (Yatani, 2015), Cohen's δ, i.e. the standardised mean difference, is to be preferred in this case. For a paired t-test, Cohen's δ can be calculated by dividing the mean of the differences between the two samples by the standard deviation of the differences between the two samples (ibid.). However, there is debate on how to calculate Cohen's δ in correlated designs, i.e. when the two groups are somewhat dependent or matched (Becker, 2015). Here, Dunlop et al.'s (1996) recommendation was followed and the original SD rather than the paired t-test value was used. Cohen (1988) suggested that $\delta = 0.2$ be considered a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size.

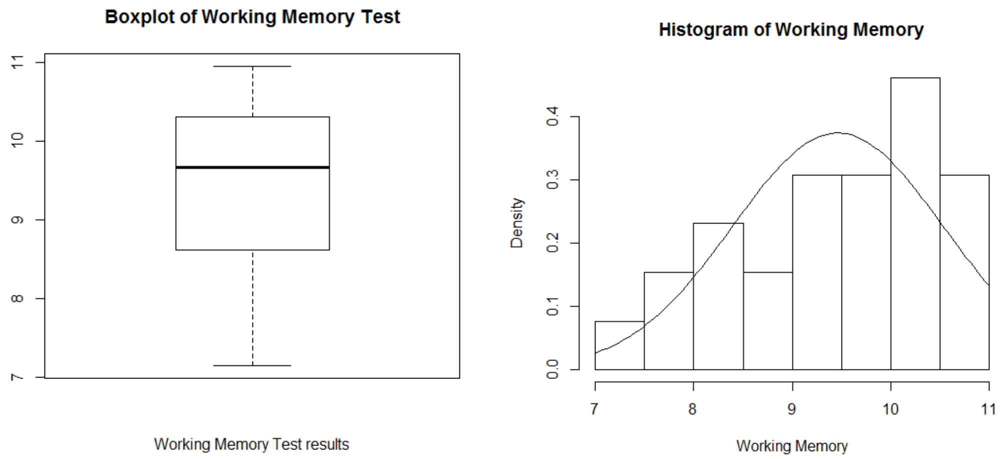


Figure 10. Descriptive graphs for WM data: boxplot and histogram.

The WM data above is appropriately spread out in the scatterplot (not reported), yet the boxplot in fig. 10 shows a difference in the length of the whiskers, with the bottom one being longer than the top one, suggesting a negative skew. This can also be noticed in the histogram, which has a more pronounced tail to the left, indicating negative skew. However, the relevant test shows only a mild skew (-0.4), fully below the reference threshold of 1. In the WM histogram (fig. 10) the peak of the curve appears slightly flat, thus confirming the kurtosis level (-0.9), which is nevertheless not alarming. Moreover, mean and median are very similar (9.462 and 9.670 respectively) and data follow the normality line quite closely in the Q-Q plot (not reported). Finally, the Shapiro-Wilk test confirms normality ($W = 0.95$, $p > 0.05$). Next, recall accuracy was considered.

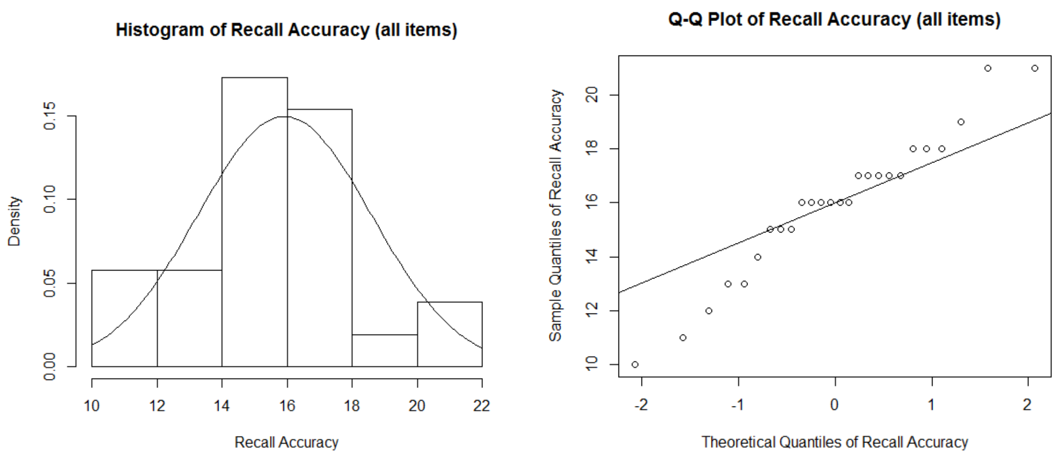


Figure 11. Histogram and Q-Q plot for recall accuracy.

The total overall accuracy scores obtained by each participant on all items in the memory test are considered, irrespective of translation condition. Mean and median are very similar (15.92 and 16.00 respectively), which suggests a normal distribution. Skewness and kurtosis tests did not reveal any abnormality in the data, and this can be verified visually in the histogram above (fig. 11), where a very symmetrical and mesokurtyc curve is displayed (i.e. the curve almost matches that of a Gaussian distribution). The Shapiro-Wilk test ($W = 0.95$, $p > 0.05$) further confirms that this variable roughly follows a normal distribution. It is interesting to notice that, although the data points in the Q-Q plot (fig. 11) show some departure from the reference line, the data is nevertheless normal. This is often the case with small samples, even when taken from a known normal distribution, and aptly illustrates how a single descriptive measures is relative and needs to be considered in conjunction with others to provide an unbiased picture of the situation.

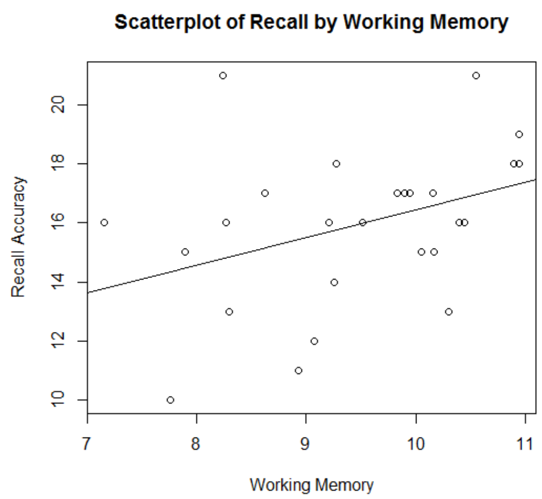


Figure 12. Scatterplot of recall accuracy by working memory. The tilt of the line provides a visual estimate of the correlation between the variables.

WM was plotted against recall accuracy to obtain a visual representation of their relationship and examine to what extent the variables vary together. Although the data points do not lie in a perfectly straight line, there is a relationship between the two variables. The scatterplot above shows that, as WM increases, recall accuracy also increases; subjects with higher WM scores therefore tend to achieve better performance in the post-test. Fig. 12 also reveals that there is no other type of non-linear relationship in the data, e.g. a curvilinear or U-shaped distribution. Therefore, the linear relationship between these two variables can be appropriately tested through correlation. The Pearson's product-moment correlation between working memory and total recall accuracy approaches significance (95% CI: -0.01, 0.66; $r = 0.37$, $n = 26$, $p = 0.06$).

Interestingly, although the confidence interval formally crosses zero, it is extremely close to be entirely positive, and the Pearson's correlation coefficient shows a considerable effect size. Although the test is not statistic, a positive relationship between WM and recall accuracy is confirmed by the scatterplot and has an effect between medium and large ($R^2 = 0.14$)[24]. The reason why the p-value is not statistic can be found in what appears to be a clear issue of power and sample size. As demonstrated by the power analysis, at the standard ∝ level of 0.05, despite a correlation coefficient of 0.37, with only 13 subjects per group the test has very little power (0.24) to detect genuine differences and patterns occurring in the data. A power level of 0.24 means a 76% change of not finding the real result, despite the effect size being large. Moreover, the CI (-0.01, 0.66) is very close to being entirely positive. This suggests that, if more participants had taken part in the experiment, the results of this test may show a significant test statistic, with a CI not spanning zero.

4.2.3 Language Proficiency

Both graphical and numerical descriptives confirmed that Italian proficiency data is roughly normally distributed. The variable was then plotted against recall performance.

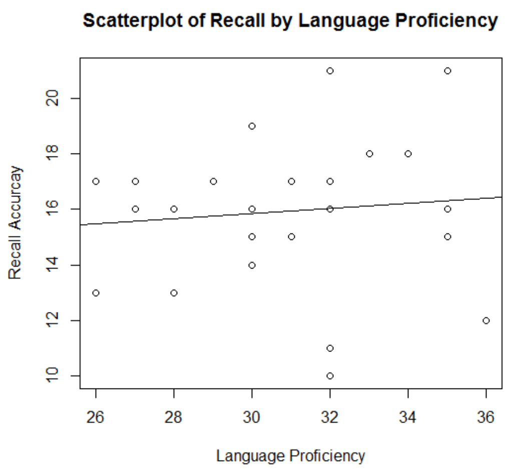


Figure 13. Scatterplot of recall accuracy by language proficiency.

---

[24] R^2 is calculated by squaring the correlation coefficient r and represents a measure of the proportion of variance in one variable that is accounted for by the other variable (Coe, 2002). According to Cohen (1992), $R^2 = 0.01$ indicates a small effect size, $R^2 = 0.09$ a medium effect and $R^2 = 0.25$ or higher a large effect. Cohen also encouraged research authors to identify effect magnitude scales in their own fields of study. In second language research, there is agreement that R^2 of 25% or more is to be considered large, while R^2 between 1-5% denote a small effect (Larson-Hall, 2010: 162).

From visual inspection of fig. 13, no clear relationship between the variables emerges. The slope of the line suggests that there may be some correlation but the data points remain scattered, and many sit far away from the reference line. A correlation test showed a very small relationship between the variables (95% CI: -0.29, 0.46; $r = 0.1$, $n = 26$, $p = 0.6$). The Pearson's coefficient is small ($r = 0.1$), consequently the R^2 indicates a small effect size ($R^2 = 0.01$). It seems that this test of language proficiency has little predictive power on recall performance. That is to say, while some proficient Italian L2 users also registered a high recall score in the verbatim memory test, other proficient L2 users recalled relatively few items correctly (for the proficiency test, see Appendix A.1; for the verbatim recall test, see Appendix A.3). The scatterplot also shows some subjects who obtained a middle score in the language test but recalled more items correctly than other more proficient Italian L2 users. However, a power analysis demonstrates that this test has almost negligible chances (power = 0.06) of finding an effect. A power of 6% is equivalent to a 94% chance of failure at detecting a true relationship in these data. Further power testing shows how, to detect a medium effect size ($r = 0.03$) at the standard α level with a desirable power of 0.8, one would need to gather data from 84 participants.

4.2.4 Fixation Measures (by-subject)

4.2.4.1 Fixation Count and Recall Accuracy

The descriptive statistics and graphics reveal fixation count to be a normally distributed variable. The scatterplot of overall fixation count and recall accuracy is presented below.

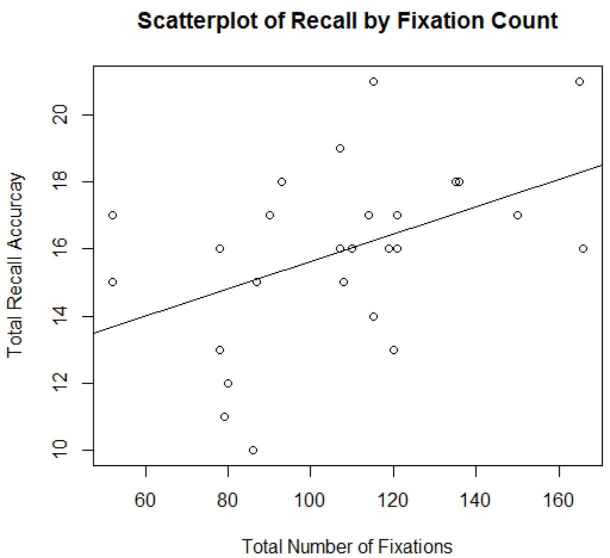


Figure 14. Scatterplot of recall accuracy by fixation count.

The graph in fig. 14 reveals the presence of a pattern between the variables. As seen previously, although the points do not fall exactly on the line, there is a clear linear relationship, with the pronounced tilt in the line indicating a positive correlation. The graph shows that the more fixations a subject makes on an item, the more likely they are to correctly recall its form in the post-test. The relevant correlation test confirms this assumption (95% CI: 0.07, 0.71; r = 0.4, n = 26, p = 0.02) and reveals a high correlation between fixation count and overall accuracy, with a large effect size ($R^2 = 0.20$). It is therefore possible to conclude that looking more at the subtitles results in more chances of their wording being remembered, regardless of the way they were translated.

4.2.4.2 Fixation Count and Translation Condition

The number of fixations for each translation condition were examined. First, the two relevant variables were plotted and checked for normality.

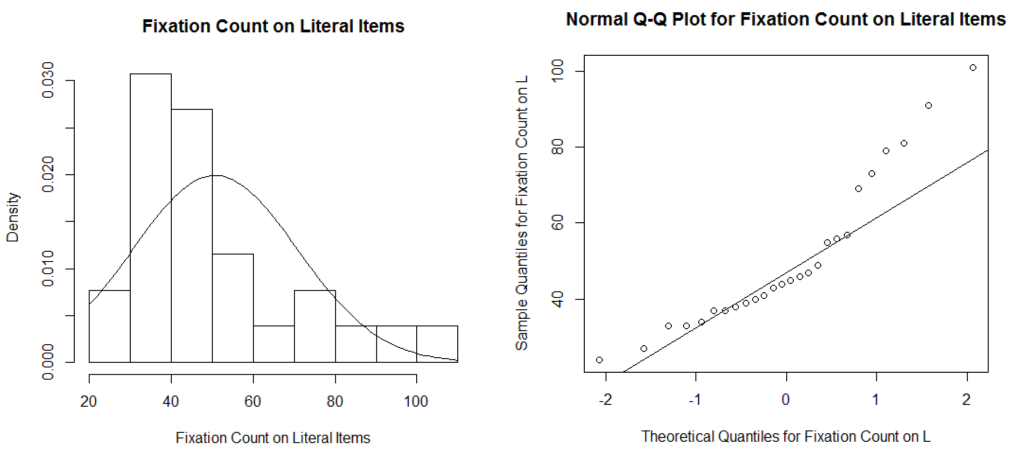


Figure 15. Histogram and Q-Q plot of fixation count on literal items.

As far as fixation count on literal items is concerned, the data showed some departure from normality. Firstly, mean and median are quite different (50.73 versus 44.50). In the Q-Q plot (fig. 15), the points follow the normality line quite closely but at the top of the graph there is a clear upwards bend. According to Verzani (2004), if the right tail is long the Q-Q line will curve up. This is verified visually in the histogram, where some degree of positive skew is evident (the right tail of the curve is longer). In Bulmer (1979), Porte (2002) and Larson-Hall (2010) amongst others, with skewness levels under 1 normality is not violated. In this dataset, while the Shapiro-Wilk goodness-of-fit test indicates a departure from normality (W = 0.89, $p < 0.05$), a formal skewness test reveals a positive skew level (0.95) that remains under 1. On the other hand, fixation count on non-literal items showed a much more clearly normal pattern. Mean and median are closer to each other (see table 8), skewness (0.39) and kurtosis (-0.38) levels are very low and fig. 16 reveals that there are no outliers in the boxplot under N. The Q-Q plot and histogram show a very mild right skew and the Shapiro-Wilk test confirms normality. A comparison table is also provided below.

Description	Fixation Count (L)	Fixation Count (N)
Mean	50.730	56.346
Median	44.5	53.5
Mode	33 and 37	46 and 65
Variance	401.164	447.675
Standard deviation	20.029	21.158
Sum of Squares	10029.12	11191.88
Standard Error	3.928	4.149
Min value	24	19
Max value	101	109

Table 8. Descriptives for by-subject total fixation count on literal and non-literal items.

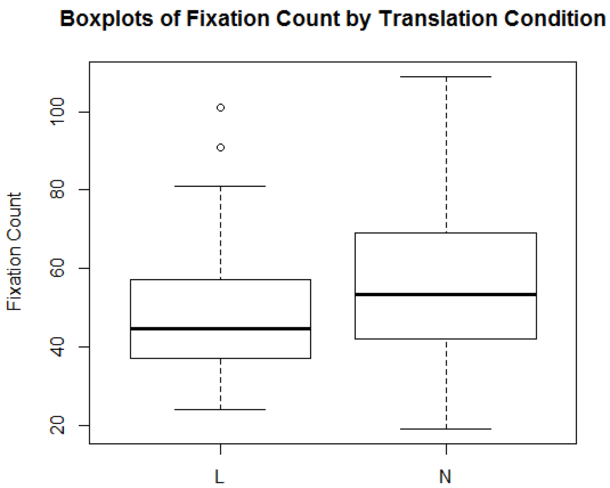


Figure 16. Boxplots of fixation count on literal and non-literal items.

The two variables were plotted together in fig. 16. The joint boxplots show that there is a wider range of values for non-literal items, i.e. overall subjects fixate a minimum of 19 and a maximum of 109 times on divergent items, whereas they fixate a minimum of 24 and a maximum of 101 times on literal items. Both variables have a slight positive skew, their medians and means are quite different and there are two extreme values in L, meaning two subjects made an unusually high number of fixations on items translated literally. Before carrying out a paired t-test between the number of fixations made by subjects on literal and non-literal items, test assumptions were checked. The paired t-test assumes that the *differences* between pairs of scores are normally distributed. After a visual and numerical examination of fixation count difference through a histogram, skewness (-0.02) and kurtosis (-0.38) tests, Q-Q plot and Shapiro-Wilk test ($W = 0.92$, $p > 0.05$), not enough evidence of non-normality was found. The t-test assumptions have therefore been met. The t-test shows no sign of a significant difference (95% CI -17.22, 5.99; $t = -0.9$, $df = 25$, $p > 0.05$) between total fixation count for literal ($M = 50.73$, $SD = 20.02$) and non-literal ($M = 56.34$, $SD = 21.15$) items. The negative sign of the t-statistic indicates that subjects fixated more on non-literally translated items, although this difference appears not significant. Cohen's δ is 0.19, indicating an effect of small magnitude. The power of the paired t-test was 0.09 so very small, suggesting that, even if there was a significant difference in fixation number between translation conditions, this particular test would have very slim chances of finding it.

4.2.4.3 Fixation Duration and Recall Accuracy

All the descriptive statistics and graphics reveal fixation duration to be a roughly normally distributed variable. The scatterplot of total fixation duration and recall accuracy is presented below.

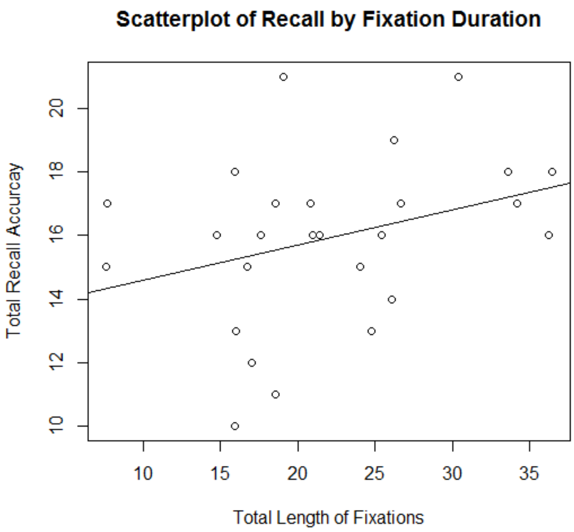


Figure 17. Scatterplot of recall accuracy by fixation duration (seconds).

Although the points do not fall exactly on the regression line, fig. 17 shows the presence of a linear positive pattern between the variables. It appears that, the longer the fixations a subject makes on the test items, the more likely they are to correctly recall their form in the post-test. The Pearson's correlation test confirms this assumption and reveals a moderate positive correlation between total fixation duration and overall accuracy, whose coefficient is however not statistic (95% CI: -0.06, 0.63; $r = 0.32$, $n = 26$, $p > 0.05$). Despite the non-significance of the p-value, the confidence interval almost does not cross zero and the correlation is of medium size ($R^2 = 0.10$). Again, the reason why the p-value is not statistic may be due to an issue of power and sample size. As demonstrated by the power analysis, at the standard α level of 0.05, despite a correlation coefficient of 0.32, with only 26 subjects power is not high (0.36) and may not detect genuine trends in the data. A power level of 0.36 means a 64% change of not finding the real result, despite the effect size being moderate. In conclusion, it appears that looking at a subtitle for longer results in more chances of its wording being remembered, regardless of the way it was translated. This relationship does not appear to be statistic, but the test does not have sufficient power to make this conclusion fail-safe. That is to say, if more participants had taken part in the experiment, the test results may have produced a significant test statistic with a CI not spanning zero.

4.2.4.4 Fixation Duration and Translation Condition

The length of the fixations (in seconds) made on each translation condition were examined. First, the two relevant variables were checked for normality.

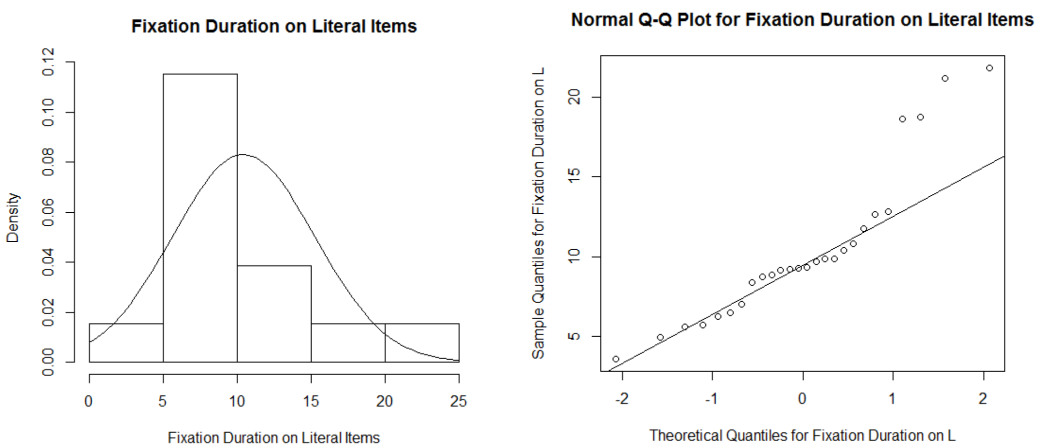


Figure 18. Histogram and Q-Q plot of fixation duration on literal items.

As far as fixation duration on literal items is concerned, the data showed some departure from normality, although mean and median are not too far apart from each other (9.27 versus 10.42). Similarly to what was found for fixation count on literal items, the points in the Q-Q plot (fig. 18) follow the normality line quite closely but at the top of the graph there is a clear upwards bend, indicating a longer right tail. This can be seen in the histogram as well, where the right tail of the curve is longer, again suggesting some degree of positive skew. The Shapiro-Wilk goodness-of-fit test confirms departure from normality ($W = 0.86$, $p < 0.05$), and a skewness level just over 1 also formally indicates a violation of normality, although not extreme. On the other hand, fixation duration on non-literal items showed a normal pattern. Mean and median are very close to each other (see table 9), skewness (0.2) and kurtosis (-0.6) levels are very low and the Shapiro-Wilk test ($W = 0.97$, $p > 0.5$) confirmed normality. Overall, a pattern similar to that found for fixation count was found for fixation duration; both fixation number and length appear to be departing slightly from normality when investing literal items, whereas they follow a normal distribution in the case of non-literal items. A comparison table is also provided below.

Description	Fixation Duration (L)	Fixation Duration (N)
Mean	10.415	11.626
Median	9.275	11.985
Mode	NA	NA
Variance	23.081	26.673
Standard deviation	4.804	5.164
Sum of Squares	577.040	666.833
Standard Error	0.942	1.012
Min value	3.58	2.72
Max value	21.83	23.68

Table 9. Descriptives for by-subject total fixation duration on literal and non-literal items.

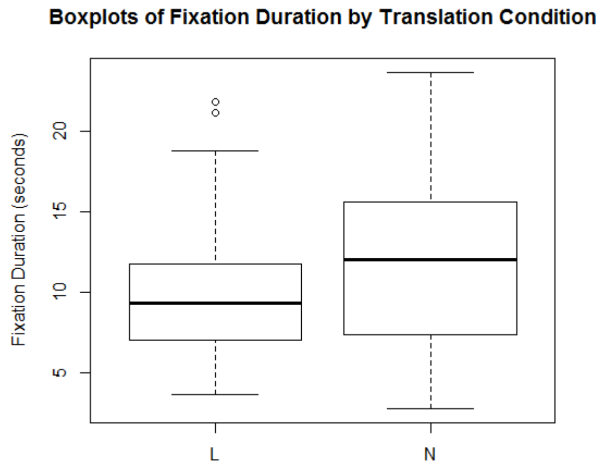


Figure 19. Boxplots of fixation duration on literal and non-literal items.

The two variables are plotted together in fig. 19. As was the case with fixation count, the fixation duration boxplots also show that there is a wider range of values for non-literal items. That is to say, subjects fixate for a minimum of 2.72 seconds and a maximum of 23.68 seconds on divergent items, whereas they fixate a minimum of 3.58 seconds and a maximum of 21.83 seconds on literal items. Both variables have a slight positive skew and there are two extreme values in L in fig. 19, meaning two subjects made unusually longer fixations overall on items translated literally. Before carrying out a paired t-test between total fixation length on literal and non-literal items, test assumptions were checked. Fixation duration differences were calculated and enough evidence that these differences follow a normal distribution was gathered through a histogram, skewness and kurtosis tests, Q-Q plot and Shapiro-Wilk test ($W = 0.94$, $p > 0.05$). No significant difference was found by the paired t-test (95% CI -3.71, 1.29; $t = -0.9$, $df = 25$, $p > 0.05$) between total fixation duration on literal ($M = 10.41$, $SD = 4.8$) and non-literal ($M = 11.62$, $SD = 5.1$) items. The negative sign of the t-statistic indicates that subjects fixated for longer on non-literally translated items, although this difference appears not significant. Cohen's δ is 0.19, indicating an effect of small magnitude, like in the t-test carried out on fixation counts. The power of the t-test was very small (0.09), again suggesting that, even if there was a significant difference in fixation duration between translation conditions, this test would have little chance of finding it. The t-test result for fixation duration mirrors that of fixation count quite closely. It makes sense for this to be the case, since, when a higher number of fixations occur, the overall duration will also be higher. There should therefore be a clear and very strong linear relationship between fixation count and duration. Said relationship is verified both by the clear linear trend in fig. 20 and the highly significant results of the correlation test (95% CI 0.8, 0.95; r = 0.90, n = 26, p < 0.001) between overall fixation count (M = 107, SD = 29.51) and duration (M = 22, SD = 7.81), with an R^2 of 0.22, meaning that a large percentage of the variance (22.68%) is explained by the correlation between these two variables.

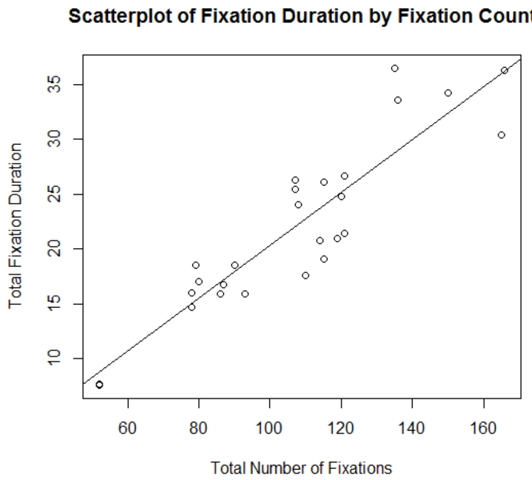


Figure 20. Scatterplot representing the linear relationship between fixation duration (y-axis) and fixation count (x-axis).

4.2.5 Order of Presentation and Fixation Measures

The counterbalanced design of this study allowed us to explore the experimental stimulus in greater detail. Subjects were randomly assigned to one of two groups: group G1L was exposed to literal subtitles in the first half of the clip and non-literal in the second; the opposite applied for group G2N (see 3.5). This allowed to factor order of presentation of the subtitled stimuli in the fixation analysis.

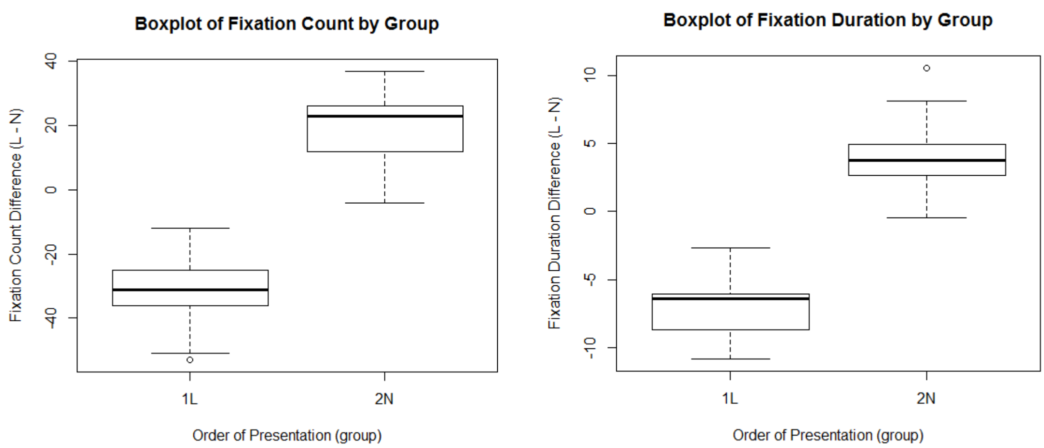


Figure 21. Boxplots of fixation count and fixation duration differences by order of stimulus presentation.

The independent two-sample t-tests on both fixation count (95% CI -61.07, -40.15; $t = -9.98$, $df = 24$, $p < 0.001$) and duration (95% CI -13.17, -8.58; $t = -9.80$, $df = 23$, $p < 0.001$) differences by group were highly significant. The two boxplots (fig. 21) reveal a similar situation. As it is shown in the fixation count plot, the difference in number of fixations for subjects in group G1L is negative (M = -30.92), whereas for group G2N it is positive (M = 19.69). Said difference is calculated by subtracting total fixation count on non-literal items from total fixation count on literal items (L - N) in each subject, so if the sign of the final figure for a subject is negative, that subject overall made more fixations on N items, and vice versa if the figure is positive, they made more fixations on L items. The difference in sign in the boxplots between G1L and G2N means that subjects from both groups made more fixations on the second half of the clip, which contained nonliteral (N) items for G1L and literal (L) items for G2N. Summary table 10 further illustrates the situation for fixation count:

	Fixation Count on all test items (n = 22)	**Fixation Count on 1st half test items (n = 11)**	**Fixation Count on 2nd half test items (n = 11)**
Group G1L	Both halves (total)	LITERAL	NONLITERAL
	1444 (sum)	521 (sum)	923 (sum)
	111.07 (mean)	40.07 (mean)	71 (mean)
Group G2N	Both halves (total)	NONLITERAL	LITERAL
	1340 (sum)	542 (sum)	798 (sum)
	103.07 (mean)	41.69 (mean)	61.38 (mean)

Table 10. Number of fixations by group (order of stimulus presentation).

The areas with the highest fixation count per translation condition are highlighted for both groups. Group G1L fixated more times on items in N (appearing in the second half of the clip, marked in green in the table) than on items in L: on average, subjects in this group made 71 total fixations on N items and only 40 on L items. Conversely, group G2N fixated more times on items in L (also appearing in the second half of the clip, marked in blue in the table) than on items in N: on average, subjects in this group made 61 total fixations on L items and only 42 on N items. So although the total fixation counts on all 22 test items do not show a remarkable difference between groups (average 103 total fixations per subject in G2N against 111 in G1L), a more in-depth analysis of the stimulus reveals a pattern behind order of presentation. The same applies to fixation duration (table 11).

	Fixation Duration (s) on all test items (n = 22)	**Fixation Duration (s) on 1st half test items (n = 11)**	**Fixation Duration (s) on 2nd half test items (n = 11)**
Group G1L	Both halves (total)	LITERAL	NONLITERAL
	304.24 (sum)	108.94 (sum)	195.4 (sum)
	23.40 (mean)	8.38 (mean)	15.03 (mean)
Group G2N	Both halves (total)	NONLITERAL	LITERAL
	268.74 (sum)	106.89 (sum)	161.85 (sum)
	20.67 (mean)	8.22 (mean)	12.45 (mean)

Table 11. Duration (sec) of fixations by group (order of stimulus presentation).

On average, subjects in G1L fixated for 15 seconds overall on N items (second half) and only 8s on L items. Conversely, people in G2N fixated 12s overall on L items (second half) and only 8s on N items. Thus, an analysis of eye movements by group seems to confirm that subjects fixate more and for longer on subtitles appearing in the second half of the clip, regardless of translation condition, which is explained by the fact that, in the second half of the clip, subtitle duration is on average noticeably higher. This is due to more dialogue occurring in the source video, which is inevitably linked to the number of words appearing in each subtitle. The by-item analysis allowed a much more in-depth analysis of this finding, making it possible to draw more detailed conclusions on whether this lack of balance between halves had an impact on recall accuracy. To this purpose, correlations and t-tests on a number of by-item variables were carried out, i.e. subtitle duration, literal and non-literal subtitle length as well as linguistic category (4.3.6).

4.3 By-item Analysis

4.3.1 Recall Accuracy and Translation Condition (by-item)

Test item	AOI	Recall accuracy (no. Y/13)		Recall accuracy (%)		Difference in %	Tot Accuracy (no. Y/26)
		Literal	Non-literal	Literal	Non-literal		
1	5	2	7	15.3%	53.8%	-38.5%	34.61%
2	13	9	11	69.2%	84.6%	-15.4%	76.92%
3	17	12	12	92.3%	92.3%	0.0%	92.30%
4	19	12	8	92.3%	61.5%	30.8%	76.92%
5	26	11	5	84.6%	38.4%	46.2%	61.53%
6	32	11	4	84.6%	30.7%	53.9%	57.69%
7	39	11	6	84.6%	46.1%	38.5%	65.38%
8	42	9	10	69.2%	76.9%	-7.7%	73.07%
9	45	10	10	76.9%	76.9%	0.0%	76.92%
10	51	9	6	69.2%	46.1%	23.1%	57.69%
11	53	11	8	84.6%	61.5%	23.1%	73.07%
12	56	12	9	92.3%	69.2%	23.1%	80.76%
13	60	13	11	100%	84.6%	15.4%	92.30%
14	65	9	8	69.2%	61.5%	7.7%	65.38%
15	70	12	9	92.3%	69.2%	23.1%	80.76%
16	85	10	9	76.9%	69.2%	7.7%	73.07%
17	89	11	5	84.6%	38.4%	46.2%	61.53%
18	91	10	11	76.9%	84.6%	-7.7%	80.76%
19	94	11	8	84.6%	61.5%	23.1%	73.07%
20	96	11	9	84.6%	69.2%	15.4%	76.92%
21	99	12	7	92.3%	53.8%	38.5%	73.07%
22	105	12	11	92.3%	84.6%	7.7%	88.46%
TOT	**All subs**	**230**	**184**	**80.4%**	**64.3%**	11.9%	72.37%

Table 12. By-item recall performance table by translation condition (across participants). The column 'Recall accuracy' shows how many participants out of the total in each group (13) correctly identified (Y) each item.

The cells in dark blue in table 12 indicate the highest recall rate for each item. The colour coding shows the mnemonic superiority of literal (formally similar) items in the post-tests. The vast majority of subtitles (16 out of 22) are remembered more accurately when translated literally than non-literally. The only four exceptions are AOI.5, AOI.13, AOI.42, AOI.91, which are recalled better when translated non-literally. In the discussion (chapter 5), more space will be given to these items, in particular AOI.5. In two cases (AOI.17 and AOI.45, marked in light blue), recall accuracy is equal between the two conditions. The last column of the table shows the total accuracy for each individual item regardless of translation condition, measured as number of correct responses over all participants (n = 26). The second-last column of the table shows accuracy difference in percentages; here, numbers with a negative sign refer to those test items with more accurate recall in the diverging translation condition (N). For these, however, higher non-literal accuracy is only slight; only AOI.5 shows a fairly substantial 38% difference, whereas the non-literal AOI.13 differs by 15% from its literal counterpart and AOI.42 and AOI.91 only by 8%. In most of these exceptions, therefore, the proportion of accurate answers in N is only slightly higher than in L. Moreover, there are other items worth pointing out from table 12 (AOI.26, AOI.32, AOI.39, AOI.51, AOI.89). Not only are these remembered correctly by the majority of people when translated literally, but also, interestingly, they have the poorest recall rate when translated non-literally. From the 46.1% for AOI.39 and AOI.51 to the lowest 30.7% for AOI.32, these subtitles are recalled correctly by less than half the participants who were exposed to them in the non-literal condition. The non-literal versions of these items might create confusion, resulting in most people not being able to correctly recognise the divergent translation in the recall test, further reinforcing the superiority of the literal condition. These by-item findings confirm those of the by-subject analysis, strongly suggesting that, after one single exposure to reversely subtitled input, literal translations of most items are remembered more precisely.

		Recall		**Totals**
		Correct (Yes)	**Incorrect (No)**	
Translation condition	**Literal**	230	56	286
	Non-literal	184	102	286
Totals		414	158	572

Table 13. Summary table of overall recall accuracy by translation condition.

In the literal condition, an overall 80% accuracy (correct recall n = 230) was registered across participants and items, against a 64% accuracy in the divergence condition (correct recall n = 184). The total responses elicited were 572 (26 subj x 22 sub), 286 per translation condition (see table 13). Of these, 158 (27.6%) were incorrectly and 414 (72.3%) correctly identified. Recall accuracy data was further analysed to gather information on its shape and suitability for inferential statistical testing. Further descriptives for the variable recall accuracy are presented in table 14.

Description	Recall accuracy (L)	Recall accuracy (N)
Mean	10.45	8.364
Median	11	8.500
Mode	11	8,9,11 (multimodal)
Variance	4.926	5.004
Standard deviation	2.219	2.237
Sum of Squares	103.454	105.090
Standard Error	0.473	0.476
Min value	2	4
Max value	13	12

Table 14. Descriptives for by-item recall data with literal and non-literal subtitles.

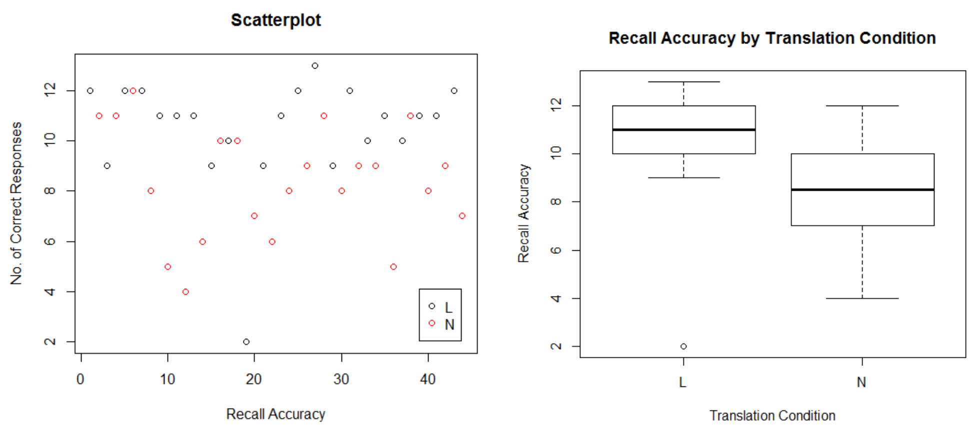


Figure 22. Scatterplot and boxplots of recall accuracy by translation condition. In the first plot, all 44 renderings are included, the 22 literal renderings are marked in black, the 22 non-literal ones are marked in red.

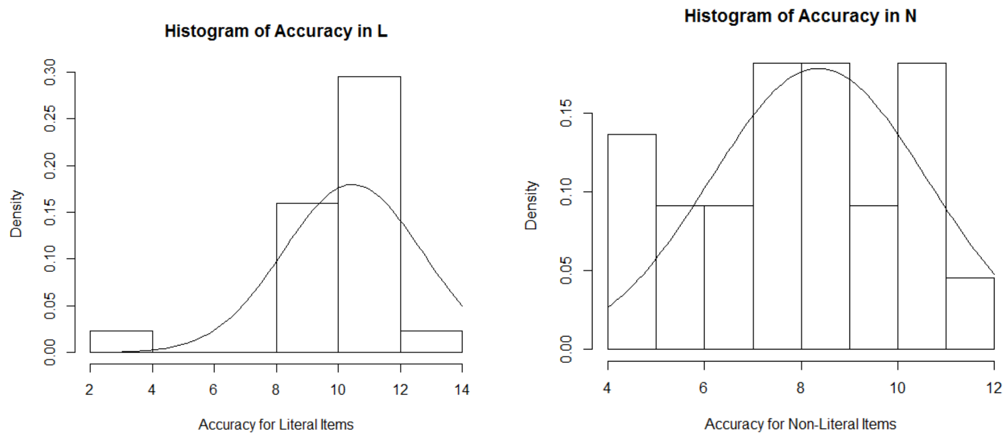


Figure 23. Histograms of recall accuracy for literal and non-literal items respectively.

From the scatterplot in fig. 22, one can gather that the shape and distribution of data changes between the two conditions. Whereas non-literal data (red dots) are scattered, literal data (black dots) are concentrated at the top of the graph. The boxplots visually reiterate that, overall, items are better recalled when translated literally. The single dot reported at the bottom of the literal boxplot indicates an outlier, AOI.5, as can be expected from inspection of table 7. With the exception of this subtitle, the minimum accuracy value for literal items is 9 (out of 13 subjects), whereas the minimum for non-literal items is 4. A similar (but less pronounced) situation occurs for maximum values. Whereas the literal item with the highest accuracy rate was recalled by all 13 subjects in that condition (100%), the non-literal item with the highest accuracy rate was recalled by 12 subjects (92.3%). The top and bottom of the boxplot whiskers show these differences, visually indicating that the full range of values is greater with diverging items. Thus, recall scores clearly vary more in the non-literal condition. The medians are in the middle of the box in both cases, which would suggest that the data is not strongly skewed. Moreover, mean and median are similar to each other in both cases (Med = 11, M = 10.45 in L; Med = 8.5, M = 8.36 in N), which suggests an approximation of normality. For non-literal data, the Q-Q plot, histogram and relevant tests confirm normality (skewness: -0.2; kurtosis: 1; Shapiro-test: $W = 0.95$, $p > 0.05$). However, in the literal data, the tests reveal some anomalies (skewness: -2; kurtosis: 6; Shapiro-test: $W = 0.70$, $p < 0.001$). These were confirmed visually by the Q-Q plots and the histograms. The literal data in fig. 23 appears to be leptokurtic, i.e. to have a sharper peak compared to a normal distribution, and a visible negative skew (longer left tail). After this graphical and numerical exploration of the variables, the mean differences were compared to ascertain whether this noticeable difference in recall between conditions is significant. The t-test shows a highly significant result (95% CI 0.80, 3.38; $t = 3.3$, $df = 21$, $p = 0.002$) between recall accuracy for literal ($M = 10.45$, $SD = 2.21$) and non-literal ($M = 8.36$, $SD = 2.23$) items. The test has a large effect size (Cohen's $\delta^2 = 0.93$), and the power of the test was .79.

4.3.2 Fixation Measures (by-item)

4.3.2.1 Fixation Count and Translation Condition

The total number of fixations on the whole clip (110 subtitles) by all participants is 14513. Items translated literally are fixated 7176 times, non-literally translated ones 7337. Each item is fixated 65.23 times across participants in the literal and 66.7 in the non-literal condition. Scatterplots, histograms and Q-Q plots showed that fixation count data in both literal and non-literal conditions are not normally distributed and positively skewed. Formal tests of skewness and kurtosis are above 1 in both conditions, and Shapiro tests in L ($W = 0.92$, $p < 0.001$) and N ($W = 0.91$, $p < 0.001$) also confirm non-normality.

Because the data cannot be considered normally distributed, a parametric t-test cannot be applied. One of the most common alternatives to the unpaired two-sample t-test is the Wilcoxon rank-sum test, also known as the Mann-Whitney test (Whitley and Ball, 2002). On these data, the Mann-Whitney test showed no significant difference (95% CI -9, 7; W = 5953, p > 0.5) between total fixation count on literal (M = 65.23, SD = 36.12) and non-literal (M = 66.7, SD = 35.83) items.

The same relationship between translation condition and fixation count was then investigated considering only the 22 subtitles on which participants answered recall questions. The total number of fixations on these test items by all participants is 2765. Items translated literally are fixated 1307 times in total, non-literally translated ones 1458. Each item is fixated 59 times across participants in the literal and 66 in the non-literal condition. At it is evident from the boxplot, this difference is not very large and may therefore not be significant.

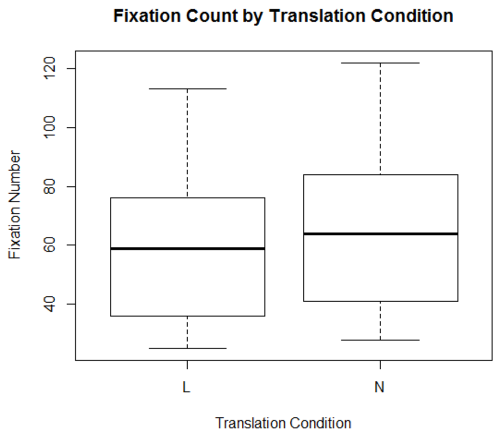


Figure 24. Boxplots of fixation count on literal and non-literal items respectively (22 test items).

The variable fixation count was found to be normally distributed in both translation conditions in all graphical and formal normality tests. The homoscedasticity assumption was also met, as indicated by the Levene's test (F = 0.674, p = 0.4). The two group variances are equal, so both independent t-test assumptions are met. The boxplot (fig. 24) shows that the medians of the two groups are close. So are the two ranges as well as the maximum and minimum values. Test items are allocated a similar number of looks in both conditions, and this difference in amount of fixations is not significant, as demonstrated by the t-test (95% CI -22.42, 8.69; t = -0.89, df = 42, p > 0.05).

4.3.2.2 Fixation Count and Recall Accuracy

Test item	Subtitle	Fixation Counts by Recall Accuracy		Fixation count difference (acc-Y – acc-N)
		Yes	**No**	
1	AOI.5	28	46	-18
2	AOI.13	48	14	34
3	AOI.17	67	7	60
4	AOI.19	105	44	61
5	AOI.26	78	50	28
6	AOI.32	40	23	17
7	AOI.39	82	28	54
8	AOI.42	51	11	40
9	AOI.45	83	24	59
10	AOI.51	35	21	14
11	AOI.53	127	42	85
12	AOI.56	192	43	149
13	AOI.60	150	9	141
14	AOI.65	85	26	59
15	AOI.70	168	37	131
16	AOI.85	117	35	82
17	AOI.89	73	38	35
18	AOI.91	176	20	156
19	AOI.94	105	33	72
20	AOI.96	115	37	78
21	AOI.99	112	40	72
22	AOI.105	90	10	80
	Totals	**2127**	**638**	

Table 15. Fixation count by recall performance (across participants).

In table 15, the number of fixations is split between items recalled correctly (Y) and incorrectly (N). The 22 test items are fixated 2765 times in total across participants. Of this total, correctly recalled items receive 2127 fixations, whereas incorrectly recalled items receive only 638. Therefore, the more a subtitle is looked at the higher the chances of it being remembered accurately. It is interesting to notice that these findings are very consistent, the only exception being AOI.5, which was remembered more accurately when fixated less times. Fixation count data were described both visually and numerically in each group (acc-Y and acc-N, i.e. correct or accurate and incorrect or inaccurate recall, respectively) and were not found to violate the normality assumption. Data in both these conditions have a negligible level of skew (0.4 in acc-Y and -0.02 in acc-N), and while kurtosis levels are slightly higher (-0.7 and -1.3 respectively) the data still follow a roughly normal distribution, as confirmed by the Shapiro tests ($W = 0.95$, $p > 0.05$ in acc-Y; $W = 0.94$, $p > 0.05$ in acc-N). The independent t-test assumes an approximately normal distribution of the dependent variable in both groups as well as homogeneity of variance. The Levene's test reveals heteroscedasticity ($F = 14.777$, $p < 0.0005$), so homogeneity of variance cannot be assumed and the t-test default option with degrees of freedom adjustment in R was therefore chosen. To get a more visual idea of the size of this difference in fixation number, the data are displayed together in fig. 25. These boxplots visually confirm that correct items (acc-Y) have a much wider range of values and receive a much higher number of fixations compared to their counterparts (acc-N).

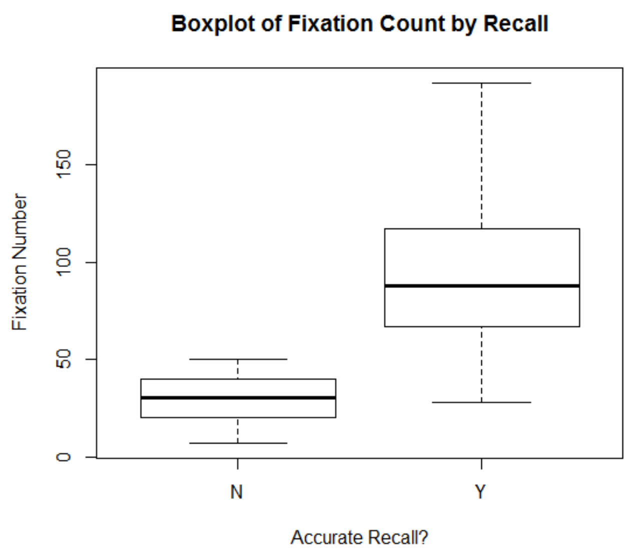


Figure 25. Boxplots of fixation count by recall accuracy (22 subtitles).

The independent t-test was carried out to establish whether this difference in fixation number between accurately and non-accurately remembered items is of statistical significance. A very significant difference in fixation number (95% CI 46.75, 88.61; $t= 6.66$, $df = 24.5$, $p < 0.001$) between correct ($M = 96.68$; $SD = 45.72$) and incorrect ($M = 29$; $SD = 13.29$) items emerged, with a large effect size ($\delta = 2.0$) and power .99. It is possible to conclude that the more frequently a subject looks at a subtitle, the more chances they have of remembering its wording.

4.3.2.3 Fixation Duration and Translation Condition

On average, each fixation made during the whole video (110 subtitles) was around 217ms long. A total of 26min (1,552,511ms) were spent fixating literal items, against 27min (1,607,414ms) on non-literal items. Although non-literal items are fixated more, this difference does not look very large. Before carrying out the relevant t-test, descriptives for the variable fixation duration were analysed. Histograms and Q-Q plots revealed a positive skew for fixation duration on both literal (1.14) and non-literal items (1.23). High kurtosis levels were recorded both in L (1.56) and N (1.46). This departure from normality is confirmed by the Shapiro-Wilk tests ($W = 0.91$, $p < 0.001$ in L; $W = 0.90$, $p < 0.001$ in N). Because it is not possible to conclude that these fixation duration data are normally distributed, the non-parametric equivalent of the t-test was computed, namely the Mann-Whitney test. Its results showed no significant difference (95% CI -1948, 1509; $W = 5939$, $p > 0.5$) between total fixation duration on literal ($M = 14113$ms, $SD = 7365$) and non-literal ($M = 14612$ms, $SD = 7681$) items.

The same relationship between translation condition and fixation duration was then investigated considering only the 22 subtitles on which participants answered recall questions. The total sum of fixation durations on all these test items in both conditions by all participants is roughly 10min (606180ms). Items translated literally are fixated for 4.76min (286637ms) in total, non-literally translated ones for 5.31min (319543ms). Each item is fixated for around 13sec (13028.95ms) across participants in the literal and 14sec (14524.68ms) in the non-literal condition. At it is evident from the boxplot in fig. 26, this difference is not very large and may therefore not be significant.

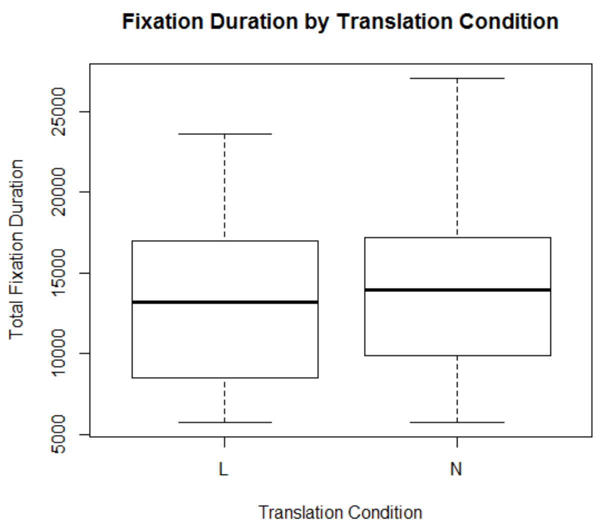


Figure 26. Boxplots of Fixation duration (ms) by translation condition (22 subtitles).

Fixation duration on the 22 test items was found to be normally distributed in both translation conditions in all graphical and formal normality tests. The homoscedasticity assumption was also met, as indicated by the Levene's test ($F = 0.45$, $p = 0.5$). The two group variances are equal, so both independent t-test assumptions are met. Fig. 26 shows that the medians of the two groups are similar. So are the two ranges as well as the minimum values. Test items are fixated for similar amounts of time in both conditions, and this difference in fixation duration is not significant, as demonstrated by the t-test (95% CI -4762.93, 1771.47; $t = -0.92$, $df = 40$, $p > 0.05$). A similar situation to that of fixation count was therefore found for fixation duration. That is to say, the total number and length of the fixations made by the participants do not change significantly depending on translation condition, neither in the test subset (22 subtitles) nor throughout the whole experimental stimulus (110 subtitles).

4.3.2.4 Fixation Duration and Recall Accuracy

Test item	Subtitle	Total Fixation Durations by Recall Accuracy (ms)		Fixation duration difference (acc-Y – acc-N)
		Yes	No	
1	AOI.5	6936	11466	-4530
2	AOI.13	12432	4047	8385
3	AOI.17	17608	1782	15826
4	AOI.19	21511	11068	10443
5	AOI.26	17978	10527	7451
6	AOI.32	7212	4264	2948
7	AOI.39	17582	5280	12302
8	AOI.42	11517	2697	8820
9	AOI.45	17536	5070	12466
10	AOI.51	8810	3931	4879
11	AOI.53	27555	9080	18475
12	AOI.56	41102	9601	31501
13	AOI.60	30998	1333	29665
14	AOI.65	17962	5830	12132
15	AOI.70	35141	8096	27045
16	AOI.85	23442	7221	16221
17	AOI.89	15248	7895	7353
18	AOI.91	36644	3939	32705
19	AOI.94	22158	7876	14282
20	AOI.96	26188	9092	17096
21	AOI.99	28060	7867	20193
22	AOI.105	22216	2382	4879
	Totals	**465836**	**140344**	

Table 16. Fixation duration (ms) by recall performance (across participants).

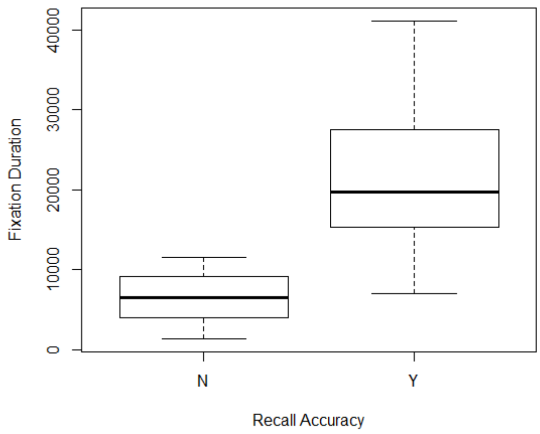


Figure 27. Boxplots of fixation duration (ms) by recall accuracy (22 subtitles).

Table 16 reveals that in almost all cases correctly recalled items are fixated for longer (with the exception of AOI.5). Overall, correct items (acc-Y) are fixated for 7.7 min (465836ms) in total, while incorrect ones (acc-N) for 2.3 min (140344ms). The boxplot in fig. 27 indicates a considerable difference in the duration of fixations allocated to correct items compared to incorrect ones. The size and position of the boxes and their black middle line show this difference graphically, with the two medians being far apart (Med = 19744ms for accurate items, Med = 6525ms for non-accurate items). The different lengths of the whiskers in the boxplot also show that fixation durations on correct items span over a much larger range of values than incorrect ones. The two means are also quite different (M = 21174ms and M = 6379ms respectively). Before carrying out an independent t-test to investigate whether this difference is statistic, test assumptions were checked. As far as normality is concerned, all plots and descriptive statistics suggest that the data follows a roughly normal distribution. The homoscedasticity assumption was tested using Levene's test. Based on the result ($F = 14.49$, $p < 0.0005$), we have to accept the alternative hypothesis that there is a significant difference in the variances between the two groups. Therefore, the independent t-test was carried out assuming inequality of variances and the degrees of freedom were adjusted according to the Welch correction for non-homogeneity of variance. The independent t-test (95% CI -19145, -10445; $t = 6.99$, $df = 25.4$, $p < 0.01$) confirms that there is a highly significant difference in fixation duration between items that were correctly recalled (M = 21174.36, SD = 9419.65) and those that were not (M = 6379.27, SD = 3099.35). Effect size for this test was 2.1 and the power of the test was also high (0.9). The results mirror those found for fixation count, suggesting that if items are fixated more times and for longer, they have a significantly higher chance of being remembered in a verbatim memory post-test.

4.3.3 Subtitle Duration

4.3.3.1 Subtitle Duration and Recall Accuracy

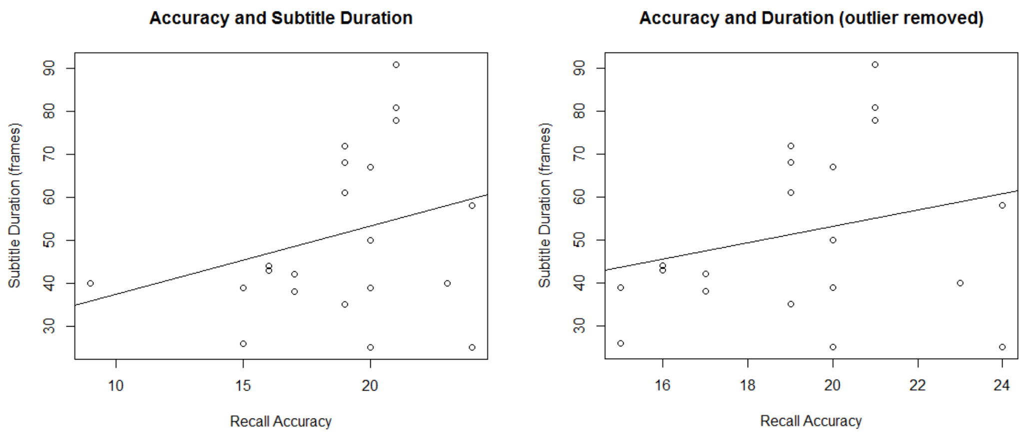


Figure 28. Scatterplot of recall by subtitle duration, measured in frames (fr). The outlier (AOI.5) has a duration of 40fr and a recall accuracy of 9.

Figure 29. Scatterplot of recall by subtitle duration (fr). The outlier has been removed.

The relationship between subtitle duration and accuracy is represented in fig. 28, which shows a feeble linear relationship. Fig. 28 confirms that AOI.5 (located at the bottom right of the plot) behaves differently from the other subtitles, being recalled correctly by only 9 subjects out of 26. A scatterplot of all 22 test items without the outlier is presented in fig. 29. However, because there is no strong basis on which to remove the point from the analysis, AOI.5 was left in the dataset. Moreover, the presence of this point does not considerably affect the normality distribution of these data. The data for both subtitle duration and recall accuracy followed a roughly normal distribution. Because removing the data point AOI.5 from this analysis is not justified theoretically, the non-parametric Spearman's rank-order correlation, which is much less sensitive to outliers, was run to determine the relationship between subtitle duration and recall accuracy for the 22 test items. No significant correlation between the variables was found ($r_s = 0.26$, $n = 22$, $p > 0.05$), i.e. as duration of the subtitles becomes longer, there is no systematic accuracy increase or decrease. Therefore, while the scatterplots show a linear relationship between these two variables, this is quite weak and not significant.

4.3.3.2 Subtitle Duration and Subtitle Length

While subtitle duration remained the same in the two translation conditions, subtitle length had to be changed between translation conditions to allow for variation between literal and non-literal renderings. To keep the two translation conditions comparable, renderings differed by one, two or three words maximum. A total of 57 subtitles presented a difference between number of words between conditions: 34.5% of these (38 on 110) present a one-word difference, 12% (13 on 110) a two-word difference and 4.5% (5 on 110) a three-word difference between renderings. On the other hand, almost half of the subtitles, i.e. 48% of the total (53 on 110) presented no difference between conditions, i.e. had the same number of words in both literal and non-literal renderings. Only in one case (1% of the total) subtitle length difference exceeded three words: in subtitle 41 (AOI 29) there is a six-word difference between conditions, as the literal translation is composed of 12 words whereas the non-literal one of six. AOI 29 was therefore excluded from this particular correlation analysis[25]. A substantial correlation between duration and length of the subtitles is to be expected, since the more words a subtitle contains, the longer it will have to remain on screen for comfortable reading on the part of the viewer. Correlation assumptions were checked first. The data clearly violated normality in both variables. Subtitle duration is naturally positively skewed as, conventionally, subtitles have a minimum duration of 20 frames (just under a second). Skewness (1.21) and kurtosis (1.35) tests have relatively high values and this is graphically represented in the histogram (fig. 30), where a clear right-tail and leptokurtic shape of the curve are visible. The Q-Q plot in fig. 30 also suggest departure from normality, which is further confirmed by the Shapiro-Wilk test ($W = 0.89$, $p < 0.001$).

[25] The total number of subtitles in the subtitle length analysis is 220 since each of the 110 items in the clip has been translated literally and non-literally, *de facto* resulting in 220 different wordings to be considered. Excluding AOI 29 means that the data set used in this specific correlation analysis has 218 subtitles in total rather than 220.

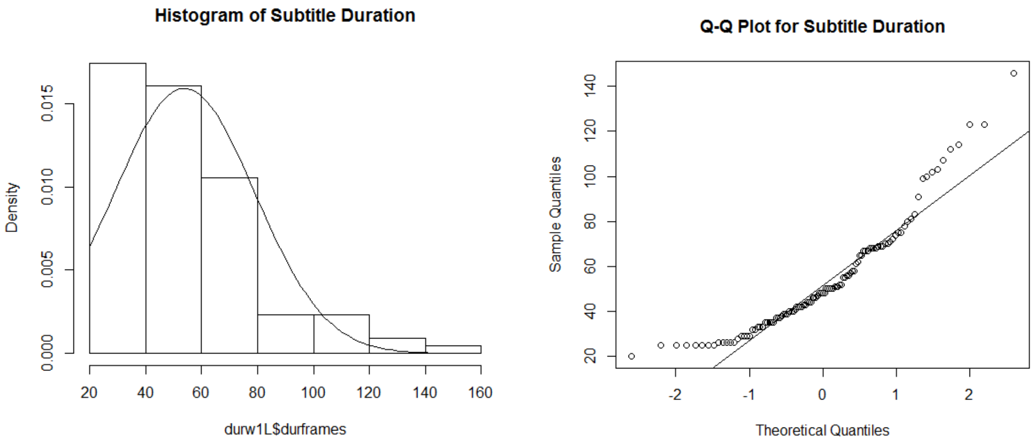


Figure 30. Histogram and Q-Q plot of subtitle duration (fr).

Subtitle length differs between translation conditions, therefore all graphs and plots were produced for literal and non-literal subtitles. In fig. 31, one can clearly see that literal length data are right-skewed and mildly leptokurtic, while non-literal data are right-skewed and platykurtic. All other descriptive measures also verified a non-linear pattern.

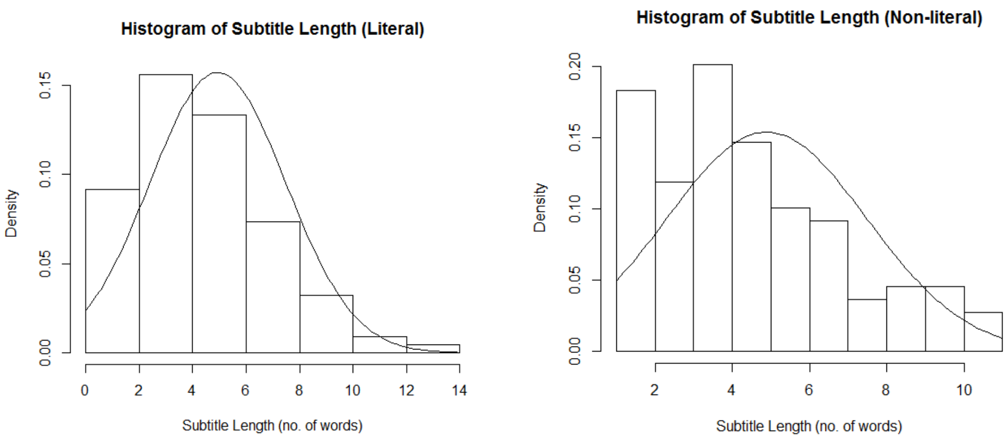


Figure 31. Histograms of subtitle length (no. of words) in the literal and non-literal wording respectively.

Because these data do not follow a normal distribution, the Pearson's product-moment correlation is not the best test to assess the relationship between subtitle duration and length. Its non-parametric alternative, the Spearman's rank-order correlation test is more appropriate in cases where data markedly violates normality assumptions. As expected, the results of this test confirm the presence of a clear positive and very strong relationship between subtitle duration and length ($r_s = 0.80$, $n = 218$, $p < 0.0001$). The power of a two-tailed correlation test with 218 observations and an r_s coefficient of 0.80 is 100% at the conventional alpha level of 0.05.

4.3.3.3 Subtitle Duration and Fixation Measures

4.3.3.3.1 Subtitle Duration and Fixation Duration

A high correlation between subtitle and fixation duration is to be expected since the longer a subtitle remains on screen the more time is available for a subject to fixate its words, and vice versa, the shortest its permanence on screen, the lesser the overall duration of a subject's fixations on that subtitle can be. Furthermore, it is commonly accepted both in academia (see Munday, 2009: 149) and in the industry (see, for example, Ofcom, 2015) that if subtitles stay on screen for too long, viewers might be inclined to re-read them (De Linde and Kay, 1999), which also results in an increase of fixation duration as subtitle duration itself increases. The main study data were aggregated as to obtain total fixation duration by all participants on each of the 110 subtitles.

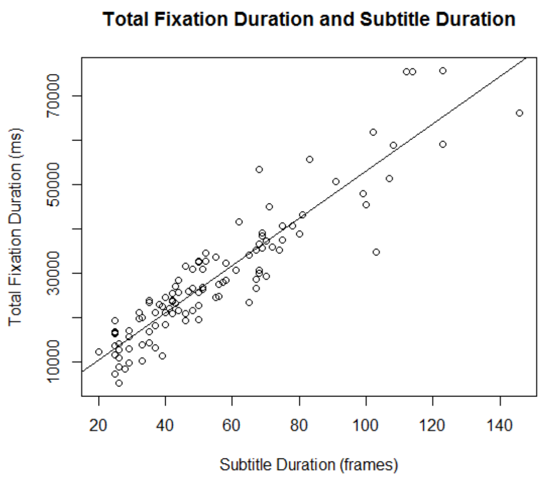


Figure 32. Scatterplot of subtitle duration (frames) and total fixation duration (ms).

Fig. 32 visually reveals the presence of a strong linear relationship between these variables. All graphical and numerical descriptive measures computed indicated that the Pearson's correlation assumptions were violated, since both variables do not follow a normal frequency distribution. Therefore, the Spearman rank-order test was used to assess the relationship between subtitle and fixation durations. As expected, a very strong and highly significant positive correlation ($r_s = 0.93$, $n = 110$, $p < 0.0001$) between the variables in question was confirmed by the test, with 100% power at the conventional significance level of 0.05.

4.3.3.3.2 Subtitle Duration and Fixation Count

A high correlation between subtitle duration and number of fixations is also to be expected, since the longer a subtitle remains on screen the higher the number of fixations that can potentially be made on it, and vice versa, the shortest it remains on screen, the lower the number of fixations that can physically be made on it. The same correlation analysis between subtitle duration and fixation duration in 4.3.3.3.1 was carried out between subtitle duration and fixation count for all 110 subtitles. It was found that neither of the two variables follows a normal distribution. The scatterplot in fig. 33 reveals a clear relationship between the variables, as expected. The non-parametric Spearman's test confirmed a statistic and very large positive correlation between fixation count and subtitle duration ($r_s = 0.93$, $n = 110$, $p < 0.0001$), again with 100% power at the conventional significance level of 0.05.

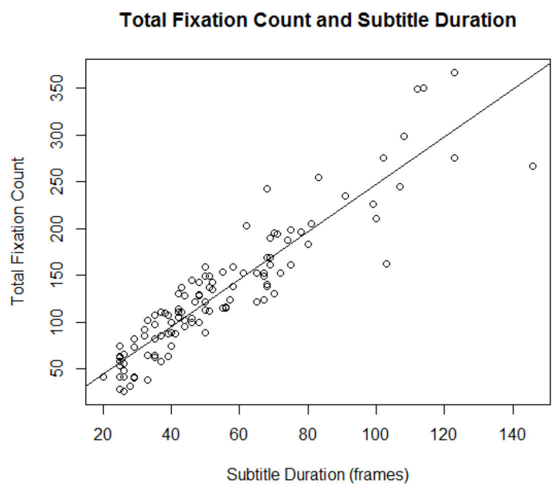


Figure 33. Scatterplot of subtitle duration (frames) and total fixation count.

4.3.4 Subtitle Length

4.3.4.1 Subtitle Length and Recall Accuracy

Since subtitle duration and length are strongly correlated, it is reasonable to assume that correlational results between length and accuracy are similar to those found for duration and accuracy. First, plots and normality tests were produced for the 22 subtitles on which recall data was collected. Neither subtitle length nor recall accuracy follows a normal distribution, with accuracy in particular being platykurtic (kurtosis = -1). The Shapiro-

Wilk tests confirms non-normal patterns (for subtitle length: $W = 0.92$, $p < 0.05$; for accuracy: $W = 0.91$, $p < 0.02$).

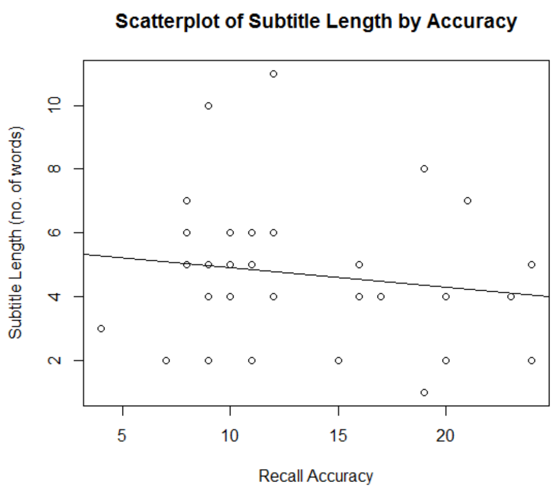


Figure 34. Scatterplot of subtitle length (no. of words) and recall accuracy.

Fig. 34 represents the relationship between subtitle length and recall. The angle of the line indicates a very slight negative correlation, yet the data are scattered throughout and do not clearly follow a linear relationship. The Spearman's test ($r_s = -0.12$, $n = 22$, $p > 0.05$) further demonstrates that there is no significant relationship between subtitle length and recall. The negative sign of the statistic confirms what is represented graphically in the scatterplot, namely that the direction of the relationship is negative, indicating that longer subtitles are remembered less accurately. However, this could be entirely due to chance, since the relationship is not statistically significant. This result is consistent with the non-significant results found between recall accuracy and subtitle duration (4.3.3.1).

4.3.4.2 Subtitle Length and Fixation Measures

In section 4.3.3.3 it was found that the longer a subtitle remains on screen, the more and the longer fixations it attracts. Furthermore, subtitle duration was found to correlate highly with subtitle length (4.3.3.2). Subtitle length should also correlate with fixation measures, since the more words there are in a given subtitle, the longer subjects may take to read it and the more fixations they may make on it, especially as the number of content words increases.

4.3.4.2.1 Subtitle Length and Fixation Duration

There are two sets of different translations for all 110 items, resulting in 220 total different wordings to be analysed. Visual plotting confirmed that the data are right-skewed in both cases, especially the fixation duration data, whose skewness (1.20) and kurtosis (1.57) are both at concerning levels. Non-normality is further demonstrated through the Shapiro-Wilk test for both subtitle length ($W = 0.94$, $p < 0.0001$) and fixation duration ($W = 0.91$, $p < 0.0001$). Since these data depart from normality, the correlation test adopted was the Spearman's rank-order test. As expected and as it can be inferred from fig. 35, a high positive correlation ($r_s = 0.86$, $n = 220$, $p < 0.0001$) was found between subtitle length ($M = 4.95$, $SD = 2.59$) and total fixation duration ($M = 14363.3$ms, $SD = 7511.72$). The size of the effect is large (0.86) and the power of the correlation test on 220 items at an α level of 0.05 is 100%.

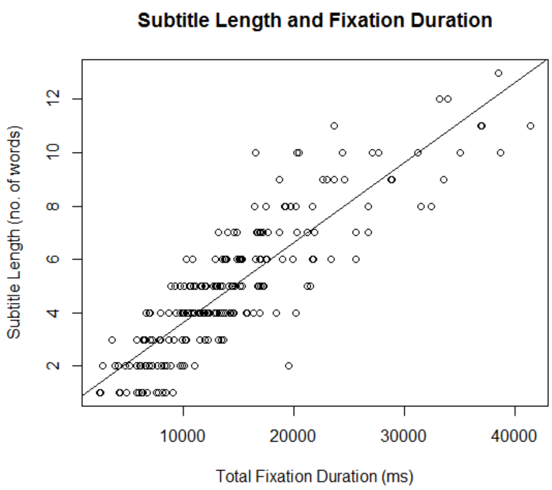


Figure 35. Scatterplot of subtitle length (no. of words) and total fixation duration (ms).

4.2.4.2.2 Subtitle Length and Fixation Count

The relationship between fixation count and subtitle length was also explored. The subtitle length data remain the same as in the previous section, so they are known not to be normally distributed. The Shapiro-Wilk test indicates that total fixation count also departs dramatically from normality ($W = 0.92$, $p < 0.0001$). Moreover, there are high levels of skew (1.12) and kurtosis (1.39). Fig. 36 visually represents the relationship between total fixation count and subtitle length. The data reveals a clear linear relationship, which is confirmed by the Spearman rank-order test ($r_s = 0.90$, $n = 220$, $p < 0.0001$). The size of this effect is very large (0.9) and the power of the correlation test on 220 items at α level = 0.05 is again 100%. The results thus establish that subtitle length behaves similarly to subtitle duration: the number of words in a subtitle have no significant relationship with recall accuracy, while being strongly correlated to fixation measures.

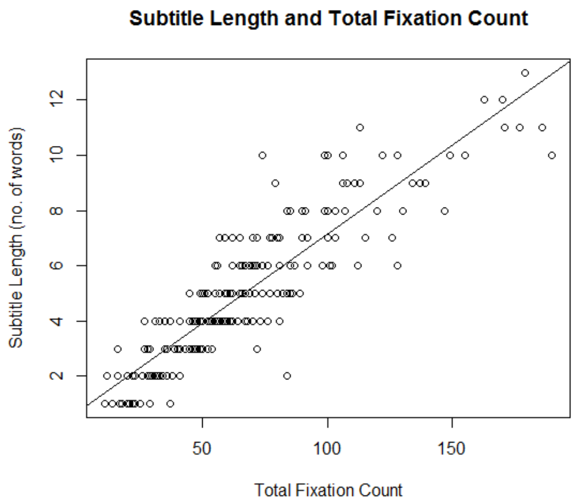


Figure 36. Scatterplot of subtitle length (no. of words) and total fixation count.

4.3.5 Linguistic Category

4.3.5.1 Linguistic Category and Recall Accuracy

Test item	Subtitle	Linguistic Category	Total accuracy (Y/26)	Total accuracy (%)
1	AOI.5	LEX	9	34.6%
2	AOI.13	LEX	20	76.9%
3	AOI.17	LEX	24	92.3%
4	AOI.19	LEX	20	76.9%
5	AOI.26	LEX	16	61.5%
6	AOI.32	LEX	15	57.6%
8	AOI.42	LEX	19	73%
11	AOI.53	LEX	19	73%
17	AOI.89	LEX	16	61.5%
19	AOI.94	LEX	19	73%
20	AOI.96	LEX	20	76.9%
22	AOI.105	LEX	23	88.4%
		TOTALS	**220**	
7	AOI.39	SYN	17	61.5%
9	AOI.45	SYN	20	76.9%
10	AOI.51	SYN	15	57.6%
12	AOI.56	SYN	21	80.7%
13	AOI.60	SYN	24	92.3%
14	AOI.65	SYN	17	65.3%
15	AOI.70	SYN	21	80.7%
16	AOI.85	SYN	19	73%
18	AOI.91	SYN	21	80.7%
21	AOI.99	SYN	19	73%
		TOTALS	**194**	

Table 17. By-item recall performance table by linguistic category (across participants). The numbers under 'Total accuracy' show how many participants out of the total (26) correctly identified (Y) each item.

Each of the 26 participants answered 22 verbatim memory questions, 12 on lexical and 10 on syntactic items respectively. Thus, the total possible responses per linguistic category is 12*26 = 312 for lexical and 10*26 = 260 for syntactic items. Lexical items were recalled correctly 220 times out of 312, whereas syntactic items were recalled correctly 194 times out of 260 (see table 17 above). Therefore, accuracy rates are 70% for lexical and 74% for syntactic items. It appears that overall, syntactic items were recalled slightly more accurately than lexical ones. If the potential outlier (subtitle AOI.5, lexical) is removed from the analysis, the situation remains the same; out of 21 total test items participants answered questions on 11 lexical and 10 syntactic items, with an accuracy rate of 73% and 74% respectively. Recall data by linguistic category were explored to determine if they follow a roughly normal distribution. Visual and numerical descriptives did not evidence lack of normality. Moreover, both Shapiro tests (W = 0.91, p > 0.05 for accuracy with lexical items; W = 0.96, p > 0.05 for accuracy with syntactic items) confirmed that the data does not violate this assumption. The variances are equal in both groups, as demonstrated by the Levene's test (F = 0.4284, p = 0.5). Fig. 37 graphically represents the information reported in table 17, visually confirming that the two linguistic categories are both recalled quite successfully and they both have the same maximum accuracy value. They also have the same minimum accuracy value, if one excludes the outlier. This extreme data point is AOI.5, the only subtitle correctly recalled by merely 9 subjects out of the total. Thus, the boxplots show two similar overall accuracy ranges as well as similar medians, suggesting that although syntax is recalled more accurately than lexicon overall, this difference may not be significant. Indeed, the t-test shows no statistical effect (95% CI -4.01, 1.87; t = -0.75, df = 19, p > 0.05) in levels of accuracy between lexical (M = 18.33, SD = 3.9) and syntactic items (M = 19.4, SD = 2.5). Effect size is small (Hedges' g[26] = 0.31) and so is the power of the test (0.1). This test does not, however, take into account translation manipulation into literal and non-literal items; there could be an interaction between translation condition and linguistic category as far as recall scores are concerned. This point is developed in the next section.

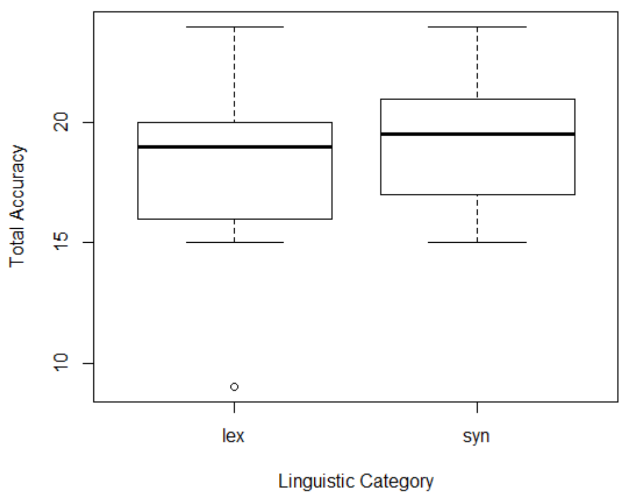


Figure 37. Boxplots of linguistic category and recall accuracy.

[26] In this dataset, the number of items differs slightly between the two groups (12 lexical and 10 syntactic). Hedges' g weighs effect size based on the relative size of each sample, and therefore constitutes an alternative to Cohen's δ when sample sizes differ (Stangroom, 2015).

4.3.5.2 Recall by Linguistic Category and Translation Condition

Test item	Subtitle	Linguistic Category	Recall accuracy (Y/13)		Total accuracy (Y/26)	Total accuracy (% correct)
			Literal	Non-literal		
1	AOI.5	LEX	2	7	9	34.6%
2	AOI.13	LEX	9	11	20	76.9%
3	AOI.17	LEX	12	12	24	92.3%
4	AOI.19	LEX	12	8	20	76.9%
5	AOI.26	LEX	11	5	16	61.5%
6	AOI.32	LEX	11	4	15	57.6%
8	AOI.42	LEX	9	10	19	73%
11	AOI.53	LEX	11	8	19	73%
17	AOI.89	LEX	11	5	16	61.5%
19	AOI.94	LEX	11	8	19	73%
20	AOI.96	LEX	11	9	20	76.9%
22	AOI.105	LEX	12	11	23	88.4%
		TOTALS	**122**	**98**	**220**	
7	AOI.39	SYN	11	6	17	61.5%
9	AOI.45	SYN	10	10	20	76.9%
10	AOI.51	SYN	9	6	15	57.6%
12	AOI.56	SYN	12	9	21	80.7%
13	AOI.60	SYN	13	11	24	92.3%
14	AOI.65	SYN	9	8	17	65.3%
15	AOI.70	SYN	12	9	21	80.7%
16	AOI.85	SYN	10	9	19	73%
18	AOI.91	SYN	10	11	21	80.7%
21	AOI.99	SYN	12	7	19	73%
		TOTALS	**108**	**86**	**194**	

Table 18. By-item recall performance table by linguistic category and translation condition (across participants). The numbers under "Recall accuracy" show how many participants out of the total in each translation condition group (13) correctly identified (Y) each item. The numbers under "Total accuracy" show how many participants out of the total (26) correctly identified (Y) each item.

Table 18 presents information regarding accurate recall broken down by subtitle, where the highest recall rates for lexical items are marked in orange, for syntactic items in green. This table allows to identify accuracy within each individual subtitle in a given linguistic category and translation condition. It also shows that syntactic item AOI.91, as well as lexical items AOI.5, AOI.13 and AOI.42 are better recalled non-literally, and that there are two subtitles where accuracy is equal (AOI.17 and AOI.45).

Translation condition	Category	Recall		Totals
		Yes	No	
L	lex	122 (78%)	34 (22%)	**156**
L	syn	108 (83%)	22 (17%)	**130**
Tot		**230**	**56**	**286**
N	lex	98 (63%)	58 (37%)	**156**
N	syn	86 (66%)	44 (34%)	**130**
Tot		**184**	**102**	**286**

Table 19. Summary table of recall accuracy by linguistic category and translation condition (across participants and items).

Table 19 presents a count summary collapsed across participants and items, where additional information about inaccuracy can be presented. Both tables show that, overall, accuracy levels in the literal condition are higher than in the non-literal condition regardless of linguistic category. When translated literally, both lexical and syntactical items are recalled quite well, i.e. the vast majority of responses are correct rather than incorrect (122 right against 34 wrong responses for lexical, and 108 against 22 for syntactical items). The total lexical observations are 312, of which 156 in the literal and 156 in the non-literal condition. The total syntactic observations are 260, of which 130 in the literal and 130 in the non-literal condition. Because these totals are different, the values in the table cannot be directly compared, so recall percentages have been added in. They show that syntactic items are recalled more accurately than lexical ones in both translation conditions (83% versus 78% in the literal condition, 66% versus 63% in the non-literal condition). This is particularly marked with formally transparent, literal items. To test whether the 12 lexical items are recalled better when translated literally or when translated non-literally, this subset of the data was extracted and analysed.

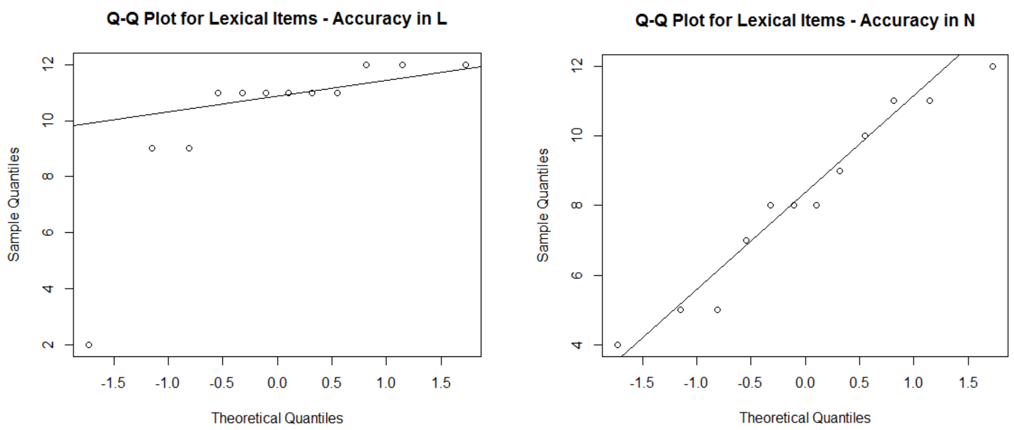


Figure 38. Q-Q plots of recall accuracy for lexical items in the literal and non-literal translation conditions respectively.

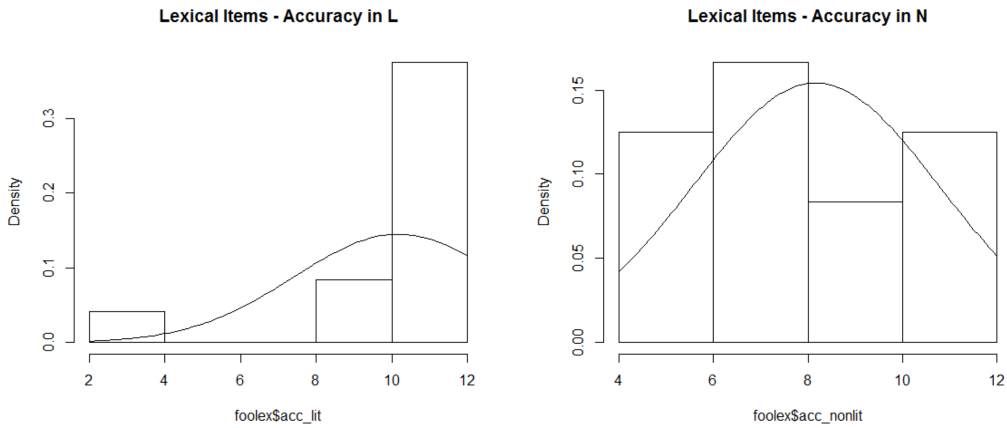


Figure 39. Histograms of recall accuracy for lexical items in the literal and non-literal translation conditions respectively.

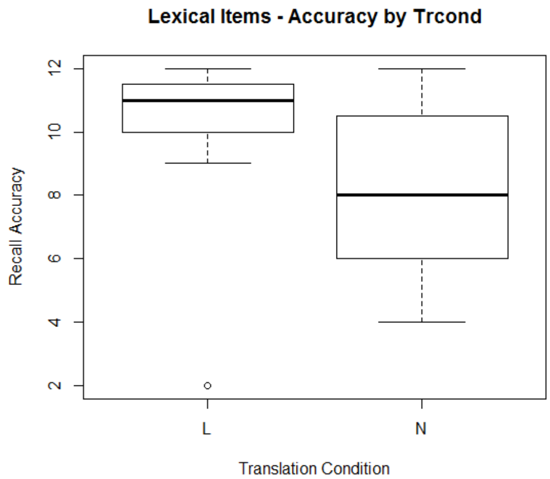


Figure 40. Boxplots of recall accuracy for lexical items in the literal and non-literal translation conditions respectively.

From an analysis of the relevant histogram and Q-Q plot above, as well as Shapiro, skewness and kurtosis tests, lexical data in the non-literal condition can be considered normal. Lexical data in the literal condition, however, clearly departs from normality (Shapiro test: W = 0.61, $p < 0.0005$), as it is also visible graphically in the relevant Q-Q plot and histogram in fig. 38 and 39. The latter, in particular, shows a pronounced level of negative skew (-2) in the histogram curve. The variances for recall of lexical items are equal in both groups, according to the Levene's test (F = 0.6069, p = 0.4), thus ensuring homoscedasticity. Fig. 40 clearly highlights the difference in recall scores between literal and non-literal lexical items, where the former are much higher and much less spread out than the latter. The two medians are also quite far apart in the plot. The t-test reveals no significant difference (95% CI -0.26, 4.26; t = 1.83, df = 21, p = 0.08) in levels of accuracy between lexical items translated literally (M = 10.16, SD = 2.75) and non-literally (M =

8.16, SD = 2.5). However, sample size limitations (n = 12) are evident here, making the test unlikely to be unable to detect a real effect. Moreover, data in one of the two groups is not normally distributed, so these results cannot be considered conclusive. Therefore, although the null hypothesis cannot be rejected, this does not mean that no effect exists, just that one could not be found in the small sample analysed. Finally, although the p-value is not statistic, the result approaches significance (p = 0.08) and the confidence interval is very close to being entirely positive (not crossing zero). The effect size was calculated for power analysis purposes and turned out to be medium-to-large (δ = 0.74). The power of the test is, as expected, quite low (0.41), meaning there is a 59% chance of not finding a genuine effect, even if there is one and its size is considerable.

To test whether the 10 syntactic items are recalled better when translated literally or when translated non-literally, the same analysis as above was carried out for this data subset.

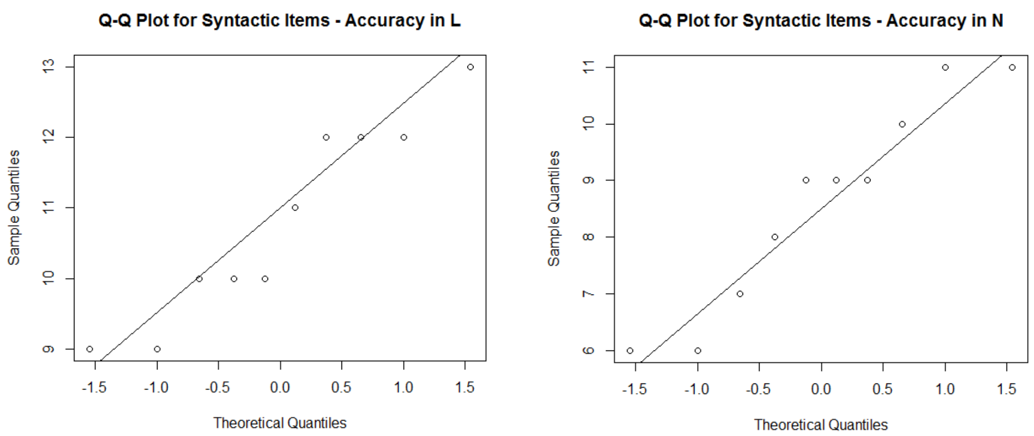


Figure 41. Q-Q plots of recall accuracy for syntactic items in the literal and non-literal translation conditions respectively.

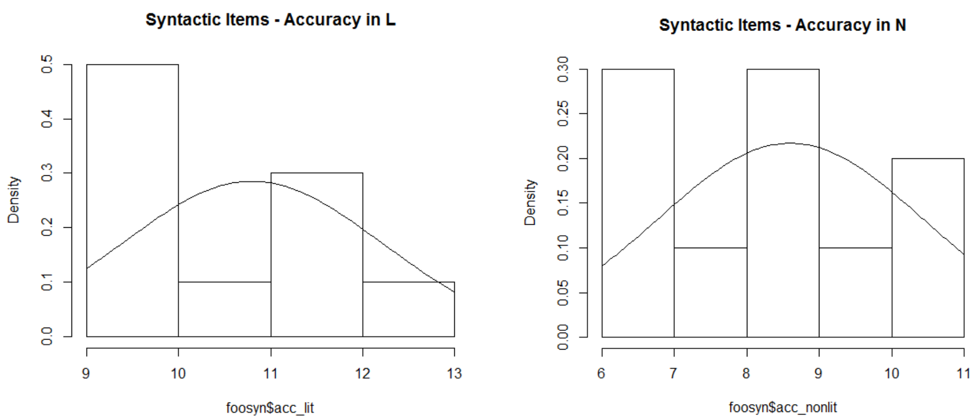


Figure 42. Histograms of recall accuracy for syntactic items in the literal and non-literal translation conditions respectively.

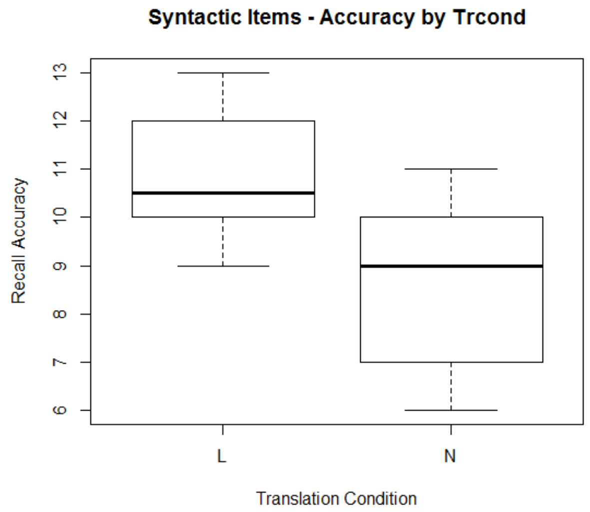


Figure 43. Boxplots of recall accuracy for syntactic items in the literal and non-literal translation conditions respectively.

The visual inspection of the plots (histograms and Q-Q plots in fig. 41-42), as well as skewness and kurtosis tests, suggest that syntactic data follow a roughly normal distribution. Despite relatively high levels of kurtosis (-1.64 in L and -1.52 in N), the Shapiro tests of normality suggest that the syntactic recall data nevertheless follows a roughly normal distribution in both L (W = 0.90, p = 0.26) and N (W = 0.91, p = 0.33). The homoscedasticity assumption is met (Levene's test: F = 0.21, p > 0.05) so the variances can be considered equal in both groups. Fig. 43 shows that the two medians are different and the accuracy score range is higher for literal items. The t-test found a significant difference (95% CI 0.65, 3.74; t = 3.01, df = 17, p = 0.007) in accuracy between syntactic items translated literally (M = 10.8, SD = 1.39) and non-literally (M = 8.6, SD= 1.83). Effect size is large (Cohen's δ = 1.34) and the test has a satisfactory level of power (0.80).

From the linguistic category analysis above, it emerges that accuracy is not statistically different in literal and non-literal lexical items whereas it is for syntactic ones. Translation condition therefore seems to play a significant role and have a statistical effect on the recall of syntactic structures. There are reasons to believe that this may also be true of individual lexical words, given the CI range and the fact that the t-test approaches significance. However, as we have seen above, such test does not have enough power to clarify the relationship between lexicon and translation condition. Moreover, the non-normality of one of the two groups (lexical literal items) means that its results are not fully reliable. More data on lexical items would have to be gathered to shed light on this point. Finally, the interaction of these two variables (translation condition and linguistic category) was analysed by means of t-tests. However, carrying out several t-tests inflates the risk of Type I error (Gravetter and Wallnau, 2013: 391), i.e. the chances of incorrectly rejecting a true null hypothesis (considering an effect significant when in reality it is not). When testing for mean differences, the probability of type I error is usually 5%, and there is a formula for determining the new error rate for multiple t-tests. By running three t-tests on the same data as it was done here, the probability would be 14.3%. This error rate is unacceptable, as it does not allow enough confidence that any significant result found are not due to chance alone. Because the data in question consists of three categorical variables, a possible alternative test could be log-linear analysis. However, this type of analysis can only be used in between-subjects designs and, importantly, it demonstrates association between variables without making any clear-cut distinction between predictor and outcome (Ebrahim, 1999). If the variables are explicitly independent and dependent, logistic regression should be used instead (Jeansonne, 2002). The results of the relevant logistic regression are therefore presented in the next section.

4.3.5.3 Evaluating the Effects of Linguistic Category and Translation on Recall: Logistic Regression

4.3.5.3.1 Analysis

Binomial logistic regression was used to ascertain whether recall scores could be predicted by translation condition and linguistic category[27]. The effects of these two categorical predictors on the likelihood that participant correctly recall subtitle wordings were therefore analysed. The outcome variable was coded as accurate (Yes = 1) and non-accurate answer (No = 0) and the two predictors were translation condition (x_1= trcond variable) and linguistic category (x_2 = cat variable). The cat distribution was 312 (55% of the total) lexical items and 260 (45%) syntactic ones in total. The complete breakdown of observed values for all these variables are reported in table 18. The trcond predictor was coded as 1 = literal and 0 = non-literal, whereas the cat predictor was coded 1 = syntactic and 0 = lexical. The analysis was carried out using a generalised linear model (glm) in the programme R (R Core Team, 2015) version 3.2.3 in the Windows 2000 environment. The logistic regression equation is as follows:

---

[27]Henceforth, abbreviated to 'trcond' and 'cat' respectively.

$$logit(Y) = natural\log(odds) = ln\frac{\pi}{1-\pi} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Where the logit is the log of the odds, ln is the natural logarithm, π is the probability of an event occurring over the probability of the event not occurring $(1 - \pi)$, α is the coefficient for the Y axis (intercept) and βs are the coefficients for the relevant predictors X_s. Because the dependent variable is dichotomous, logistic regression predicts *the probability* of outcome Y occurring given known values of X_s, rather than predicting *the value* of Y, such as in OLR (ordinary linear regression) with continuous dependent variables. The natural log transformation is necessary to make the relationship between said categorical outcome and its predictor(s) linear.

The simplest logistic regression (model 1, translation condition only) was run first and was significant, as per the likelihood ratio test $\chi 2(1) = 18.71$, $p < 0.001$. The coefficients and odds ratios for this model are presented in Table 20. Translation condition was a significant predictor (β = 0.82, z = 4.25, p < 0.0001). The coefficient β represents the change in the logit of the outcome variable associated with a unit change in the predictor. Negative coefficients represent a negative relationship between the probability of success (i.e. providing an accurate answer) and the independent variable, whereas positive values represent a positive relationship. In model 1, given the baseline level (trcond N = 0), the probability of success increases by going from non-literal to literal condition. In other words, there is a positive relationship between literally translated items and accurate recall. With categorical predictors, the coefficients *are* the log-odds ratios. Therefore, they can be converted manually to odds ratios by exponentiation ($e^{0.59} = 1.80$, $e^{0.8228} = 2.2768$). Odds ratios are crucial to the interpretation of logistic regression. They are an indication of how much the odds increase multiplicatively with a one-unit change in the independent variable. Odd ratios are always positive values and the distinction between a positive and negative relationship depends on which side of 1 they fall on. If an odds ratio is larger than 1, the probability of the outcome happening increases as the predictor increases (positive relationship). Here an odds ratio of 1 would mean that the odds of accurate and inaccurate response are the same. As it can be seen in table 20, in both cases (intercept and translation condition) the odd ratios are larger than one. In fact, the odds of a literally-translated item being recalled correctly were 2.27 times higher than those of non-literally translated items. Therefore, literal translations were 2.27 times more likely to be correctly recalled than divergent translations. Confidence intervals were also produced for said odd ratios.

If a confidence interval contains 1, one cannot be sure that the predictor leads to a higher or lower probability of an accurate response. Notice how in table 20 below the CIs do not cross 1.

Predictor	*β* Estimate	SE	Z value	p-value	Odds ratios	CI	
						2.5%	97.5%
Intercept	0.5900	0.1234	4.779	p < 0.0001***	1.803922	1.419800	2.304863
Translation Condition (0 = N, 1 = L)	0.8228	0.1935	4.252	p < 0.0001***	2.276786	1.563617	3.341941

Table 20. Summary table for the logistic regression model 1 (accuracy ~ translation condition).

The model explained 4.6% of the variance in accuracy scores (Nagelkerke's $R^2 = 0.046$). The fact that the pseudo R^2 is quite low is understandable as this is a one-predictor-only model and evidently there are other variables which contribute to explaining the variance in recall. Unfortunately, there are many pseudo R^2 measures available (see Menard, 2000), and therefore R^2 in logistic regression does not have one clear univocal definition (Bartlett, 2014). Moreover, there is no consensus on which pseudo R^2 is best (Allison, 2013). These pseudo measures were devised to provide a measure of proportion of variance explained, but they cannot be deemed equivalent to R^2 in OLR (Long, 1997: 104-109), so several researchers warn that they should be used and interpreted with caution. As Hosmer and Lemeshow report in their volume on logistic regression: "[…] low R^2 values in logistic regression are the norm and this presents a problem when reporting their values to an audience accustomed to seeing linear regression values. […] Thus we do not recommend routine publishing of R^2 values with results from fitted logistic models." (2000: 167). According to Peng et al., R^2 should therefore be used as supplementary to other, more useful evaluation indices (2002: 6). One of these is the Akaike Information Criterion (AIC). The AIC exists to solve another problem with pseudo R^2, namely that the latter becomes larger every time a variable is added (Field et al., 2012: 318). A logistic model provides a better fit to the data if it constitutes an improvement over the intercept-only model (null model), which is considered a good reference because it has no predictors (Peng et al., 2002: 5-6). This was the case for model 1, whose AIC value (659.51) was smaller than that of the null model (676.22), indicating that model 1 provides a better fit to the data. The deviance statistic was also considered. This is another way of assessing the fit of the model by comparing it to a baseline state using the likelihood ratio (Field et al., 2012: 316). This measure is minus two times the log-likelihood and is often referred to as -2LL. Larger values indicate models that provide a poorer fit. Here, the null model deviance is 674.22 and model 1 deviance 655.51, which confirms the improvement from the former to the latter. Note that the difference in deviances (674.22 - 655.51 = 18.71) corresponds to the model chi-square.

Based on the results of the t-tests in the previous sections, linguistic category could play a role on memory accuracy and was therefore added as a predictor together with its interaction with translation condition (model 2). The addition of these variables did not significantly improve model fit, as demonstrated by the likelihood ratio test, $\chi 2(2) = 1.42$, p > 0.05.

Predictor	***β* Estimate**	**SE**	**Z value**	**p-value**	**Odds ratios**	**CI**	
						2.5%	**97.5%**
Intercept	0.5245	0.1657	3.166	p < 0.002***	1.689655	1.2258269	2.350156
Translation Condition (0 = N, 1 = L)	0.7531	0.2551	2.953	p < 0.004***	2.123649	1.2945299	3.525839
Linguistic Category (0 = lex, 1 = syn)	0.1456	0.2486	0.586	p > 0.05	1.156772	0.7114659	1.888208
Cat*Trcond	0.1678	0.3926	0.427	p > 0.05	1.182694	0.5494237	2.567334

Table 21. Summary table for the logistic regression model 2 (accuracy ~ translation condition + linguistic category).

As one can see from table 21, the coefficient for translation condition has decreased slightly (from 0.82 to 0.75) but remains highly significant, while the coefficients for linguistic category and the interaction term are not (p > 0.05 in both cases). The odd ratios are all larger than one, indicating positive relationships. As far as linguistic category is concerned, the analysis shows that the probability of providing an accurate answer in the post-test is higher with syntactic items. Specifically, syntactic items are 1.15 times more likely to result in an accurate answer compared to lexical items. However, this odds ratio is not reliable as its confidence interval is narrow and contains 1, which means that there cannot be certainty that linguistic category leads to a higher or lower probability of an accurate response. The same applies to the interaction term, whose CI is also quite narrow and contains 1. The odd ratios thus confirm that the addition of these two predictors does not significantly explain variance in accuracy scores. The deviance statistic for model 1 is 655.51 and for model 2 is 654.09, so model 2 would seem to provide a slight improvement over model 1. However, this difference is quite small and the β coefficients for linguistic category and the interaction term are not significant.

Moreover, the AIC value for model 2 (662.09) is larger than that of model 1 (659.51), again suggesting that model 1 provides an overall better fit to the data. Logistic regression results support the proposition that the simplest model (translation condition only) remains the best at predicting the probability of the outcome Y (recall accuracy).

4.3.5.3.2 Diagnostics

After the analysis, casewise statistics were calculated to see how well model 1 fits the observed data. This is required because logistic regression will generate a model regardless of the observed data, yet the real-life value of the model may be hampered if its predictions do not fit the collected observations well. In logistic regression, fitted values are the calculated probabilities of outcome Y occurring given the values of predictor X for a given subject (Field et al., 2012: 338). The predicted probabilities created by the model can be contrasted with the observed outcome to discover whether higher probabilities indeed correspond to an event occurring (in our case accurate recall). The analysis of predicted probabilities for model 1 revealed that when an item is translated literally, the probability of it being remembered correctly is 0.804, whereas when an item is non-literal, the probability is 0.643. That is to say, roughly 80% of subjects recall correctly literal items, whereas around 64% recall correctly divergent ones. Since a probability of 0 means no chance of correct recall and a probability of 1 means that the subject will definitely recall an item accurately, it becomes clear that both literal and non-literal items are recalled quite accurately overall, although literal ones are associated with a significantly higher probability of recall success. Since Model 1 has only one categorical predictor – translation condition – with only two levels (dichotomous variable), only two fitted probabilities were returned. For reference, the predicted probabilities for model 2 are also presented (table 22) and reveal that when a lexical item is translated literally, the probability of it being remembered correctly is 78%, whereas when translated non-literally the probability is 63%. When a syntactic item is translated literally the probability of correct recall is 83% and when translated non-literally 66%. In both translation conditions, syntactic items have a higher probability of being remembered correctly, thus reiterating the (albeit non statistical) finding in the coefficients table for model 2 and supporting the t-test results addressed in the above sections.

Translation Condition	Linguistic Category	Accuracy	Fitted Values
L	lex	Y	0.7820513
L	lex	N	0.7820513
L	syn	Y	0.8307692
L	syn	N	0.8307692
N	lex	Y	0.6282051
N	lex	N	0.6282051
N	syn	Y	0.6615385
N	syn	N	0.6615385

Table 22. Predicted probabilities for logistic regression model 2.

The residuals were also examined. They are the difference between the data points collected in the sample and those predicted by the model, and, as such, they represent the error. Standardised residuals are the residuals divided by an estimate of their standard deviation. Studentised residuals are the difference between the adjusted predicted value and the observed value divided by the standard error, and are called studentised because they follow a Student's t-distribution. Both types of residuals allow to compare different models, and studentised residuals are also useful to judge the influence of a particular case on the ability of the model to correctly predict that case (Field et al., 2012), i.e. to identify influential cases. Only 5% of the standardised residuals should have values outside ±1.96 (rounded to 2), and only around 1% should be outside ± 2.58. Cases close to 3 require inspection and cases above 3 (or below -3) constitute a cause for concern. Model 1 residuals are slightly unusual in the sense that there is not much variability in their values. The reason for this is that they are based on a single predictor that is categorical (Field et al., 2012). Residual data was inspected and no cases outside ±1.96 were found (see R code below).

```
> step7reg$stdz.resid<-rstandard(model1)
> step7reg$stu.resid<-rstudent(model1)
> step7reg$stdz.resid > 2 | step7reg$stdz.resid < -2
# stored as TRUE or FALSE
> step7reg$large.stdz.resid<-step7reg$stdz.resid > 2 |
step7reg$stdz.resid < -2
# count large standardized residuals (no. of TRUEs):
> sum(step7reg$large.stdz.resid)
> [1] 0
```

Other residual and influence statistics are leverage, Cook's distance and DFBeta. Leverage assesses the influence that an observed value has on the outcome Y over the predicted values (Field et al., 2012: 269). Leverage values are also called hat values. They can range between 0 (no influence) and 1 (complete influence). If no case has an undue influence on the outcome, all values should be closed to the average value (k + 1)/N, where k is the number of predictors and N is the sample size (in our case 1+1/572 = 0.003496503). Influential cases will have values twice or three times the average leverage (i.e. 0.006 or 0.010 respectively). Leverage inspection revealed no hat value was larger than 0.007 (see R output below), confirming the lack of influential cases.

```
> step7reg$leverage<-hatvalues(model1)
# average leverage = (k+1)/n
> (1+1)/572
[1] 0.003496503 # average hat value
# influential cases: 2(k+1)/n or 3(k+1)/n
> 2* 0.003496503 - [1] 0.006993006
> 3* 0.003496503 - [1] 0.01048951
> step7reg$leverage >= 0.007
> step7reg$large.leverage<-step7reg$leverage >= 0.007
> sum(step7reg$large.leverage)
[1] 0 # no value above 0.007
> step7reg$cooks.dist<-cooks.distance(model1)
> step7reg$cooks.dist > 1
# no case > 1
> step7reg$dfbeta<-dfbeta(model1)
> min(step7reg$dfbeta): [1] -0.01596825
> max(step7reg$dfbeta): [1] 0.01051866
```

Cook's distance is another measure of influence and gauges the effect of each case on the whole model. Values greater than 1 should alert the researcher (Cook and Weisberg, 1982). In model 1 no value exceeds 1 so there is no cause for concern. DFBeta measures the influence of a specific case on the β coefficient of a regression model. Values larger than 1 indicate an influential case. DFBetas do not exceed 1 in model1.

4.3.5.3.3 Assumptions

A number of premises need to be met for logistic regression results to be valid. Firstly, there should be a linear relationship between any continuous predictor and the logit of the outcome variable. Because in this analysis there were no continuous predictors, there is no risk of infringing the linearity assumption. Secondly, errors should be independent, meaning that, for any one pair of observations, the residuals should not be correlated. This assumption can be checked through the Durbin-Watson test. With a test statistic of 1.841 and a related p-value of 0.064, the null hypothesis that the errors are correlated can be rejected. Residuals are considered independent and the validity of model 1 is therefore further confirmed. Thirdly, predictors should not be too highly correlated. In this analysis, model 1 only had one predictor so this assumption does not apply. In model 2, multicollinearity can be tested using VIF (Variable Inflation Factor). Although in some cases recommended maximum VIF values of 5 (Rogerson, 2001) and 4 (Pan and Jackson, 2008) are present in the literature, the maximum VIF has been most commonly set at 10 (Hair et al., 1995; Myers, 1990; Neter, Wasserman and Kutner, 1989, amongst others). In model 2, both trcond and cat have very low values compared to these recommendations (1.73 and 1.67 respectively, see R output below), so it is possible to conclude that the predictors are not correlated.

```
> install.packages("car")
> library(car)
> durbinWatsonTest(model1)
lag Autocorrelation D-W Statistic p-value
1      0.07639611      1.841568   0.064
Alternative hypothesis: rho != 0
> vif(model3)
trcond        cat   trcond:cat
1.732679   1.671623   2.350594
```

4.3.5.3.4 Conclusions

Logistic regression is a powerful tool but requires a number of aspects to be assessed in the analysis, namely (1) overall evaluation of the model (with significance testing to compare it against the intercept-only or a previously run model), (2) significance testing for each predictor's coefficient, (3) goodness-of-fit measures, (4) diagnostics and (5) assumptions verification. In section 4.3.5.3, a logistic regression was run on translation condition and linguistic category data (model 2). Coefficients were analysed to see what predictors contribute substantially to the model. It was found that, despite the previous t-test results, linguistic category and its interaction with translation condition are not significant contributors to the model's ability to predict recall accuracy. Model 2 did not constitute an improvement over model 1 (translation only), as highlighted by the likelihood ratio test, $\chi^2(2) = 1.42$, $p > 0.05$. The only significant predictor was translation condition. The simplest model proved the best ($\chi^2(1) = 18.71$, $p < 0.001$) at predicting the probability of the outcome Y and was therefore retained. Estimate coefficients and related p-values were produced and used to define this regression model. Odds ratios and their confidence intervals were also analysed, and showed that literal items are 2.27 times more likely to be remembered correctly than non-literal items. Other measures such as pseudo R^2, AIC and the deviance statistic (-2LL) were then used to assess model fit. Moreover, several diagnostic tools were employed to assess the quality of the model. Predicted values, standardised and studentised residuals, Cook's distance, leverage and DFBeta values all confirmed that the error in the model is largely within acceptable limits and there are no cases of undue influence. Finally, a number of assumptions were to be checked for the regression results to be valid. The Durbin-Watson test and the Variable Inflation Factor were computed and confirmed all pre-requisites were met, ensuring model validity. The multiple evidences herein summarised suggest that linguistic category does not statistically contribute to the probability of accuracy success, whereas translation condition is confirmed as a significant predictor of recall accuracy.

4.3.6 Order of Presentation

4.3.6.1 Subtitle Duration by Half

During pilot analysis, it emerged that subtitle duration differed between the first and second part of the experimental stimulus. Subtitle durations for the 22 subtitles that were chosen as test items were analysed. A number of assumptions were checked to assess whether a t-test is the most appropriate test for group differences. The equality of variance assumption can be tested via the Levene's test, which confirms that the data are homoscedastic ($F = 1.45$, $p = 0.2$). Subtitle duration data for the first half of the video is slightly skewed (skewness = 0.8) and for the second half is leptokurtic (kurtosis = 1.4). However, the scatterplots, histograms and Q-Q plot examined revealed roughly normal patterns and the Shapiro-Wilk tests confirmed normality (for durations in the first half: $W = 0.88$, $p > 0.05$; in the second half: $W = 0.94$, $p > 0.5$). No strong evidence of non-normality for either of the two groups was found. On average, subtitle duration for the 11 test items in the in the second half ($M = 63.72$, $SD = 16.87$) was roughly one second (24.7 frames) longer than for those in the first half ($M = 39$, $SD = 12.51$). This difference is represented in fig. 44. The independent two-sample t-test revealed a significant difference (95% CI -38.01, -11.44; $t = -3.09$, $df = 18$, $p < 0.001$) in subtitle duration between the first and second part of the video, as far as the 22 test items are concerned. The test has a large effect size (Cohen's $\delta = 1.6$) and good power (94%).

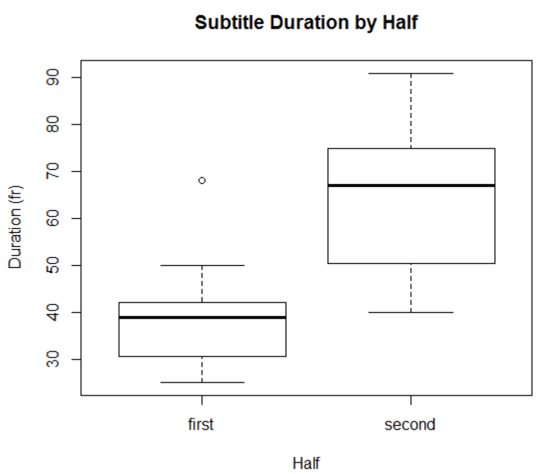


Figure 44. Boxplots of subtitle durations in the first and second half of the video stimulus (22 subtitles).

The analysis above was then repeated for the whole experimental stimulus. On average, subtitle duration was only 7 frames longer in the second (M = 58.19, SD = 27.08) than in the first half of the video clip (M = 50.55, SD = 23.35). Throughout the video clip, this difference is much less marked than in the test items alone, as it is visible by comparing the boxplots in fig. 44 and 45.

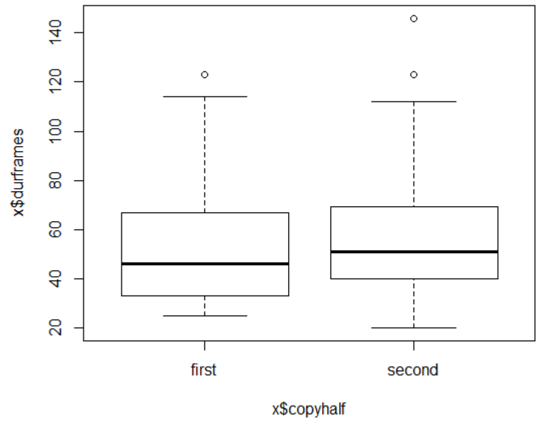


Figure 45. Boxplots of subtitle duration in the first and second half of the video stimulus (110 subtitles).

The variances of the two populations from which the samples are taken are equal as per the Levene's test (F = 0.91, p = 0.3). During the visual inspection of the plotted data, the histograms showed a clear positive skew and the Q-Q plots further confirmed a departure from normality. Skewness and kurtosis tests support the findings, in particular for durations in the first half of the video (skewness = 1.22; kurtosis = 1.18). The Shapiro-Wilk normality tests are also significant for both halves (first half: W = 0.87, p < 0.0005; second half: W = 0.91, p < 0.0005), confirming that the data cannot be considered normally distributed. For this reason, the t-test is not the best-suited way of assessing group differences and a non-parametric alternative should be used. The Wilcoxon rank-sum test (W = 1236, n = 110, p > 0.05) confirmed the result of the t-test, providing no evidence to support the notion that subtitle duration changes significantly between the first and second part of the experimental stimulus as a whole. Therefore, although the 22 test subtitles differ statistically in duration between halves, overall duration of the subtitles is evenly spread throughout the clip.

4.3.6.2 Subtitle Length by Half

As expected, a very similar situation to subtitle duration arose during pilot analysis for subtitle length. In the main study, this variable also significantly changed between halves, as far as the 22 test items were concerned. The 11 subtitles in the first half of the clip had on average 3.6 words, against 5.3 words in the second half. This difference is represented visually in fig. 46. Subtitle length data were found to deviate from normality in the plots and numerical tests, including the Shapiro-Wilk test ($W = 0.92$, $p < 0.05$). Therefore, the Wilcoxon rank-sum test was used to assess the mean difference between groups. The test confirmed that there is a statistical difference between test items in the two halves ($W = 132$, $n = 22$, $p < 0.05$).

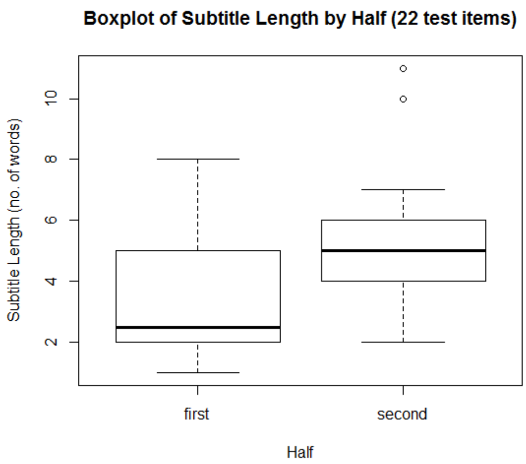


Figure 46. Boxplots of subtitle length in the first and second half of the video stimulus (22 subtitles).

The analysis above was repeated for the whole experimental stimulus (110 subtitles). On average, subtitle length was very similar, being only 0.2 words longer in the second ($M = 4.08$, $SD = 2.62$) than in the first half of the video clip ($M = 5.00$, $SD = 2.56$). Subtitle length data for the whole stimulus material do not follow a normal distribution according to graphical and numerical descriptives including the Shapiro-Wilk test ($W = 0.94$, $p < 0.0005$), so the mean difference was tested through the Wilcoxon Rank-Sum test, which found no significant difference between the groups ($W = 5627.5$, $n = 110$, $p > 0.05$). Therefore, although there is a statistical difference in number of words in the 22 test items between halves, the length of the subtitles is also overall evenly spread throughout the clip.

4.3.6.3 Fixation Measures by Half

Since subtitle measures (duration and length) correlate with fixation measures (duration and count), the latter are also likely to vary according to stimulus presentation as far as the 22 test items are concerned. The relevant data subsets were analysed for the two fixation metrics. Both totals and averages were investigated.

As far as fixation count is concerned, the subtitle data in each half ($n = 11$) were found to follow a roughly normal distribution for both totals and averages. Therefore, t-tests were computed. The results confirm that both *total* fixation count (95% CI -96, -23.44; $t = -3.43$, $df = 20$, $p < 0.005$) and *average* fixation count (95% CI -3.69, -0.90; $t = -3.43$, $df = 20$, $p < 0.005$) differed significantly between halves. The negative sign of the t-statistic indicates that participants looked more times in the second half of the clip. Indeed, the mean number of fixations on all 11 items appearing in the first half of the clip was 96, in the second 155. On average subjects made 4 fixations on each subtitle in the first half and 6 on each subtitle in the second half. Cohen's δ was very similar (1.46 for *total* fixation count and 1.44 for *average* fixation count), indicating a large effect. The power of these tests was 87% and 90% respectively.

As far as fixation duration is concerned, data were also found to be roughly normally distributed. The t-test results for *total* fixation duration (95% CI -19728.62, -4708.10; $t= -3.39$, $df = 20$, $p < 0.005$) reveal that the total duration of all fixations registered across participants on the 11 test items appearing in the first half of the clip (M = 21444ms, i.e. 21s) was significantly shorter than the total fixation duration registered on items in the second half (M = 33662ms, i.e. 34s). The power of this test is 89%, and Cohen's δ is 1.44, indicating a large effect. On the other hand, the t-test results for *average* fixation duration (95% CI -9.18, 27.16; $t = 1.04$, $df = 16$, $p > 0.05$) did not indicate any significant difference between the two parts of the video. In fact, the positive sign of the t-statistic indicates that the average fixation duration made by participants on each of the 11 subtitles appearing in the first half (M = 227ms) was longer than the average fixation duration on those in the second half (M = 218ms). Interestingly, when fixation duration is analysed *per se* – i.e. in isolation from fixation count – and is simply added up to form totals, there seems to be a significant difference in duration between halves, but when the measure is averaged, said difference disappears. This is to be expected if one considers that, as it was established in sections 4.3.6.1 and 4.3.6.2, the 22 test subtitles are longer (subtitle length) and stay on screen for longer (subtitle duration) in the second half, thus attracting more looks. Yet, the second half of the video does not attract more looks for any other reason than the fact the subtitles are longer. In other words, if fixation duration is not considered in isolation but in relation to its corresponding number of fixations (average), then this measure is no longer affected by order of presentation.

Just like for subtitle duration and length, the fixation analysis above was repeated for the whole experimental stimulus (n = 110). Only data for fixation duration is presented hereafter, as the process yielded equivalent results for fixation count. First, total fixation duration is considered – that is, the raw sum of all fixation durations on all subtitles by all participants; then average fixation duration is considered – that is, the sum of all fixation durations by all participants on each subtitle divided by the number of fixations made by all participants on that subtitle.

Total fixation duration data have homogenous variances (Levene's $F = 0.2$, $p > 0.05$) but do not follow a normal distribution in any of the two groups (first and second half). Therefore, the Wilcoxon rank-sum test was used in place of the t-test, as the former does not assume normality of the data. The sum of all fixation durations (across participants) is around 28 minutes for the second half and around 25 min in the first. This overall difference between halves is not very large, and was found to be non significant by the Wilcoxon rank-sum test ($W = 1364$, $n = 110$, $p > 0.05$). It appears that total fixation duration does not change significantly between the two parts of the stimulus.

Average fixation duration was then analysed. These data have homogenous variances (Levene's $F = 1.79$, $p > 0.05$) but graphical and numerical description evidenced an acute departure from normality in both groups (first and second half). Therefore, the Wilcoxon rank-sum test was used to asses mean group differences. Average fixation duration is around 226ms in the second half and around 220ms in the first half. Therefore, there is only a 6ms fixation duration difference between halves, which was found to be non-significant by the Wilcoxon rank-sum test ($W = 1595$, $n = 110$, $p > 0.05$). It appears that neither total nor average fixation duration change significantly between the two parts of the video. However, to make sure that the differences detected between halves do not influence experimental post-test results, the effect of order of presentation on recall was appraised.

4.3.6.4 Recall Accuracy by Half

On average, subtitles appearing in the second half of the clip were recalled more accurately than those in the first (77% versus 69% accuracy respectively). Fig. 47 shows the variation in range. The item with the lowest accuracy rate in the first part of the video

(AOI.5) was recalled correctly by 9 subjects out of 26, whereas in the second part the lowest-accuracy item (AOI.89) was nevertheless correctly identified by 16 out of 26 subjects. That is to say, subtitles in the first half had a wider range of accuracies compared to the second, where subjects generally achieved higher performance. However, the wider range of values in the first half of the clip (see fig. 47) is largely due to the presence of AOI.5. If AOI.5 were excluded, the minimum recall for items in the first half would be 15, i.e. very similar to the minimum recall score for items in the second half. Moreover, the graph shows that the medians of the two groups are fairly similar, which would suggest this group difference to be due to chance. Indeed, the difference in recall accuracy between halves was not significant according to the independent two-sample t-test (95% CI -5.27, 0.54; t = -1.71, df = 16, p > 0.05).

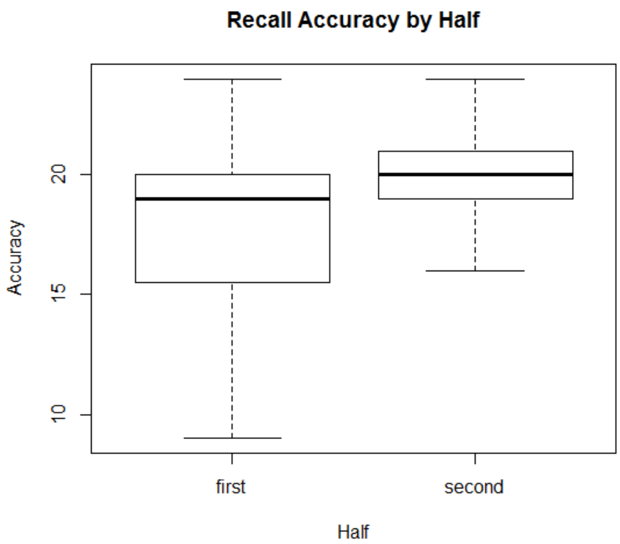


Figure 47. Boxplots of recall accuracy in the first and second half of the video stimulus (22 subtitles).

The required t-test assumptions were also tested, and no clear evidence of non-normality was detected in graphical and numerical tests. Recall data for the second half of the video is mildly leptokurtic (kurtosis = -1.02), the other skewness and kurtosis levels are negligible. The Levene's test confirmed the homoscedasticity assumption (F = 1.04, p = 0.3). Recall accuracy follows a normal probability distribution and the samples have equal variances, thus confirming that the t-test was the most appropriate test to use and its results are reliable.

4.3.7 Relative Frequency

4.3.7.1 Frequency and Translation Condition

Throughout this section, AOI.60 will be excluded from the analyses since, as explained in section 3.14 of the methodology chapter, the item is not lexicalised. This means that the only frequencies that could be extracted for this item were those for the word pair *ma – però [but]*, which is not representative of the whole subtitle. Therefore, the total subtitles to be analysed hereafter are 42 overall (21 wordings per translation condition) rather than 44. In 9 cases out of 21, the subtitles had a higher relative frequency in the non-literal condition. The remainder and majority of the test items (12 cases on 21) had a higher relative frequency in the literal condition. While this difference is not very large, it may introduce a confound if one cannot safely distinguish between a translation condition effect and a frequency effect. That is to say, if there is a significant difference in frequency between literal and non-literal items, for example if literal items are consistently more frequent than non-literal ones, one cannot be sure of whether the differential memory effect found is attributable to the experimental manipulation (the literal and non-literal wordings) or just item frequency. Therefore, the relationship between these two variables was also considered.

The variable frequency is not normally distributed, as it is visible from the scatterplot and Q-Q plot below (fig. 48). The former revealed how the data is concentrated in the lower part of the graph, indicating that most subtitles have relative frequencies between 0 and 150 IPM. The Q-Q plot also shows a different shape from that of a Gaussian distribution, since the data points do not follow the reference line. The levels of skewness (2.59) and kurtosis (8.02) are very high. As the boxplot in fig. 49 also shows, the data is heavily skewed in both groups. There is also an extreme point (IPM 333.07, AOI. 94) at the top of the graph. However, there is no valid reason to remove this point from the analysis. Moreover, replotting the data without AOI.94 (see fig. 50) shows that the data remain heavily left skewed in both groups. The median in fig. 49 is similar but slightly lower in N. The size of the box is larger and the whisker is taller in L, indicating a wider data range in this condition. The lower whiskers are almost non-existent, especially in N, meaning that the lowest value is part of the most frequent values (the box within which 50% of the data falls). Due to a violation of the normality assumption in both translation conditions, the non-parametric alternative to the independent t-test, the Mann-Whitney (or Wilcoxon rank-sum) test, was used in this analysis. Results showed that relative frequency for subtitles translated literally (M = 44.3, SD = 76.45) does not change significantly (95% CI: -4.271, 16.912; W = 259; p > 0.05) from that of subtitles translated non-literally (M = 33.08, SD = 51.78).

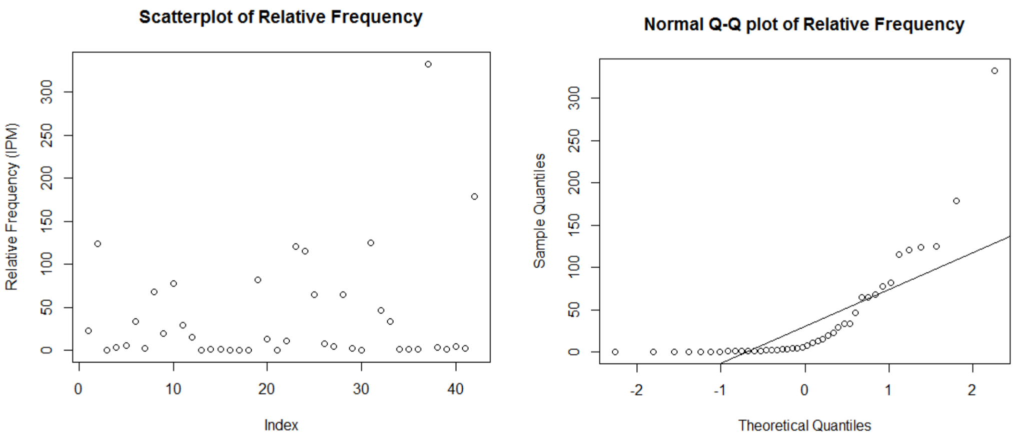
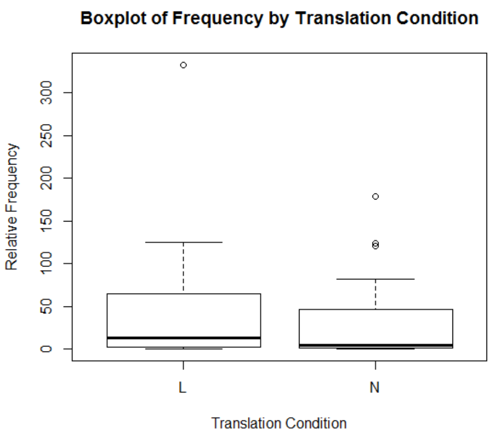


Figure 48. Scatterplot and Q-Q plot of relative frequency (22 subtitles).



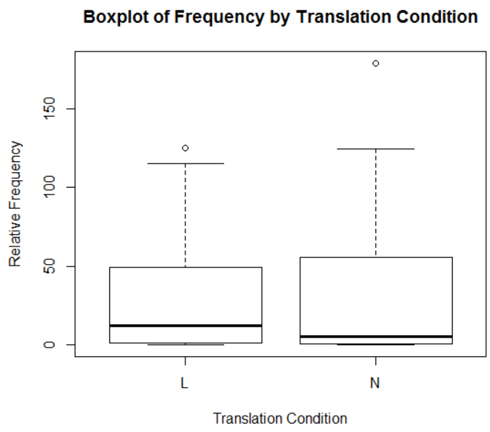


Figure 49. Boxplots of relative frequency by translation condition (includes extreme data point). Here the full spectrum of data is plotted, including AOI.94 (333.07 IPM).

Figure 50. Boxplots of relative frequency by translation condition (excludes extreme data point). Here AOI.94 has been excluded from the data, so the maximum y-axis value is 200 IPM.

4.3.7.2 Frequency and Recall

The aim of this analysis is to compare accurate scores by frequency, so recall accuracy was aggregated into a numerical variable to gather information on how many subjects out of the total correctly recalled a subtitle in each of the two frequencies. The accuracy figures for each item are therefore aggregated over 26 subjects, 13 in each group. In the previous section, relative frequency was found to violate the normality assumption (see fig. 48). Accuracy data were also found not to follow a normal distribution in both graphical and numerical examinations, including the Shapiro-Wilk normality test (W =

0.89, $p < 0.005$). From the scatterplot in fig. 51, no strong relationship between accuracy and relative frequency emerges. The correlation line is positive but almost flat and the graph shows that some highly frequent items (for example AOI.99, 179 IPM) are recalled by relatively few subjects (7 people out of 13), whereas other much less common items (with a relative frequency between 0-50 IPM) are recalled more accurately (by 12 subjects out of 13). The graph suggests no significant relationship between the variables, which is confirmed by the Spearman's non-parametric correlation test ($r_s = 0.09$, $n = 42$, $p > 0.05$).

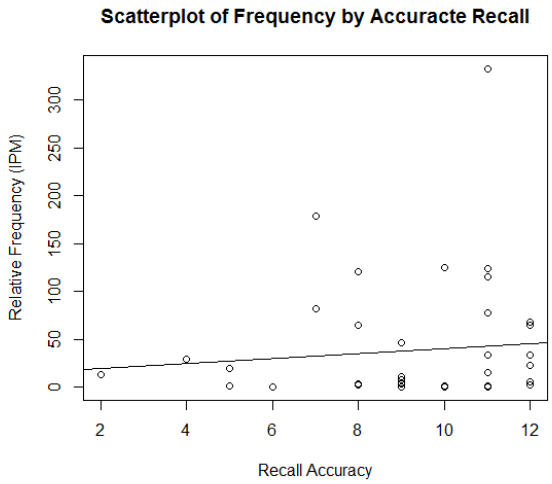


Figure 51. Scatterplot of relative frequency by recall accuracy.

4.3.7.3 Frequency and Linguistic Category

Descriptive statistics reveal that relative frequency violates the normality assumption in both the lexical and syntactic category. Fig. 52 shows very skewed data in both groups, as well as the presence of extreme data points. The median is similar in the two groups but slightly lower for syntactic items. In lexical items, the size of the box is slightly larger and the whisker is taller, indicating a wider data range in this condition. The lower whiskers are almost non-existent, especially in syntactic items, which indicates left skew in the data. Due to the violation of the normality assumption, the non-parametric Mann-Whitney test was computed and showed that although relative frequency for lexical subtitles ($M = 46.33$, $SD = 73.61$) tends to be higher than for syntactic ones ($M = 28.50$, $SD = 50.86$), this difference is not significant (95% CI: -0.672, 26.673; $W = 282$; $p > 0.05$).

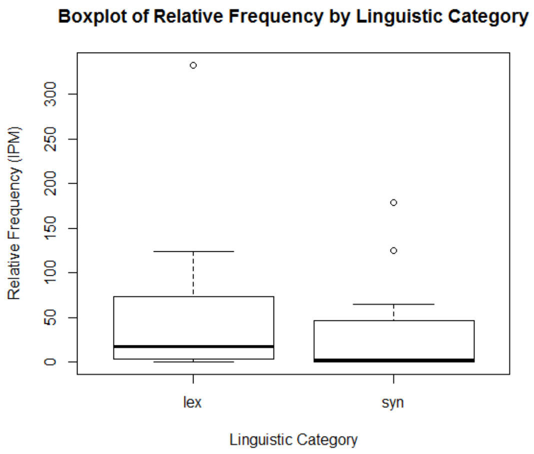


Figure 52. Boxplots of relative frequency by linguistic category.

4.3.7.4 Frequency and Order of Presentation

Frequency was analysed in relation to order of presentation of the experimental stimulus as well. As in the previous sections, frequency is not normally distributed and heavily skewed in both first and second half of the video. Fig. 53 highlights that subtitles appearing in the second half have a wider range of frequencies, the medians are similar and data in both groups is visibly left-skewed. The Mann-Whitney test showed that relative frequency for subtitles appearing in the first half (M = 27.32, SD = 38.90) does not change significantly (95% CI: -29.605, 7.357; W = 178; $p > 0.05$) from that of subtitles appearing in the second half of the clip (M = 51.19, SD = 84.01).

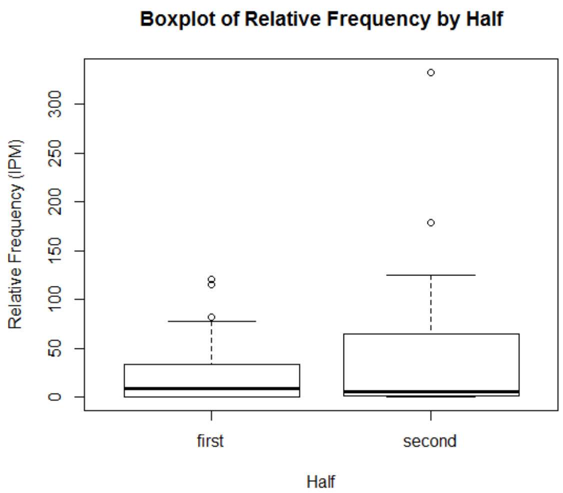


Figure 53. Boxplots of relative frequency by order of presentation.

4.3.7.5 Frequency and Fixation Measures

4.3.7.5.1 Frequency and Fixation Count

Fixation count data was found to be normally distributed, yet relative frequency is known to depart from normality and be heavily skewed. For this reason, the Spearman's rank-order correlation was computed. A visual feel for the relationship of these two variables can be obtained from fig. 54. Perhaps surprisingly, this scatterplot of fixation count and relative frequency reveals a positive but tenuous correlation between the variables. Results show no significant relationship ($r_s = 0.21$, $n = 42$, $p > 0.05$) between total fixation count and relative frequency of the subtitles.

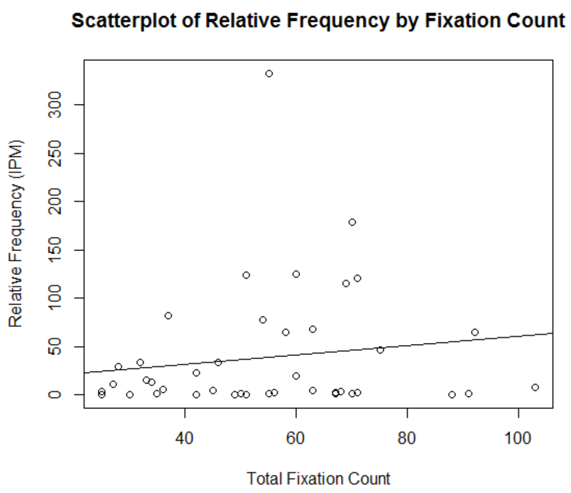


Figure 54. Scatterplot of relative frequency by fixation count.

4.2.7.5.2 Frequency and Fixation Duration

The same analysis for fixation count was carried out for fixation duration, both for totals and averages. Total fixation duration data were found to be normally distributed, yet since frequency departs from normality the Spearman's rank-order correlation was computed. In fig. 55, frequency and total fixation duration are plotted together. Just like for fixation count, this scatterplot reveals a positive but tenuous correlation between the variables. Test results show no significant relationship ($r_s = 0.20$, $n = 42$, $p > 0.05$) between total fixation duration and relative frequency.

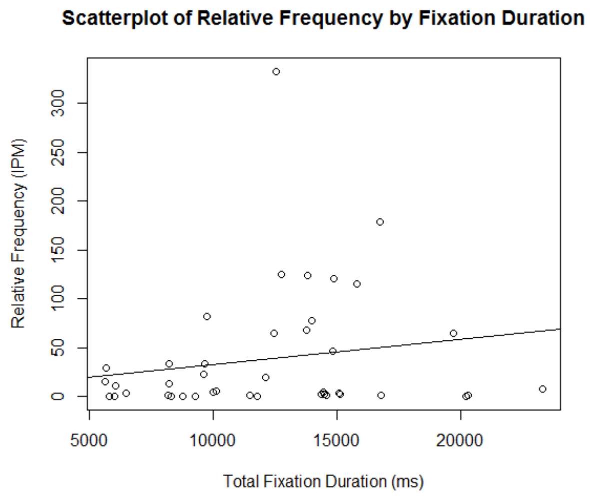


Figure 55. Scatterplot of relative frequency by total fixation duration.

An analogous situation arises for mean fixation duration. The data were found to be normally distributed, yet since frequency departs from normality the Spearman's rank correlation was computed. In fig. 56 frequency and mean fixation duration are plotted together. Just like for fixation count, this scatterplot reveals virtually no relationship between the variables, as it is indicated by the reference line being almost entirely flat and the data being scattered rather than tightly clustered around the line. Correlation results show confirm there is no significant relationship ($r_s = -0.02$, $n = 42$, $p > 0.05$) between mean fixation duration and relative frequency. Interestingly, the very small relationship is negative, meaning that the most frequent a subtitle is, the least time on average people spend looking at it.

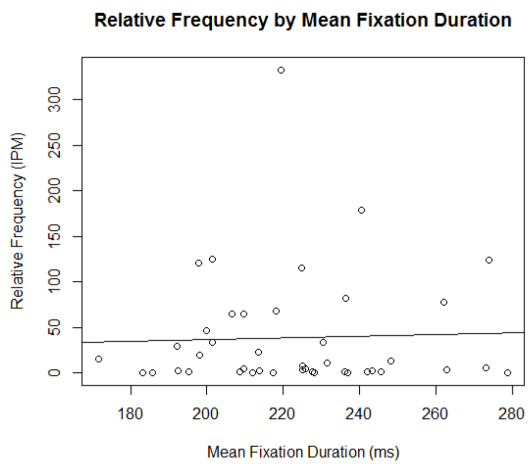


Figure 56. Scatterplot of relative frequency by mean fixation duration.

4.3.8 Generalised Linear Mixed-Effect Modelling: Logistic Regression Analysis

4.3.8.1 Introduction

So far, this chapter mostly featured tests of relationships (correlation) and differences between mean recall (t-tests), consistent with those used by Ghia (2012a) in her parallel experiment on standard subtitles. However, these tests consider only a pair of variables at the time, i.e. they do not do not allow the simultaneous analysis of how two or more variables affect the same outcome. As we have seen in 4.3.5.2, carrying out multiple t-tests can cause type I error so, traditionally, ANOVA analysis (Clark, 1973) has been used to work around this issue and explore how much of the total variance in the dependent variable is accounted for by two or more individual predictors. However, in practice, by-subject (F_1) and by-item (F_2) analyses of variance are not performed on raw data but first require aggregation (e.g. across items for F_1 analyses) as to obtain one observation per subject per cell and enable ANOVA testing on these cell means (Barr et al., 2013: 260). Aggregation *de facto* means disregarding important information, as these procedures reduce variation by collapsing the data into averages (Winter, 2013). Moreover, categorical and count data by definition cannot be normally distributed (Quené and van den Bergh, 2008), which is the main assumption of most parametric statistical tests used in language research, including t-tests and ANOVAs. Although ANOVA procedures are still customary in psycholinguistic research (Locker et al., 2007), they have several inherent limitations (see Locker et al., 2007; Barr, 2008; Dixon, 2008; Quené and van den Bergh, 2004 and 2008; Winter, 2013; Barr et al., 2013), which have lead researchers to turn to a relatively new set of statistical tools that can both overcome the averaging issue and cope with categorical dependent variables: generalised linear (mixed-effect) models. Generalised Linear Models (GLMs or GLiMs) allow to run a regression-like analysis where the dependent variable does not need to be normally distributed but is typically assumed to follow an exponential family distribution, e.g. normal, binomial, negative-binomial, multinomial, gamma, Poisson, inversed-Gaussian and so on (Turner, 2008). GLMs allow modelling count data and binary responses as a function of covariates without assuming normal distribution of the errors, nor a linear relationship between the dependent and independent variables. However, these models do assume independence of errors, which is still unrealistic in many real-life research applications. This is where Generalised Linear Mixed-Effect Models (GLMEMs, GLMMs or GLiMMs) come into play. In this class of analyses, a non-independent or clustered binary response is modelled as a function of covariates taking into account the idiosyncrasies of each individual cluster, which are investigated directly through the addition of random effects. This last part of the data analysis therefore describes how a generalised linear mixed-effect was fitted to the accuracy data in this experiment.

In our dataset, recall data from 26 participants is crossed with 22 test items so that there are 572 unique accuracy observations. The dataset has multiple accuracy responses from each subject, which are likely to be correlated and cannot be considered independent. A random intercept for subject was therefore added to the model, which makes it possible to assume a different baseline accuracy for each subject, thus accounting for participant-specific differences. A by-item random effect was also added to capture subtitle-specific differences. Having random by-subject and by-item intercepts means that subjects are allowed to vary around the intercept β_0 (Barr et al., 2013: 259). In building the model, we followed Dixon's recommendation: "[…] an appropriate strategy with logistic regression models is to proceed incrementally by adding effects one at the time until the most parsimonious fit is obtained." (2008: 455). Thus, the model was built step by step and upwards (bottom-up), guided by the experiment design and the previous explorations of the data addressed in this chapter. Predictors and interactions were added one by one, while keeping all other variables (fixed and random effects) constant. Factors were retained if they improved model fit, which was assessed by running ANOVA comparisons to check whether reduction in the residual sum of squares between models was statistically significant. The *glmer()* function[28] in the *lme4* package (Bates, 2010) was used to fit the models presented below. The statistical tool used is R (R Development Core Team, 2017), specifically R Version 3.4.0 "You Stupid Darkness" (2017-04-21). All tests were conducted in RStudio, a popular open-source user interface for R. RStudio was chosen because it is designed to make working in R more productive and intuitive, it facilitates data viewing, it includes a console and an editor that allows for syntax-highlighting, it supports direct code execution, and it integrates tools for plotting, activity history and workspace management (RStudio, 2016).

[28] The function *lmer()* produces a *linear* (i.e. Gaussian) mixed effects model while *glmer()* produces a generalized mixed effects model with a *non-linear distribution*, for example from the Poisson or binomial family. Although the *glmer()* function is more appropriate to the data structure of the present experiment, if *lmer()* is called with a non-Gaussian family argument specified, R automatically applies *glmer()*, and if *glmer()* is called with no family or a Gaussian family specified, R reverts to *lmer()*. So either of the two can be used as long as the correct family of distributions is specified.

4.3.8.2 GLMEM Analysis

In this logistic regression analysis, the effects of translation condition and other independent variables – both participant- and subtitle-related – on recall accuracy scores are ascertained. To obtain the regression parameter estimates, all models were fit by maximum likelihood (see Bolker et al., 2008) using the Laplace Approximation, the default iterative algorithm for likelihood estimation with *glmer* in R. Since the outcome variable in this analysis is dichotomous (accurate vs. inaccurate recall), the family of distribution specified for this regression analysis was binomial.

The null model (Model 0) was fit first, which had no predictors but only the intercept and the two random effects, one for subject and one for item. Model building started by adding translation condition to this baseline regression equation, thus obtaining Model 1. The estimated fixed effects for this model are presented in table 23, where β is the coefficient estimate for each fixed effect, SE is the standard error and the Wald z-score (henceforth Wald Z) is the coefficient estimate divided by the estimate for its standard error. In absolute value, i.e. regardless of the value sign, z-scores describe how far from zero is the SE of the coefficient estimate (Jaeger, 2008: 440). The model significance depends on how large this standardised distance from zero is, and "coefficient significantly larger than zero increase the log-odds (and hence odds) of the outcome" (ibid.), i.e. an accurate response in the recall post-test. As expected based on previous analyses, there was a very visible main effect for translation condition ($\beta = 0.90$, $SE = 0.20$, Wald $Z = 3.37$, $p < 0.0005$), which confirms this variable is a significant predictor. Positive β values represent a positive relationship. In Model 1, the baseline level for translation condition is the non-literal level ($N = 0$). The probability of recall success is calculated by going from 0 to 1 which, in our case, means going from the non-literal to the literal ($L = 1$) condition. In other words, the positive relationship between formal similarity (literal transfer) and accurate recall is confirmed by Model 1. As far as the random effects are concerned, their variance and standard deviation is reported in table 24.

Predictor	β Coefficient	SE	Wald Z	p-value
Intercept	0.6493	0.1925	3.373	0.000743 ***
Translation Condition = Literal	0.9003	0.2026	4.444	8.82e-06 ***

Table 23. Summary table of fixed effects for Model 1, i.e. a generalised mixed-effect model of response accuracy in the verbatim recognition post-test, with translation condition as predictor (baseline level = non-literal). Number of observations n = 572; deviance 642.1.

Random effect	s^2	SD
Subject (intercept)	0.1948	0.4414
Subtitle (intercept)	0.2750	0.5244

Table 24. Summary table of random effects for Model 1, where s^2 is the variance and SD is the Standard Deviation.

When running analyses of variance (ANOVA) between pairs of models, a series of model assessment indices (e.g. AIC, BIC, deviance statistic) are produced. As explained in 4.3.5.3.1, the deviance statistic is minus two times the log-likelihood (-2LL), and a value increase signals a poorer model fit. The AIC (Akaike's Information Criterion) and the BIC (Schwarz's Bayesian Criterion) are adjusted versions of the log-likelihood that correct for model complexity by taking into consideration the number of parameters in the model (Field et al., 2012: 867-868). The BIC is comparable to the AIC but is more conservative and is to be used with large sample sizes and a small number of parameters (ibid.). Neither of these two values are interpretable *per se*, but they are useful to compare models, and the model that provides the better fit to the data will always have a lower value. Model 1 significantly improved the fit to the data compared to Model 0, as demonstrated by a highly significant AIC difference (AIC Model 0 = 668.52; AIC Model 1 = 650.08; $p < 0.0001$) and the reduction in deviance statistic (-2LL Model 0 = 662.5; -2LL Model 1 = 642.1).

Then the effect of participant-related factors on recall scores were explored. Proficiency did not have a main effect on recall scores (β = 0.052, SE = 0.051, Wald Z = 1.019, $p > 0.3$), nor did its interaction with translation condition (β = -0.087, SE = 0.070, Wald Z = -1.237, $p > 0.2$) or that with working memory (β = -0.006, SE = 0.040, Wald Z = -0.161, $p > 0.8$). No main effect was registered for age (β = -0.012, SE = 0.028, Wald Z = -0.433, $p > 0.6$) or gender (β = -0.253, SE = 0.222, Wald Z =-1.138, $p > 0.2$), i.e. recall accuracy scores did not vary between male and female participants, and younger subjects performed similarly to older subjects in this data set. From the by-subject analysis (4.2.5), it emerged that subjects in both group G1L and G2N looked at subtitles significantly more and for longer in the second half of the clip. The by-item analysis (4.3.6.3 and 4.3.6.4) seemed to confirm this for the test items ($n = 22$), but not for the whole dataset ($n = 110$), where no significant difference in fixation duration was found between the first and the second half of the clip (the same applied to fixation count). A generalised mixed-effect model revealed no main effect for the variable group ($\beta = -0.325$, SE = 0.364, Wald Z = -0.89, $p > 0.3$), nor did its interaction with translation condition ($\beta = 1.082$, SE = 0.562, Wald Z = 1.924, $p = 0.054$). However, since the interaction approached significance, this model was compared to the Model 1 to see if the addition of the interaction term improved model fit. No such improvement was recorded (AIC Model 1 = 650.08; AIC Model 2 = 650.26; $p > 0.1$) and the interaction term was therefore excluded from subsequent models.

Significant main effects emerged instead for fixation measures (Model 3) and working memory (Model 4). Both fixation count and duration significantly improved model fit and had a significant main effect when added to previous model equation individually. However, since we know that these two fixation measures are very highly correlated (see fig. 20 in 4.2.4.4), only one was kept in the model to avoid biasing the model estimates. Since fixation duration is calculated in milliseconds, it results in very high values on the totals (1 minute = 60,000ms), which creates some issues when fitting *glmem* models (large Eigen values) and can make the model unidentifiable. Therefore, total fixation count was chosen. ANOVAs were run for comparisons and these models significantly improved the fit to the data compared to the previous ones, as evidenced by lower AIC and deviance values. Due to space constraints, the intermediary steps are omitted, but summaries of fixed and random effects for this final optimal model (Model 4) can be found in tables 25, 26 and 27. Translation condition, working memory (WM) and total fixation count emerged as highly significant predictors of verbatim recall scores. As far as the fixed effects are concerned, all β estimates are highly significant and no predictors are correlated.

Predictor	β Coefficient	SE	Wald Z	p-value
Intercept	-2.8731	1.02033	-2.816	0.00486 **
Translation Condition = Literal	1.03132	0.20906	4.933	8.09e-07 ***
WM	0.27023	0.10222	2.644	0.00820 **
Total Fixation Count	0.18963	0.04518	4.197	2.70e-05 ***

Table 25. Summary table of fixed effects for Model 4, i.e. the optimal generalised mixed-effect model of response accuracy in the verbatim recognition post-test, with translation condition (baseline level = non-literal), working memory and fixation count as predictors. Number of observations n = 572; deviance 619.97.

Random effect	s^2	SD
Subject (intercept)	0.03745	0.1935
Subtitle (intercept)	0.19077	0.4368

Table 26. Summary table of random effects for Model 4, where s^2 is the variance and SD is the Standard Deviation.

Correlation of Fixed Effects:			
	Intercept	Trcond = L	Working Memory
Trcond = L	-0.226		
Working Memory	-0.964	0.112	
Total Fixation Count	-0.296	0.199	0.087

Table 27. Correlation of fixed effects for Model 4.

The positive sign of all coefficients in table 25 (intercept aside) reflect a positive relationship between the variables, indicating that the higher the number of fixations on a subtitle and the WM of the subject, the higher the probability of correctly recognising that subtitle in the post-test. For translation condition the positive sign means that, as we move to literal items, the probability of recall success significantly increases. Model 4 (in which working memory was added to the regression equation) significantly improved model fit compared to Model 3 (translation condition and fixation count) as demonstrated by the significant difference in AIC (Model 3 = 636.24; Model 4 = 631.97, p = 0.012) and the log-likelihood reduction (-22L Model 3 = 626.24; -2LL Model 4 = 619.97). Both working memory and total fixation count are therefore important predictors that need to be kept in the model.

Item-related factors were also considered. The variable half (first and second half of the video) was also investigated and did not add any explanatory power to the model (β =

0.325, SE = 0.364, Wald Z = 0.893, $p > 0.3$). No significant effect was found for the interaction of half with translation condition (β = 0.432, SE =0.531, Wald Z = 0.814, $p > 0.4$) or half with fixation count (β = 0.09015, SE = 0.096, Wald Z = 0.932, $p > 0.3$), confirming the pattern of results emerging from 4.3.6. Linguistic category did not affect recall performance (β = -0.084, SE = 0.329, Wald Z = -0.258, $p > 0.7$), not even in interaction with translation condition (β = 0.133, SE = 0.415, Wald Z = 0.321, $p > 0.7$), upholding the non-significant results found in 4.3.5.3. Nor did linguistic category interact with viewing order (group) (β = -0.277, SE = 0.434, Wald Z = -0.637, $p > 0.5$). There is no main effect for subtitle duration (β = -0.002, SE = 0.01, Wald Z = -0.253, $p > 0.8$), nor does the interaction between subtitle duration and viewing condition (group) contribute to recall accuracy scores (β = -0.009, SE = 0.0148, Wald Z = -0.652, $p > 0.5$). Subtitle duration does not interact with linguistic category either (β = 0.019, SE = 0.0154, Wald Z = 1.249, $p > 0.2$). Subtitle length does not affect recall accuracy individually (β = -0.129, SE = 0.080, Wald Z = -1.610, $p > 0.1$), i.e. it does not have a main effect on the outcome, but its interaction with translation condition appeared to have some significance (β = 0.223, SE = 0.102, Wald Z = 2.184, $p = 0.02$), so a comparison between this model (Model 11) and our optimal model (Model 4) was carried out. ANOVA results show a very small decrease in deviance (-2LL Model 4: 619.97; -2LL Model 11: 614.68) as well as a very slim reduction in AIC value that approaches significance but is not statistic (Model 4: 631.97; Model 11: 630.68; $p = 0.07$), suggesting that this variable does not contribute to the fit of the model and should therefore be removed. This choice is corroborated by the experimental design, where subtitle length was controlled for precisely to minimise differences in the translations. Despite subtitle length having to change from L to N to allow the creation of two different translation versions (and recall that subtitle duration was also kept identical between conditions), the difference in number of words between L and N was kept minimal, with the two renderings differing by one (34.5%), two (12%) or three (4.5%) words maximum. Moreover, in several cases (48% of the total), the number of words in a subtitle remained exactly the same between the two translation conditions, *de facto* for the most part eliminating subtitle length differences between translation conditions.

To summarise, a logistic regression analysis was run by fitting generalised mixed-effect models to the data. No main effect were found on recall scores for Italian proficiency, linguistic category, viewing order (group), clip half, gender, age, subtitle length and subtitle duration. Moreover, no noteworthy interactions between these variables emerged. The most parsimonious model that provided the optimal fit to the data remained Model 4, which included translation condition, working memory and fixations count as fixed effects, as well as random intercepts for subjects and subtitles. Significant main effects were found for all these predictors. The fact that no significant interactions between these predictors and other variables were found makes the results more directly interpretable, whilst also confirming that the experimental design was successful in controlling for confounds related to both subjects (such as proficiency) and item (such as subtitle duration). The findings once again confirm that translation condition is capable in itself of affecting recognition memory scores, and so are individual working memory differences and the amount of fixations subjects devote to the subtitles during viewing.

CHAPTER V
DISCUSSION

5.1 Introduction

This chapter will discuss all findings emerging from the data analysis in order to inform the answers to the RQs outlined in chapter 3 (for a summary, see table 2). Preliminary to that discussion, in the first part of this chapter (section 5.2), I will also address the variables that were controlled for in this study, both those related to the experimental subjects (language proficiency and working memory) and those related to the experimental stimulus (subtitle duration, subtitle length and order of presentation). Then, in the second part of the chapter (section 5.3), I will move to the main research variables. I will start by addressing the relationship between recall and translation condition (RQ 1), the primary variable of interest in this piece of research. I will present a summary of the statistical results, address a sample of test items directly and, in doing so, discuss possible interpretations of the findings. Next, I will consider the other main variables in the following order: linguistic category (RQ 2), eye movements, i.e. fixation duration and count (RQ 3), and relative frequency (RQ 4). Finally, the open-ended questionnaire data will be qualitatively analysed and discussed in two sub-sections, addressing the implications of the findings first for different types of noticing (RQ 5), and then for potential language learning. A discussion on the limitations of the present study and possible future design improvements (section 5.4) will conclude the chapter.

5.2 Control Variables

5.2.1 Subject Control

5.2.1.1 Language Proficiency Test

In section 4.2.3 of the previous chapter, it emerged that the correlation between the language proficiency scores and recall scores was small and did not reach significance. The relationship might not emerge clearly for a number of reasons. It is reasonable to believe that there is a relationship between L2 language proficiency and L2 memory attainment, and that language knowledge assists and facilitates recall. However, it is also true that irrespective of foreign language knowledge, people have different memory spans and this may have a more direct impact on verbatim L2 recall than language proficiency level. A second possibility is that the language test (see Appendix A.1) was not proportionate to the actual level of the students. The minimum pass score in the language test was 60%, i.e. 21/37 correct answers. However, the actual scores registered ranged between 26 and 36 (M = 30.8; Q1-Q3 interquartile range: 28-33) indicating that, on average, subject performance was slightly higher than the minimum allowed by the test (see also fig. 13). It could therefore be that the level of some students was actually C1 or somewhere between B2 and C1, and that perhaps a C1 level test might have had more discriminatory power. However, this numerical information can be gathered only after the tests have been administered, and some participants had to be discarded from the analysis because they did not pass the language test, so the level was deemed to have an appropriate discrimination power. Moreover, an upper-intermediate level of Italian is sufficient to deal with the non-specialised colloquial language used in the film, and an excessively high-level test would run the risk of being unnecessarily discriminatory, preventing participants from taking part in the study despite their language level being appropriate to understand the subtitles, thus losing valid data. Since language proficiency was not a main research variable but a control variable, the aim was not to recruit participants with different language levels in order to assess the effects of proficiency on recall, but rather to gather participants with as comparable a level as possible. This goal was achieved through recruitment parameters (see section 3.8 of the methodology chapter) and the introduction of the language test. The crucial finding revealed by this analysis is that, at the same level of proficiency (CEFR B2), individual language score differences do not significantly affect recall. Another subject-specific variable, working memory was found to have a stronger relationship with recall.

5.2.1.2 Working Memory

As we have seen in section 2.9, working memory has been found to play a role in language processing and acquisition, *in primis* because cognitive tasks can be successfully completed only with a sufficient ability to maintain information as it is processed (Cowan, 2010). We therefore expected that subjects with higher working memories would achieve higher performance in the recall test. The relationship between this participant variable and recall scores was addressed in the by-subject analysis through correlation. The presence of a relationship between recall post-test scores and WM was confirmed in the statistical analysis (see scatterplot in fig. 12). The Pearson correlation approaches significance although the p-value remains just above the cut-off point 0.05. As demonstrated by the power analysis (0.24), there is a clear issue of power and sample size with this test. A power level of 0.24 means a 76% chance of not finding the real result, despite the effect size being large. Another reason why the correlation is not stronger in this dataset is that some subjects achieved a relatively high accuracy despite their relatively low WM score.

The scatterplot of recall accuracy and WM in fig. 12 reveals two specular data points at the top of the graph, corresponding to two subjects (S_4 and S_{11}) who achieved the same high accuracy result in the post-test (21/22 correct responses) while displaying two rather different WM abilities (8.24 and 10.55[29] respectively). This could be explained by the fact that the information was presented in a multimodal[30] environment, where sources of information other than the written word were presented. This may have helped subjects with a lower working memory (such as S_4) achieve a relatively high recall score in the post-test. Although no text is, strictly speaking, monomodal (Gambier, 2006: 6), reading an AV text is obviously not the same as reading a paper text. Besides the medium change from analogue (paper) to digital (screen), and the fact that the support is different (not a newspaper or a book but a PC, tablet or other electronic device), it is the way in which viewers read the text that fundamentally changes. Reading is no longer linear as the eyes move continually back and forth between the images and the written text along a vertical axis (deflections) as well as along the more traditional horizontal one. Images need to be processed too (Kress and van Leeuwen, 1996) and a wealth of other semiotic resources contribute to the communication of meaning (Baldry and Thibault 2006: 18-19), which are also to be processed. For example, verbal elements such as newspaper headlines or restaurant menus intertwine with paralinguistic features such as intonation and accents as well as with non-verbal elements such as lighting, proxemics, film cuts and framing, just to name a few (for a recent review of the semiotic codes contributing to the production of meaning in AVT, see Gambier, 2012). Moreover, the original L1 aural source text and the L2 subtitles appear simultaneously, allowing for multiple comparisons of the two language codes (Mariotti, 2002; Danan, 2004). Such moment-by-moment cognitive comparisons of input are made in the STM store and characterise processing for language learning, where the increased L1-L2 mapping can result in internal representations being made and can produce new insights (Doughty, 2001: 213-214). In previous studies, this "deeper processing" (Lambert et al., 1981: 134) involved with reading L2 (reverse) subtitles was reflected in better memory and found to aid the development of contextual meaning, spelling and phrasing (ibid: 146-147). This "unusual type of reading" (Lavaur and Bairstow, 2011: 462), where information is presented

---

[29] For an explanation of how WM scores were calculated, see section 3.10 in the methodology chapter. For a breakdown of all WM test components, see Appendices A.2.1 to A.2.5.

[30] For a definition of multimodality, see the end of section 6.5 in the conclusions chapter.

multimodally, could have assisted viewers with a lower WM by having a facilitative effect on their short-term verbatim retention of words and structures.

Moreover, since this was a critical variable to be controlled for (2.9), subjects with very low and very high WM scores were excluded from the study, which meant that the data was relatively balanced, i.e. it did not contain extreme yet legitimate values that would have contributed to statistical significance in the correlation test. The fact that working memory was still found to have an effect between medium and large ($R^2 = 0.14$) on verbatim recall of L2 subtitle wordings further confirms that working memory and recognition memory are strongly linked. This strong relationship emerges clearly in the generalised linear mixed-effect model analysis (4.3.8.5), where clear main effects were found for WM ($\beta = 0.270$, SE = 0.102, Wald Z = 2.644, $p = 0.008$) on recall scores. By showing that there is a relationship between WM and verbatim recognition, this correlation is aligned with the current literature and upholds the view that WM affects second language processing and, more broadly, acquisition. The finding confirms the results of numerous other studies that highlighted the role played by working memory in several language-related spheres, such as sentence comprehension, native vocabulary knowledge, novel word learning, sentence parsing and text understanding to name but a few (see section 2.9, as well as Szmalec et al., 2013 for a review and discussion).

5.2.2 Experimental Stimulus Control

5.2.2.1 Subtitle Duration and Length

Subtitle duration and length were controlled for in order to keep the translations of the same source strings comparable and ensure the internal validity of the experiment. As predicted, subtitle duration significantly correlated with length, since the longer a subtitle is in terms of number of words, the longer it will have to remain on screen to allow for comfortable reading by the viewer. Subtitle length ranged between 1-13 words. The minimum duration of a subtitle was 20fr (just under a second), while the maximum was 5sec 21fr (just under six seconds). From all analyses, it appeared that neither subtitle duration nor length significantly affect recall accuracy. From 4.3.3 and 4.3.4, no significant correlation between the variables emerged. Logistic regression analysis (4.3.8.5) confirmed that the only interaction found (between subtitle length and translation condition), did not significantly improve model fit to the recall accuracy data. Moreover, no main effects were found for subtitle length and subtitle duration, and no significant interactions between subtitle duration and other variables. These results mean that, in this experiment, recall does not depend on how long a subtitle is, be it physically or temporally. Therefore, in this dataset, a 13-word subtitle has the same chances of being remembered correctly as a one-word one. Conversely, the shortest subtitle, remaining on screen for under a second, has the same chances of being retained in memory as the longest one, which remains on screen for just under six seconds, i.e. is more than five times as long. The analysis also revealed that subtitle duration significantly correlated both with fixation count and duration. This was expected, as the longer a subtitle, the more likely it is to attract more looks and the longer the total duration of those fixations will be. Furthermore, as previously mentioned, both in academia (see Munday, 2009: 149) and in the industry (see, for example, Ofcom, 2015) it is commonly accepted that if subtitles stay on screen for too long in relation to their length, viewers might be inclined to re-read them (De Linde and Kay, 1999), which also results in an increase of fixation duration and length as subtitle duration itself increases. The findings confirm that these item-specific variables do not present a confound and do not have a significant impact on the outcome of interest.

5.2.2.2 Order of Presentation

This study adopted a standard 2x2 Latin square design with complete counterbalancing (see section 3.5). This design allows us to look directly at potential effects of the order in which the participants were exposed to the translations by including 'order of presentation' (i.e. group assigned) as an independent variable in the analysis (Field and Hole, 2003: 82). During pilot analysis, it was found that subjects, regardless of the experimental group they were assigned to, devoted significantly more and longer looks to the second half of the clip, regardless of the way in which it was translated. As far as fixation duration is concerned, the by-subject analysis on the 22 test items found that, on average, subjects in G1L fixated for 15s in total on subtitles in the second half of the video and only 8s in the first half, while those in G2N fixated 12s in total on subtitles in the second half of the video and only 8s in the first half (see table 11 in 4.2.5). A similar relationship was found for fixation count (see table 10 in 4.2.5): on average, the 13 subjects in G1L looked 71 times in total at the subtitles appearing in the second half of the video and 40 at those in the first half. The 13 subjects in G2N behaved similarly, looking 61 times in total at subtitles appearing in the second half of the video and 41 at those in the first half. The reason for such eye behaviour stems from the higher speech density of the source audio in the second half of the clip, where there are fewer silent pauses (absence of dialogues) and conversations between characters are denser, i.e. more words are uttered in verbal exchanges. Because dialogues in the L1 source audio are mirrored by the corresponding L2 translations, this speech density should result in longer subtitles (both temporally and physically) in the second half, where subjects would have more written L2 text to read, in turn leading to more and longer fixations overall. Order of presentation was further explored through by-item tests in order to establish whether this apparent lack of balance between halves was real and had an impact on recall. The by-item tests allowed to obtain more fine-grained results, since subtitle measures could also be included in the analysis and information on the whole experimental stimulus (110 subtitles rather than just the 22 test items considered in the by-subject analysis) could be gathered.

Subtitle duration (4.3.6.1) and length (4.3.6.2) behaved similarly with regard to order of presentation: the t-tests revealed a significant difference between halves for the 22 test items (fig. 44 and 46), which was to be expected, given the aforementioned direct correspondence between fixation and subtitle measures. However, for both subtitle duration and length, this difference disappeared when the whole experimental stimulus was analysed: when all 110 subtitles are considered, subtitle duration in the second half of the video was on average only 7 frames longer than in the first half. Conversely, subtitle length in the second half was on average only 0.2 words longer than in the first half.

Fixation measures were also reanalysed in relation to order of presentation in the by-item analysis (4.3.6.3). Since fixation count behaved like fixation duration, I will use the latter as means of example in this discussion. In the 22 test items analysis, *total* fixation duration was significantly longer in the second half compared to the first (34s vs. 21s). *Average* fixation duration, on the other hand, was not very different between halves: each subtitle was fixated on average for 227ms in the second half and 217ms in the first, and this difference of 10ms was not significant. Moreover, when the whole experimental stimulus is considered, neither total nor average fixation duration display a significant difference between halves. On average, each subtitle is 226ms long in the second half and 220ms in the first. These results show that the difference in fixations between halves is significant only if one considers the raw totals in isolation. In other words, when durations on a subtitle by a subject are *not* considered out of context but in relation to the number of fixations made by that subject on that subtitle, fixation durations are not significantly affected by order of presentation anymore. That is to say, if the overall fixation durations are divided by the number of fixations they correspond to, the average time a person spends looking at subtitles between halves is quite similar.

Finally, in order to assess whether the main outcome variable was affected by the differences in halves detected, recall accuracy was analysed in relation to order of presentation (4.3.6.4). Subtitle wordings were recalled slightly more accurately in the second part of the clip, but recall rate was comparable between halves and the t-test results clearly indicate that the difference is not significant. It is therefore possible to conclude that order of presentation does not influence recall and experimental validity is maintained. A mixture of harder- and easier-to-recall items are present throughout, and which part of the video a subtitle belongs to does not affect its chances of being remembered correctly.

In summary, the order of presentation analysis revealed the presence of some differences in subtitle-specific variables between the two halves of the video clip, showed that such differences do not impinge on participant variables such as fixation duration and length, and established that order of presentation of the stimuli does not affect the main outcome variable in this study, recall accuracy. This type of analysis is much overlooked in psycholinguistic experiments involving subtitles, despite its crucial importance for the intrinsic validity of any study. Without controlling for order effects, there is no way of checking that the experimenter's manipulations do not accidentally affect results. In this sense, this study improves on Ghia's (2012a) noticing experiment by explicitly taking into account this control variable and reporting the results of the analysis, thus confirming that the present investigation is sound in terms of manipulation of the experimental stimulus.

5.3 Primary Research Variables

5.3.1 Recall by Translation Condition

The analyses in sections 4.2.1 and 4.3.1 established that the highest recall rate in the post-test (see Appendix A.3) is achieved with literally translated input. In the by-subject analysis, 65.4% of the subjects obtained higher recall accuracy in the literal condition (L), 15.4% obtained higher accuracy in the non-literal condition (N) and 19.2% obtained equal accuracy in both conditions. That is to say, 17 subjects out of 26 recalled items more accurately in L, while five subjects accurately recalled an equal number of items in both translation conditions. Only four subjects recalled items more accurately in N. By taking a closer look at these four subjects, it becomes evident that said inverse performance superiority is rather modest, since in most cases the proportion of accurate answers they provided in N is only slightly higher than in L. For the most part, recall is superior in the literal condition, which is reflected in the highly significant t-test result, with a large effect size and good power. The by-item analysis upheld these results, revealing that 73% of the total (n = 22) subtitles tested were recalled more accurately when translated literally. Two subtitles (AOI.17 and AOI.45, i.e. 9% of the total) were recalled equally accurately in both conditions. Four subtitles (AOI.5, AOI.13, AOI.42 and AOI.91, i.e. 18% of the total) were better recalled non-literally, but again non-literal recall is only slightly higher than literal recall in these cases (see table 29 below for details). Overall, out of all 572 responses elicited in the post-test (286 L and 286 N), literal renderings were recalled correctly in 80% of cases, non-literal ones in 64%. The t-test established that this noticeable recall difference is highly significant (see 4.3.1). Translation condition was found to have a large effect on recall ($\delta = 0.93$) and the test again had a good power. Logistic regression analysis (4.3.5.3) revealed that literal items are 2.27 times more likely to be remembered correctly compared to their non-literal counterparts. Generalised linear mixed-effect modelling (4.3.8.5) showed a main effect of translation condition on recall scores, which remained highly significant ($\beta = 1.031$, SE = 0.209, Wald Z = 4.933, $p < 0.00001$) when adding all other significant predictors in the model. Therefore, it emerged quite clearly that, after one single exposure to reversely subtitled video input, literal translations are remembered more precisely than their non-literal counterparts regardless of subtitle length, duration and order of stimulus presentation. It appears that literalness (intended as the degree of formal and functional similarity between language items) has indeed a psychological reality and can influence short-term recognition memory for exact phrasing. The relatively high correct scores across conditions reported in table 13 in the data analysis chapter are also noteworthy, amounting to 72% accuracy in total. With one single exception (AOI.5), all post-test items across translation conditions were recalled by more than half the subjects overall, indicating the viability of using this type of AVT (reverse subtitles) in second language acquisition not only at beginner and intermediate level as Danan (2004), Mariotti (2002) and Ghia (2012a) suggest, but also at more advanced stages of learning.

As mentioned above, there was a minority of subtitles which did not follow the general trend expressed above. Looking at these exceptions can provide some useful insights into the way we process multimodal information as well as into what might be the best way to conceptualise 'literalness'.

Subtitle	Cat	Source	Literal Wording	Non-literal Wording	Recall Accuracy	
Highest Recall in N					**L**	**N**
AOI.5	Lex	It's different.	È differente.	È diverso.	15%	54%
AOI.13	Lex	The prophecy?	La profezia?	Il vaticinio?	69%	85%
AOI.42	Lex	Bingo!	Bingo!	Indovinato!	69%	77%
AOI.91	Syn	Without him, we're lost.	Senza di lui, siamo perduti.	Se non ci fosse, saremmo perduti.	77%	85%
Equal Recall in L and N						
AOI.17	Lex	Of the resistance.	Della resistenza.	Della guerriglia.	92%	92%
AOI.45	Syn	They smell good, don't they?	Hanno un buon profumo, no?	Senti che buon profumo!	77%	77%

Table 28. Summary of recall exception items.

The first thing to notice in table 29 is that most of these items (4 out of 6) are lexical rather than syntactic. If one thinks in terms of Ghia's translational salience, it would seem that item salience emerging from translation discrepancy is more mnemonically effective when it applies to lexicon than syntax. Ghia (2012a: 52) defines translational salience as a type of perceptual salience, i.e. "the prominence acquired by linguistic items when input is delivered". As such, it always involves a subject-internal component, i.e. a receiver – in the case of audiovisuals, the viewer – to interact with such delivery (ibid. 53), a gradable contrast (ibid.: 51) and a deviation from expectations (ibid.: 71). Thus, translational salience is a type of salience deriving from an accentuated, visible contrast between ST, i.e. the audiovisual dialogue, and TT, i.e. the subtitles (ibid.: 71). Indeed, Pavesi and Perego highlight how lexical items are "favoured loci of salience as opposed to longer syntactic structures" (2008a: 220). Broadly speaking, syntactic structures can generally be deemed to be typically more 'volatile', less tangible than lexical items, which in turn are more concrete and picturable, therefore potentially more memorable and more easily learnable incidentally[31]. Backing to this view also comes from FLT and FLL research,

[31] Of course this is, to a certain extent, a generalisation. On the one hand, some syntactic structures can be relatively easy to learn, especially if fully lexicalised; on the other, not all lexical items are

according to which some aspects of L2 grammar are difficult for adult learners to acquire without focus on form (Doughty and Williams, 1998; Laufer, 2006; Lightbown and Spada, 1990; Spada and Lightbown, 1993; White et al., 1991), whereas many believe that a great deal of vocabulary can be learnt incidentally, i.e. when attending to meaning rather than linguistic form specifically, for instance through reading (see, for example, De Keyser, 1998: 43). Issues of perceptibility may therefore distinguish lexical from syntactic items, whereby perceptibility is defined as "the ease with which a structure can be identified in the input flow" (Ghia, 2012a: 172). Structures with low perceptibility can be difficult to learn via implicit enhancement strategies alone (Doughty and Williams, 1998a), such as the formal L1-L2 discrepancy created in non-literally translated syntactic input. These differences between lexical and syntactic items might be particularly acute in the case of audiovisuals, where the often fast pace and the transient nature of subtitled text does not allow the learner to stop and dwell on the language encountered. In this dataset, it seems that only the most striking and therefore potentially more memorable non-literal items can leave a memory trace strong enough to be retrieved at a later stage, and it appears that these items tend to be lexical in nature. This was the case for non-literal renderings *guerriglia* (AOI.17) and *vaticinio* (AOI.13). In the former, non-literal recall reached the same level as the literal item *resistenza*, making the two items overall equally memorable (92% accuracy); in the latter, although the literal counterpart *profezia* also achieved a high level of accuracy (69%), the non-literal version surpassed it: 85% of subjects recalled *vaticinio* correctly.

As far as frequency is concerned, *guerriglia* is quite a rare term (6.1 IPM[32]) and *vaticinio* even more so (0.1 IPM). On the basis of this clear low frequency corresponding to accurate recall, one might be tempted to conclude that the more infrequent a non-literal noun, the higher chances it has to be recalled precisely in the recognition post-test. However, the other two non-literal nouns in our corpus of test items do not uphold this

concrete and picturable, and therefore might be less memorable. Though the above statements are by necessity a simplification, grammatical constructions and lexical constructions exist and have been posited to be able to have separate effects, e.g. in triggering the mental representations accessed by the learner (Seilhamer, 2010: 18). Lexical items have also been found to have an advantage over grammar when occurring as recasts in a study which used eye-tracking technology to investigate noticing specifically (Smith, 2010). Lexical recasts were found to be easier for learners to notice, retain, and use productively in writing than grammatical recasts (ibid.). Therefore, a broad distinction between syntax and lexicon can be made, both in theoretical and experimental research.

[32] The reader is reminded that the corpus used to extract such frequencies is itWac (Baroni et al., 2009). See section 3.14 for details.

view: both *cuore* (124.3 IPM) from AOI.105 and *studenti* (120.9 IPM) from AOI.53 have some of the highest relative frequencies in the whole test set and yet are also recalled correctly by more than half the subjects (84.6% and 61.5% respectively). Therefore, recall of non-literal nouns in this dataset specifically is quite high and does not appear to be modulated by frequency. I say 'nouns' specifically and not lexical items more generally because the relationship between recall and frequency for other, non-noun content words in the non-literal condition is less clear-cut. Sometimes other infrequent non-literal lexical items are recalled quite well (e.g. *intensamente*, 4.02 IPM and 69% accuracy; *indovinato*, 0.78 IPM and 77% accuracy); sometimes infrequent non-literal lexical items are *not* recalled very well (*persuaso*, 1.09 IPM and 38% accuracy) and sometimes relatively frequent non-literal lexical items are also not recalled well (*individuato*, 19.07 IPM and 38% accuracy; *non possibile*, 29.2 IPM and 30.7% accuracy). In their discussion on the effects of language similarity on FLL through subtitled video exposure, Van de Poel and D'Ydewalle stress how lexical items, and nouns in particular, are "the easiest building blocks in acquiring a new language" (2001: 271). The accurate non-literal lexical results gathered herein suggest that, amongst content words, nouns specifically might indeed have a special status in the mind of the learner as "anchoring points" during reading (Smith, 2004) – in particular reading of audiovisual texts – compared to other content words, with effects on memory that cannot be predicted by frequency alone. What (low) frequency might have done is assist subjects during the MCQ in the opposite, literal condition. One subject who saw AOI.13 and AOI.17 in L (*profezia* and *resistenza)* reported that he "instantly recognised" (S_7) these words in the post-test. Besides formal similarity, one of the reasons why recall accuracy is also quite high when AOI.13 and AOI.17 are translated literally (69% and 92% respectively), could also be precisely that the other test options (*vaticinio* and *predizione* for AOI.13, and *guerriglia* and *ostilità* for AOI.17) are infrequent and striking, so viewers will immediately discard them as non-suitable candidates unless they actually encountered them during watching.

Although both Ghia's (2012a) results and mine clearly favour the mnemonic superiority of literal translations, this study does not deny the role of salience claimed by Ghia. On the contrary, it lends support to the view that non-literal items will indeed be recalled by the learners in a visual recognition post-test if they are salient during watching. However, it appears that only items with a high degree of salience manage to reach the threshold necessary for later recognition to occur, in such a fast-paced environment like that of video and after just one exposure. It is not difficult to understand why, when faced with *resistenza – guerriglia*, one would be more confident in making a choice in the MCQ then when faced with *impossibile – non possibile*, regardless of which translation condition one actually saw. While in the former pair the two words correspond to defined and vivid concepts, and a clear nuance in meaning exists between the two, in the latter such a distinction in meaning is not immediately perceivable. In this sense, the N item *guerriglia* might be intrinsically more salient than the N item *non possibile* relative to their respective audio strings *resistance* and *impossible*. One of the subjects spontaneously reports on this very point in the open questionnaire: "I fully understood what had happened and was said - however, selecting the correct phrase [in the MCQ] was difficult, especially when they were almost identical, but meant the same thing" (S_{10}). Moreover, translation divergence might not automatically correspond to salience[33], especially when items are syntactic in nature. Non-literal syntactic renderings, for example like *scegliere* [to choose] for *to make a choice* (recalled by just over 50% of subjects) are often just another way of structuring the L1 message in the L2 and if this difference in structure either (a) is not immediately visible to the viewer or (b) is visible but does not particularly strike the viewer, it might be difficult for them to shift the detected (cognitively registered) input further into consciousness so that it is memorable enough to be confidently recognised in a later unexpected memory test. However, from this dataset it emerges that, if a non-literal item is lexical – a noun specifically – regardless of frequency it may have good chances of being noticed in such a way that it will be subsequently recalled. This was the case for all four nouns in the non-literal condition, which were recalled by 61.5% or more subjects.

What is more, not only were learners able to identify correctly non-literal renderings of nouns, but they also mentioned them of their own accord in the final questionnaire. The non-literal renderings for soul (*cuore* [heart]) and kids (*studenti* [students]) were spontaneously mentioned in open answers where participants were not presented with L2 phrasings or prompts. Particularly remarkable are the words *guerriglia* (AOI.17), reported by seven people, and *vaticinio* (AOI.13) actively mentioned by nine. Some participants did not remember the exact phrasing, but felt like mentioning these items anyway, by describing the situation they occurred in and by writing the first few letters

[33] At least if salience is defined as "prominence stemming from a *visible* contrast among input components" (Ghia, 2012a: 178). It is this visible contrast that then creates a situation whereby "an expression (phoneme, syllable, morpheme, phrase, etc) is projected onto the foreground in the learner's perception." (Py, 2004: 121, in Pavesi and Perego, 2008a: 220).

they recalled[34], often adding that they found them striking. In most cases, subjects were actually able to retrieve these words freely and correctly from memory: these were some of the most frequently free-recalled items in the questionnaire. Alongside reporting these two items, most subjects also reported they encountered them for the first time during the clip. This finding suggests that non-literal renderings, when salient enough to be recognised in an MCQ, can also leave a memory trace strong enough to make free-recall possible, thus indirectly promoting acquisition (see also 5.3.6).

Another item that was mentioned several times (n = 7) in the questionnaire was the pair *bingo – indovinato* (AOI.42). Although recall is high in L (69%), a higher number of participants recalled it correctly in N (77%). The frequency of *indovinato* (0.76 IPM) is higher than that of *bingo* (0.48 IPM), but by such a small amount that it is difficult to attribute the result to this variable alone since both words are, in fact, quite uncommon in the itWac corpus. The literal *bingo* is a case of complete formal identity between L1 source and L2 target and it is an acceptable translation for the ST in this emphatic context, so one might expect this item to be archetypal of literalness and to achieve much higher recall superiority in L. However, results show that this was not the case. It is possible that participants who saw the L item processed it more shallowly due to the L1 audio-L2 text identity, so that, when they found themselves in the MCQ, they could not remember whether they had only heard or also read the word *bingo*. In a situation of uncertainty, if one were not sure whether the same English word could actually be used in Italian in the same context, this would explain why they picked one of the two remaining options (non-literal N or distractor D) rather than the apparently more obvious literal solution. Evidence towards this interpretation comes from the questionnaire. Out of the seven subjects who actively mentioned AOI.42, six encountered it in the non-literal version (i.e. they belong to G2N). Two of them mentioned explicitly that they found *indovinato* an interesting solution and that they would have struggled to find an appropriate Italian translation. One stated it seemed more formal than the original (S_{22}). Other two subjects stated that *indovinato* stuck in their minds "because it is not a word I am familiar with in this context" (S_{13}) or because "I wouldn't have thought to have said it" (S_{14}). Almost every subject who explicitly recalls the item in the questionnaire does so because they saw the non-literal rendering and found it

---

[34] One such example is S_{14}, who wrote: "I had never seen this word and can't quite remember it but remember noting it, guegl... (war?)". Another is S_5: "I had not seen the word for prophecy before - va...".

interesting. Only one subject in G1L mentions seeing *bingo*, and her comment is rather enlightening: "They [the subtitles] weren't all exact translations. (…) Some things weren't translated at all, such as 'Bingo!', which is possibly not very idiomatic in Italian (I could be wrong)" (S_2). The use of *bingo* or *tombola* as an exclamation in these ironic contexts is acceptable, especially following the recent exponential growth in popularity of the game in Italy, both in bingo rooms and online (La Storia del Bingo, 2016). However, these emphatic expressions are quite rare (itWac rankings: 30,979 and 28,405 respectively) and have a marked orality trait, so unless the learners have spent some time in the target country, they will tend not to be familiar with the words in such context. Therefore, it is likely that other participants in G1L were not sure about the acceptability of this word in Italian, doubted their memory and picked an incorrect item in the MCQ, leading to the result for AOI.42 reported in table 29 above.

Subtitle AOI.5 deserves special attention because it is has a particularly idiosyncratic behaviour compared to the rest of the items. Not only is it recalled correctly by a slight majority of subjects (54%, n = 7) in the N condition, but it is also rather poorly recalled in L (15%, n = 2) despite the orthographic similarity with the audio (see table 29). Moreover, AOI.5 is the only subtitle in which fixation number and duration are also reversed. The by-subject analysis revealed a significant positive correlation between fixation count and recall: generally speaking, the more fixations a subject makes on an item, the more likely they are to recall it correctly in the post-test, regardless of translation condition. This was found to be the trend for total fixation duration too, although the correlation was not significant. AOI.5, on the other hand, receives more looks (46) by subjects who did *not* recall the item correctly in the post-test, compared to the looks it received (28) from those who did accurately recall it. The same goes for fixation duration, which was around 7s in total across participants and translation conditions in the case of subsequent accurate recall, 11s with inaccurate recall. The two translation versions of AOI.5 received roughly the same number of looks (36 in L, 38 in N), yet looking at the item in L did not correspond to accuracy in test performance. It seems that this particular item, when translated literally (L), despite the formal near-identity between L1 audio and L2 text (*different – differente*), creates confusion in the viewers, who end up choosing N or D in the MCQ (see table 3). Moreover, and again contrary to the otherwise quite clear general trend, amongst the six subjects who saw the item in N (*diverso*) but incorrectly recalled it, not a single person chose the back-translation of the audio (*differente*). It is quite evident that *differente* does not facilitate recall, nor is it considered an acceptable option in case of uncertainty in the MCQ. On the other hand, *diverso* does not seem particularly striking an item either; it is simply the standard Italian translation for the English source, uttered in the car scene where Neo and Trinity are talking about the oracle and the concept of 'difference' *per se* does not play a major role in the conversation. The pair *differente – diverso* is comparable to *impossibile – non possibile* (AOI.32), in the sense that in both cases the nuance in meaning between L1-L2 renderings compared to the source audio is not large enough to justify choosing the non-literal translation in the MCQ on the basis of its salience alone. So why did subjects recall *diverso* more accurately than *differente*? This could be a case where frequency has a more direct influence on processing, since the frequency of *diverso* (82.15 IPM) is much higher than that of *differente* (13.32 IPM). After all, this variable plays a well-established role in reading (Clifton et al. 2007) and was also found to affect subtitle perception (Moran, 2008; 2012). Perhaps the reason why frequency did not seem to play a role in the rest of non-noun content words in the non-literal condition is that all those items had relatively low frequencies (0.7 IPM min – 29 IPM max). This raises interesting questions for what is to be considered a 'frequent' or 'infrequent' word. This issue will be addressed specifically in the section on frequency (5.3.4) later in this chapter.

Another aspect that might have contributed to the observed recall results for AOI.5 is familiarity. Although often confounded with frequency, the two are not one and the same: high frequency in a corpus does not necessarily mean a learner will have already seen and know the item in question. As a matter of fact, two words with similar frequency can indeed differ in familiarity, and this is particularly true for infrequent words (Clifton et al., 2007: 6). As we have seen, the word *diverso* is much more common than *differente*, and in this case subjects are also very likely to be more familiar with its use in this context. In fact, at the post-intermediate stage of the subjects in this study, I can confidently say that such a common term would be the target one would *expect* to read upon hearing the source string *different*. The familiarity effect would explain why subjects did not see *differente* as a possible candidate in the MCQ: despite the orthographic similarity, it is both much more uncommon and unfamiliar a term, and it might have even looked like an unidiomatic rendering to them. This is not unlikely, considered that participants were proficient learners, who, unlike beginners, are naturally less reliant on L1-L2 surface form similarities during moment-by-moment L2 sentence processing (Talamas et al., 1999). This item therefore also indirectly highlights the role played by expectations in AVT processing. Ghia (2012a: 179) states that these types of translational expectations can often be linked to literalness (learners expect a target translation to be as close as possible to the original), but points out how deviations from this situation can arise with non-literal, frequently encountered alternatives (such as translation routines or clichés) that the learner has encountered several times before and therefore is familiar with and expects to see as translation equivalents of given ST strings. This is clearly the case with *diverso*, the translation solution *tout court* for the word *different* in this context. A crucial point on which I fully share Ghia's view is that the most literal item is not always the most natural (ibid.: 178). In the translation protocol of this study (3.11) L2 literal targets were chosen on the basis of their formal and semantic equivalence with the source, which, like in the case of *differente*, can sometimes override the naturalness of a translation, thus creating unmatched expectations in the viewers. The point of contention here is rather the consequences stemming from this situation on the viewer. Ghia states that "the unexpectedness of formal equivalence in such cases might have thus resulted in greater salience" (ibid.: 179). This may be true for syntax (Ghia's statement refers to cleft sentences specifically), but not necessarily for individual content words. My lexical data show quite clearly that such a situation can instead cause confusion in the learners, who do not seem to accept *differente* as a viable solution despite its near-identity of form with the English word they heard at the same time as reading the literal Italian. Such confusion results in as little as two subjects out of 13 correctly recalling the item in L, the lowest accuracy of the whole corpus of test items in this study. To summarise, *diverso* is more frequent and more familiar an item to the subjects, who know it to be a correct and most likely solution in that context, since they will have seen it several times before, whereas *differente*, a much less used word in Italian, defies expectations in the viewers who are not confident to choose it in the post-test, even if they actually read it in the subtitle during watching.

The observed finding for AOI.5 does not, however, contradict the significant role played by formal similarity in language development. During early stages of L2 vocabulary development, learners do show mixed influences from both L1 and L2 when producing as well as recognising L2 word forms (Hall, 2002: 71). As they develop language competence through production and recognition practice, however, learners also tacitly acquire knowledge of the sequential probabilities and frequencies of the L2 words and structures (Ellis and Sinclair, 1996), so that, at later proficiency stages like that of this experiment, other factors like expectations, familiarity and frequency effects can indeed override the importance of orthographic overlap in word recognition. AOI.5 therefore provides support to the view that formal L1-L2 similarity can facilitate language processing and mnemonic retrieval as well as cause interference when it clashes with other variables that influence L2 memory and learning.

Another interesting and rather evident discovery that emerged from the analysis is the clear perceptual confusion with non-literal items that arose in the MCQ. It does indeed seem that in some cases, viewers can be "unaware of whether text or speech is the primary provider of the language" (Vanderplank 1998: 964), which confirms the findings of De Bot et al.'s (1986) study, where viewers noticed a translation discrepancy but could not tell exactly which channel it appeared in. In the present study, when subjects incorrectly recalled non-literal items, they nearly always picked the back-translation of the source audio, that is to say, the literal version. This happened in 81.5% of all cases of recognition mistakes in N. The finding is consistent with a number of studies where purely formal similarity between competitor items were found to influence performance on various translation-related tasks (Dijkstra et al. 1998; Kroll and Stewart, 1994; de Groot, 1992). Such confusion might arise especially for L2 learners (as opposed to balanced bilinguals), who are more affected by such formal similarity (Talamas et al., 1999), and provides further evidence to the view that cross-language influence (CLI) is a real phenomenon occurring not only at the initial stages of language learning but also at more advanced L2 levels. In her noticing study on standard subtitling, Ghia also reports that in most cases learners chose the back-translation of the subtitles (2012a: 86) containing the Italian L1. In the case of standard subtitling, one might expect such interference, given that it was the learners' L1 to appear in the subtitles and that visual information is believed to persist for longer in memory than aural input (Field, 2004). Moreover, the learners' native language will necessarily exert a strong influence if learners do not understand parts of the L2 audio stream and therefore heavily rely on native language subtitles to successfully follow the video. On the other hand, the outcome with reverse subtitles, where the L2 input appears in the subtitles, was not as easily predictable. However, it is now clear that a similar situation arises in the reverse condition. Participants often read a non-literally-translated subtitle but picked the literal version in the MCQ, indicating that they were translating from the audio track. A strong tendency to rely on L1 information is herein confirmed, even when this information is conveyed in the more transient aural channel. Together with Ghia's, these findings provide a more complete picture of how CLI operates in multimodal environments: it appears that the dominant L1 exerts a strong perceptual influence on recall choices regardless of the channel through which it is conveyed.

5.3.2 Linguistic Category

Irrespective of translation condition, overall accuracy for syntax is 74%, slightly higher than for lexicon (70%). However, the t-test revealed that this overall difference in accuracy by linguistic category is not statistically significant and therefore could be due to chance alone. Consistent with Ghia's results in her study on standard subtitles, literal transfer consistently promoted higher recognition of both lexical items and syntactic structures compared to non-literal transfer. Ghia also found that verbatim recall in the translation divergence condition was higher for syntax than for lexicon (2012a: 87). This result is confirmed for reverse subtitles as well, although the recall difference is not very large: the average accuracy in the non-literal condition for syntactic items is 66%, against 63% for lexical items (see table 19). What Ghia does not clarify is the corresponding situation in the literal condition in her study. My data reveals that verbatim recall for syntactic patterns with literal transfer is also higher (83%) than recall of vocabulary under the same translation condition (78%). This overall higher accuracy for syntax could be explained in a number of ways. In previous comparative studies on the perception of audiovisual texts, when auditory information was processed in the L1, subjects were found to be more successful in concentrating on the corresponding L2 written input (Holobow et al., 1984: 68). If learners have more time and cognitive resources to dedicate to foreign input, they might be enabled to better attend to higher-level structures in the L2 written text that might have gone unnoticed, had the L2 been in the audio. Secondly, the structures tested in this study were, for the most part, lexicalised, which may have contributed to their overall higher perceptibility. Thirdly, sentences were relatively short, self-contained chunks, which might have enabled quick moment-by-moment familiarity checks and comparisons with their mother tongue during watching, compared to longer, more complex syntactic structures. Finally, the relatively high level of proficiency of the subjects is also likely to have assisted them in attending syntax more than just individual items. In fact, the open-ended questionnaire revealed a propensity to notice grammatical features in the audiovisual translated input: half of the subjects mentioned having noticed syntactic differences from the dialogue in the subtitles. Subjects explicitly reported (therefore consciously noticed) encountering phrases, tenses (in particular subjunctive and imperfect) as well as word order that piqued their interest. A subject even stated: "I read the subtitles as a grammar point, to see the position of 'ne', for example[35]" (S_{10}). Seven syntactic items were free-recalled as striking, three of which were test items (AOI.51, AOI.60, AOI.91).

A t-test of recall accuracy within syntactic items found a statistical difference in recall between items translated literally and non-literally, suggesting that literal transfer plays a significant role in the recall of syntactic structures. A t-test of accuracy within lexical items did not find a difference in recall between items translated literally and non-literally, although the CI is very close to not spanning zero and the t-test p-value itself ($p = 0.08$) is not far from the cut-off point for significance. Moreover, the boxplot in fig. 40 highlights a clear difference in medians and range of values, with lexical items on average being recalled by more subjects in the literal ($M = 10.16$) than non-literal ($M = 8.1$) condition. In fig. 40, lexical N items display more variable scores, being recalled correctly by a minimum of four subjects out of 13, whereas lexical L items – if one excludes AOI.5 (the outlier in the boxplot) – were recalled correctly by a minimum of nine subjects out of 13. Logistic regression was then used to further assess the relationship between these variables. Linguistic category was not found to be a significant predictor of accuracy scores and no significant interaction with translation condition emerged from the analysis (4.3.7.5). No main effect of linguistic category nor any interaction with translation condition emerged from GLMEM analysis either (4.3.8.5). Although the interaction was not significant, the interesting trends above were observable in the data and suggest that the variables are not completely independent. By analysing linguistic category and translation condition together, a subset of the data is *de facto* considered each time (e.g. just the items that are lexical in nature, and within those, the number of correct responses in L and N respectively). This means halving the sample size (in the case of this study, for example, from 22 to 11 test items, or from 26 to 13 participants), which in turn makes it more difficult to make failsafe inferences about the data. Moreover, it is possible that our classification of linguistic category (3.13), did not provide appropriate tools for detection of existing effects, if these are more nuanced. Maintaining the macro-division between lexical and syntactic items, while justified theoretically and motivated by an intention to keep results comparable to previous work on AVT manipulation and

[35] The Italian clitic *ne* is a pronoun with interesting syntactic properties, typically difficult to master even for advanced learners of IFL (Italian as a Foreign Language). Much has been written in the literature about this topic, see for example Belletti and Rizzi (1981) or Burzio (1986).

noticing (Ghia 2012a), might not allow a fine-grained enough analysis of micro differences in the input, allowing us to observe only general tendencies.

What emerges from this overall picture is that recall superiority of literal transfer seems to apply to syntax more visibly than lexicon. As we have seen, literal transfer is a good predictor of accurate recall of structures (83%), slightly less so of lexicon (78%). That is to say, as far as syntax is concerned, eight out of ten target structures were better recalled in L (across participants), one was better recalled in N and one achieved equal recall in both conditions. Out of the 12 lexical items, on the other hand, eight were recalled better in L, but three were better recalled in N and one item was recalled evenly in both conditions. The situation therefore appears to be more clear-cut with syntax. The fact that a few lexical items were equally or better recalled non-literally (see table 29 earlier in this chapter) could have contributed to the t-test of accuracy for lexical items not finding a significant recall difference between the literal and non-literal conditions despite the visible difference in the boxplot in fig. 40: lexical subtitles were overall better recalled in L, but there were enough cases of higher or equal lexical recall in N to lower the average difference enough to not warrant statistical significance. If, as it was mentioned above, lexical items are the preferred loci of salience compared to higher-level, syntactic structures, it could be reasonable to posit that, within the realm of lexicon, the recall difference between L and N conditions will be slimmer, since salient, non-literal individual content words are more likely to be noticed and, in some cases, processed in such a way that they will be recognised in an immediate verbatim memory test. Therefore, although formal similarity ensures recall accuracy of lexical items is for the most part still higher with literal transfer, the most salient non-literal lexical items can match or outperform their literal counterparts. Conversely, a translational discrepancy created in a syntactic structure could be intrinsically less salient to the viewers who, when faced with the test item in the MCQ, end up relying on the dominant L1 input, thereby choosing the back-translation of the structure as they heard it in the audio.

5.3.3 Eye Movements

First of all, the mean fixation duration on reverse subtitles recorded in this study is 217ms. This value confirms the expectations outlined at the end of 2.7.1 and is in line with d'Ydewalle and De Bruycker (2007), who registered mean values of 185ms and 201ms for reverse one- and two-liners respectively in their adult population. The slightly higher values recorded herein are to be expected, since, unlike in the present investigation, in d'Ydewalle and De Bruycker's study reverse subtitles were completely meaningless to the participants. Viewers will lack a reason to dwell on L2 words and will be less likely to engage in further cognitive processes such as lexical access (2.6.3) if it is not clear which L2 words are to be matched to what L1 counterparts. Moreover, in the present study subtitle translation was manipulated precisely to see whether differential effects would be found for the two translation conditions, which is likely to increase the overall fixation number and duration. These two variables were thus analysed in relation to translation condition to establish whether formal similarity and formal discrepancy are registered differently by viewers. The distribution of looks across conditions was analysed for the 22 test items as well as for the whole experimental stimulus (110 subtitles). By-subject (4.2.4.2 and 4.2.4.4) and by-item (4.3.2.1 and 4.3.2.4) t-tests revealed no significant effect of translation condition on eye movements. Items received slightly more and longer looks in the non-literal condition, but these differences do not appear to be significant. In the by-subject analysis, however, the power of the test is too small (0.09) to draw any incontrovertible conclusions, since, even if there was a genuine effect of translation on eye movements, the test would have very slim chances of finding it. In the by-item analysis, as far as fixation count is concerned, and with respect to the 22 test items, literal subtitles were fixated on average 59 times, non-literal ones 66. With respect to the whole experimental stimulus, the difference is thinner, with literal subtitles receiving on average 65 fixations, while non-literal ones 67. A parallel situation arose for fixation duration. As for the 22 test items, literal subtitles were fixated for 13s on average across participants, non-literal ones 14.5s. As for the video clip as a whole, literal subtitles were fixated for 14.1s on average across participants, non-literal ones 14.6s. Eye movements are indicative of visual and cognitive processing (Ghia, 2012a: 73) and fixation durations are a marker of the ease of accessing the meaning of a word and integrating this into the current sentence (Cop et al., 2015). Overall, participants attend more to contrastive input, which may suggest that integration of non-literal items in the flow of information may be more difficult than in the non-literal condition. However, t-test results show that both conditions attracted a similar amount of attention overall. Moreover, the GLMEM analysis (4.3.8.5) revealed no significant interactions between translation condition and fixation measures. It appears that subjects equally processed both literal and non-literal items, suggesting that literal translations can be as salient and attract as many looks as non-literal ones. If a literal word or structure is interesting to the viewer, for example because they had not encountered it before (lack of familiarity) or because they did not expect that translation in that context, it will receive an amount of looks comparable to its non-literal counterpart. This finding provides evidence that different learners do find different elements of language salient. As Seilhamer put it: "The lexical or grammatical constructions that strike one learner as particularly useful or meaningful (…) may not seem terribly relevant to another learner." (2010: 23). This is also evidenced in the questionnaire, where a wide variety of words and constructions were reported, each mentioned by one or two learners (see section 5.3.5). Since the majority of subjects also explicitly stated noticing a discrepancy between the audio and the subtitles, it is possible that, as they became aware of such discrepancy, the learners were in a state of alertness (Tomlin and Villa, 1994), such that they paid attention to and read with the same curiosity all subtitles, including the literal ones.

Another possibility is that people did read non-salient, literal items faster and more smoothly, i.e. made fewer regressions during first-pass reading, which, however, meant that they had extra time available to attend to the images and then go back to the subtitle before it disappeared (shifts). As previously explained, subtitles can be re-read if they are still on screen for some time after the viewer has processed them a first time. This is recognised in the AVT industry too, through the well-known 6-second rule, whereby a subtitle (of a standard, two-line length and an average adult reading speed of 180 WPM) is not allowed to stay on screen for more than six seconds, to avoid automatic superfluous reprocessing of the written text and let the viewer enjoy the images. It is possible that smoother reading of literal items led to a subtitle-image-subtitle eye movement pattern that resulted in a similar overall number and length of fixations.

One last factor might be proposed as an explanation for the eye-tracking findings detected for literal and non-literal transfer, namely a different sensitivity to orthographic regularity, which is "a highly abstract form of metalinguistic awareness" (Koda, 1996: 454). Shared L1-L2 orthographic knowledge has been found to enhance lower-level processing skills, which, in turn, facilitate the simultaneous deployment of multiple processing skills (Koda, 1996: 455), to influence moment-by-moment word identification independently from phonological knowledge (ibid.) and to accelerate L2 lexical processing efficiency (ibid.: 456). If processing efficiency is understood as smoother reading, involving fewer regressions and therefore fewer overall fixations, and if a different baseline sensitivity to such formal L1-L2 regularities truly exist, this might explain why L and N items receive roughly the same amount of looks. According to this interpretation, non-literal items would receive more looks in light of their being more unusual or unexpected renderings of the SL, but since subjects have a different sensitivity to orthographic similarities, not everyone is as efficient at processing the L2 literal text where such similarities most visibly occur, which resulted in enough people attending equally to literal and non-literal items, albeit for different reasons (sensitivity to L1-L2 orthographic similarity versus defiance of expectations).

Eye movements were also analysed in relation to recall scores to establish how visual attention relates to mnemonic accuracy. In the by-subject analysis (4.2.4.1 and 4.2.4.3), there is a significant correlation between fixation count and recall: the more a subject looks at an item, the more likely they are to recall it in the post-test. A similar situation applies to fixation duration, although the correlation results do not achieve statistical significance. In the by-item analysis (4.3.2.2 and 4.3.2.4), t-tests on both fixation count and duration revealed a significant difference between items correctly and incorrectly identified in the MCQ. Tables 15 and 16 show consistently higher fixation count and duration values for items that were correctly recalled in the post-test, and fig. 25 and 27 represent this difference graphically. Fixation measures were found to affect recall scores in the logistic regression analysis (GLMEM) as well. Specifically, total fixation count had a strong main effect on accuracy ($\beta = 0.189$, SE = 0.045, Wald Z = 4.197, $p < 0.00001$) even after the other significant predictors were added to the model. The sign of the coefficient confirms the positive relationship that emerged in the by-subject analysis: if more (and longer) fixations occur on a subtitle, its wording will tend to be accurately recalled. On the other hand, fig. 25 and 27 indicate that when less visual attention is dedicated to a subtitle, this tends to result in incorrect recall. In other words, it seems that when viewers do not recall an item correctly, it is because they have not looked at it long enough. This finding may reflect the higher skipping rate described in the literature (d'Ydewalle and de Bruycker, 2003; Pavakanun, 1992) for reverse compared to standard subtitles. The questionnaire confirms that subtitles in this translation mode can occasionally be skipped.[36] Interestingly, images and lack of speech appeared to be another reason why some words were missed. Two subjects reported that when there was a break in the dialogue, they would start concentrating more on what was happening on screen and found themselves to be slower at going back to reading the subtitles when they appeared, which meant they missed a couple of words in some of them. This finding

[36] A minority of subjects (3/26) reported on their experience of reading reverse subtitles, e.g.: "I felt that because the clip was in English I tended to not focus on the subtitles as much." (S_{10}) and "I didn't feel the need to read the subtitles so much; I did so more to check my understanding of Italian." (S_{13}).

is in line with the longer latencies found by d'Ydewalle and de Bruycker (2007) for reverse compared to standard subtitles. In other cases, low fixation numbers were registered in cases where viewers 'were stuck' on a word, i.e. made one long fixation on it and did not have enough time to read the rest of the subtitle before it disappeared. Naturally, when subjects do not read the whole subtitle (e.g. if they are concentrating on the images or they are stuck on one word) they have less chances to accurately remember its wording, and the above questionnaire remarks provided further insights on some of the reasons why such reduced subtitle processing occurred.

In conclusion, this analysis showed that (a) some translations can be given the same amount of visual processing and yet yield different recall scores; (b) non-literal renderings attracted more and longer looks than literal ones, as hypothesised, but the difference is only slight and non-literal items were recalled less accurately in the recognition post-test overall. This would suggest that during L2 subtitle reading, if a translation is unusual or interesting to a viewer, for example because it creates a departure from their expectations, processing effort increases and the translation might even be consciously noticed by the viewer at that moment, but this cannot guarantee its accurate recall at a later stage. Thus, the present analysis suggests that there is more to the relationship between language processing and memory than meets the eye, and fixation measures aptly revealed the complexity of this relationship in multimodal bilingual environments such as that of audiovisuals.

5.3.4 Frequency

A number of analyses were carried out on this variable. First of all, no difference was found in frequency by translation condition (4.3.7.1). The fact that frequency does not differ significantly between literal and non-literal items means that the effect of translation condition is not confounded by this variable: if, for example, literal subtitles were consistently less common than their non-literal counterparts, one could not be sure of whether the differential recall found for L and N were to be ascribed to translation manipulation or to the resulting rarity of the manipulated items. The frequency analysis thus confirms the validity of the study, by demonstrating that a non-literal item is no more likely to be rare than a literal one. That is to say, a mix of high- and low-frequency items was present in both literal and non-literal renderings. It is therefore possible to conclude with a degree of certainty that the differential effect on recall found for L and N cannot be reduced to a matter of different frequency of the test items and is at least in part due to the experimental manipulation of formal similarity. The same applies to frequency by linguistic category: no statistical difference was found between the variables. The fact that frequency does not change between linguistic categories indicates that lexical and syntactic items have equal chances of being high- or low-frequency and the experimental stimuli are balanced in this respect.

Secondly, no statistical correlation was found between frequency and recognition memory. That is to say, high-frequency items had no more (or less) chances of being accurately recalled than low-frequency items. In other words, although a range of frequencies (0.01 – 333 IPM) was present in the subtitles, these were not predictive of recall outcomes. As we have seen in 5.3.1, lexical items (nouns aside) in N produced varying recall accuracy scores, independently from their frequencies. Lexical items in L also display an independence from this variable, since all items (apart from AOI.5) were recalled accurately by the vast majority of learners (min. accuracy 70%) regardless of their varying frequencies. For example, both the low-frequency word *ciecamente* (AOI.96 - 1.193 IPM) and the high frequency word *crede* (AOI.94 - 333.07 IPM) were recalled by 85% of subjects. The same applied to syntactic items. Some non-literal low-frequency structures such as AOI.91[37] (1.7 IPM) were recalled correctly by 85% of subjects, while other non-literal low-frequency items such as AOI.51 (0.074 IPM) achieved a much lower accuracy of 38%. Literal structures quite closely mirror the results observed with literal words: all items were recalled accurately by all learners (min. accuracy 70%) regardless of their varying frequencies. For example, both AOI.45 (0.373 IPM) and AOI.85 (125.319 IPM) achieved a high recall accuracy (77%) despite the wide difference in frequency.

According to the word frequency effect paradox (Mandler et al., 1982) low-frequency words are better identified in word recognition tests, whereas high-frequency words are better remembered in traditional free-recall tests. Although no comment can be made here on the traditional recall test part of the paradox, since no such test was presented in this experiment[38], the finding regarding word recognition seems to be at odds with the

[37] For the list of all AOIs (i.e. the 22 test subtitles) in this study, complete with ST, literal TT and non-literal TT, see table 5 in section 3.13.

[38] Free-recall as intended in the frequency effect paradox is fundamentally different from free-recall as intended in this study. The former refers to *elicited* recall and includes tests as fill-in-the-blanks, where learners are forced to produce from memory a set of words predetermined by the exercise. The latter, on the other hand, refers to unprompted, *spontaneous* recall, where learners have full decisional power on what items to produce from memory, and can even, therefore, not report any words or expressions at all.

results presented above, where low-frequency words are not consistently better identified in the MCQ. Lohnas and Kahana (2013), who more recently addressed the paradox, note that these results are valid only for pure lists of either low- or high-frequency items. The alleged superiority of low-frequency words is not consistent in mixed lists (when both classes of items are presented within the same list), with studies reporting diametrically opposite results, i.e. advantages of low-frequency words, high-frequency words, or no clear advantage. The MCQ presented to the subjects of the present study contained questions on all test items and was therefore also a mixed list, containing both low- and high-frequency test items, which may explain why lower-frequency words did not show a clear recognition superiority. Furthermore, these previous experiments looked at frequency effects in L1 (Lohnas and Kahana, 2013; Hicks et al. 2005; Watkins et al., 2000), whereas this study is concerned with L2 input. The situation is likely to become more complex when a foreign, non-dominant language is to be processed – especially if L1 and L2 occur simultaneously – and this will affect the way in which viewers process and remember information. Moreover, the studies above, like many others, looked at word lists containing words presented in isolation (Lohnas and Kahana, 2013; Hicks et al. 2005), whereas reading reverse subtitles in this experiment involved a rich contextual environment, which could have easily facilitated the viewers in retaining and recognising the items despite their high frequency. In fact, context-related factors such as contextual diversity (i.e. the number of repetitions of a word in different environmental contexts) have been found to play a role in recall (Verkoeijen et al., 2004) and to override frequency in predicting word processing times (Adelman et al., 2006). Another crucial point to add is that, surprisingly, studies involving frequency do not always explicitly state what 'high' and 'low' frequency are taken to mean. An exception is a study by Scarborough and colleagues (1977) on frequency and repetition effects in lexical memory, who were working with a 1-million word corpus (Kučera and Francis, 1967). They defined low frequency as less than 28 occurrences per million, the cut-off point resulting from the median split of frequencies in their experimental items. Where some indication of how low and high frequency are calculated is given, researchers often divide their stimuli into a number of bins, and define as low frequency all words in the lowest bin and high frequency those in the highest bin. Lohnas and Kahana (2013) note that one of the reasons for the contrasting results described above might be precisely the different definitions of low and high word frequency: different definitions lead to aggregation into different bins which, in turn, could easily lead to noting a low-, high- or no frequency advantage (ibid.: 1945). Moreover, such categories are by definition relative to the language sample of the study in question. Therefore, Lohnas and Kahana recommend keeping frequency as a numerical variable when analysing how frequency interacts with mnemonic performance (ibid.: 1946). For these reasons, numerical IPM values were used in all statistical analyses involving frequency in the present study. Furthermore, the range of the test items was clearly stated and a scatterplot of the frequency data (fig. 48) is reported in section 4.3.7.1, so the reader can directly see their distribution. The cut-off point was set as the average frequency (38.69 IPM), so that items below this threshold count as low-frequency, whereas items above it as high-frequency. Since the frequency of the items in this study is overall relatively low (30 items fall below the cut-off point, 12 above), I did not use the median (7 IPM) as this would have skewed the perception of frequency by counting any item above 7 IPM as common, when in fact such occurrence rate is rather rare, if one considers the ranking of words with this frequency in the corpus.

To summarise, in the present experiment words that are frequent (e.g. *diverso, ragazzi*) as well as very infrequent (e.g. *guerriglia*, *vaticinio, profezia*) were both remembered correctly by the majority of subjects, and the same applied to syntactic items. Thus, from these data, it seems apparent that frequency cannot be deemed the only variable that influences item recognition in the context of exposure to subtitled video material, at least in cases when other subtitle-specific variables, such as L2 translation, are also experimentally manipulated.

Thirdly, no correlation was found between frequency and fixation number or duration. This third finding is interesting since frequency should correlate negatively with both eye movement metrics. According to the well-known Word Frequency Effect (see 2.11), frequent words are processed faster than infrequent ones (van Heuven et al., 2014). Reading research has shown that, once word length is kept constant, frequent words are fixated less (or skipped altogether, if they are function words such as pronouns, conjunctions, exclamations and the like), while infrequent, uncommon words are fixated for longer (Rayner et al., 1989; Pollatsek et al., 2008). Thus, a high-frequency word might be fixated only once, while more difficult, less frequent words might be fixated two or even three times (Cop et al., 2015: 15). The fact that frequency and eye metrics do not significantly correlate could be due to the translation manipulation itself: enhancing the input by creating two different conditions (L and N) could have triggered a heightened state of altertness and a focused orientation of attention (and therefore eye movements) on specific words motivated by variables other than frequency alone, for example salience, context or lack of familiarity. This may be particularly plausible given that the majority of subjects (61.5%) in the questionnaire explicitly stated noticing a discrepancy between the audio and the subtitles, which would suggest that salience did make the input visible and had an effect on processing. Moreover, those who did not openly mention a discrepancy still reported (free-recalled) several words that they found interesting. These observations suggest that, in audiovisual contexts, viewers recall items based on more than just frequency information. Furthermore, the solid word frequency effect reported in the literature applies mainly to silent reading research, which does not include a source audio. Although subtitle reading is also silent, the simultaneous occurrence of source and target text in subtitled video is likely to affect such reading, *in primis* allowing the subject to (sub)consciously compare the two codes. Drawing such comparisons may involve a change in attention allocation and processing effort, resulting in frequency not having the same effect recorded in more 'traditional' silent reading studies. Another possible explanation arises from experimental design. Since this experiment is not concerned with words in isolation but with the audiovisual text as a whole, it is possible that other sentential and textual variables may also have an effect on the viewer. Indeed, word processing in natural reading is influenced by sentence parafoveal stimuli as well as context (Cop et al., 2015: 2), which might have dimmed the visibility of existing frequency effects. In fact, since certain words occur in certain collocational contexts more than others, a word classified as low-frequency in a frequency list can *de facto* have a higher occurrence when it is found in a specific context (Moran, 2012: 213). Frequency is not independent from its immediate surroundings either: as we have seen at the end of 2.6.2, fixations on a noun are affected by both the length and frequency of the adjective preceding it, which cause spill-over effects (Pollatsek et al., 2008). Variables such as parafoveal preview, spill-over effects of frequency and other reading mechanisms were not analysed in this study, since the Tobii X120 hardware cannot support what would effectively be a reading study design. A much higher eye tracker frequency is needed, typically 1000Hz/s or more, to look at these word-level measures and accurately capture miniature movements as tremors or micro-saccades. Moreover, it was never an intention of this research to investigate said miniature movements. One of the main goals was to investigate various aspects of the reception of reverse subtitles while being able to provide a direct comparison to Ghia's (2012a and 2012b) parallel investigation on standard subtitles. Specifically, the study aimed at assessing attention allocation between subtitles in order to compare different translation conditions. Looking at fixation metrics on the subtitle area is therefore appropriate to this goal as well as consistent with the research antecedent. It seems that frequency effects need more fine-grained methods of analysis (higher sampling frequency, more local measures at the word-level) in order to be detected. And, indeed, recent literature on the subject has highlighted the need for such "fine-grained linguistic data analysis in work examining the effects of frequency" (Divjak, 2016: 22).

Finally, the fact that fixations were collected on each subtitle as a whole but the frequency was calculated on its content words might have created a mismatch between how common a lexical item is and the looks it actually receives. Recall from 3.13 that, with lexical items, only one element in the sentence is changed between translation conditions, yet eye data is collected on the whole subtitle. However, because the subtitles in the two conditions will be identical apart from this one single element, a fixation difference in conditions should still emerge if the manipulated element is more striking in one of the two conditions. Moreover, this procedure works well for syntax, where the structure of the sentence has been changed and therefore the interest is in collecting eye data on the whole subtitle. This type of eye movement analysis at the subtitle-level is, however, currently inevitable, since eye-tracking technology does not allow one to analyse dynamic text (such as subtitles) with the aid of reading statistics in the same way that it can be done with most eye-tracking software during reading of static texts (Kruger and Steyn, 2014: 106). For example, AOIs can be created automatically by some eye-tracking software on each word if the stimulus is a static text, but need to be created manually by the researcher on subtitled video (see also 3.17.1). This is why, to date, most subtitle processing studies are limited to the investigation of amount of attention to the subtitles (rather than to individual words) and cannot entertain fine-grained analyses of reading processes during the consumption of audiovisual material. This point is of particular interest and will therefore be further addressed in section 5.4.

To conclude, this research does not deny a role for frequency in characterising language. As shown in 2.11, I agree with the tenet that frequency plays a major role in language processing, comprehension and learning[39]. However, the results outlined above suggest that frequency may not be telling the whole story: when language is presented audiovisually, viewers' subsequent recall depends on more than just frequency information. Although it can undoubtedly assist the learner in processing and learning

[39] For an overview of and a convincing argumentation on the implications of frequency effects, see Ellis (2002a and 2002b).

language, frequency cannot ensure that a word will be remembered: sometimes one single exposure is enough for a word to be remembered forever (Hulstijn, 2002: 271) whereas sometimes words encountered multiple times are still not acquired (Seilhamer, 2010: 23). Attentional processes can modulate frequency effects (Ellis, 2002a: 179), so more in-depth, controlled research is needed on attention allocation in audiovisual processing to reveal and separate the individual contribution of input-specific variables such as frequency and sentence context from learner-specific ones such as familiarity or individual differences in perception.

5.3.5 Noticing

The construct of noticing (reviewed in 2.8) was addressed through the analysis of orientation, detection and metalinguistic awareness (overt noticing), sub-components that, taken together, helped shed light on how processing the L2 while watching audiovisuals can influence recognition (post-test, see Appendix A.3) and free-recall (questionnaire, see Appendix A.4).

Eye movements allow us to investigate the allocation of attention in the context of noticing. In particular, fixation measures can be taken as a proxy for the attentional function of orientation (see 2.8 and 2.10). Orienting attention means "committing attentional resources to sensory stimuli" (Tomlin and Villa, 1994: 190). In this study, the stimuli viewers fixate on are words and sentences during subtitle reading. However, orientation to input as measured by eye-tracking does not automatically coincide with input detection, i.e. its cognitive registration in the memory of the viewer. If fixations are recorded on a subtitle, all one can be sure about is that attention was oriented to that subtitle. This does not *per se* inform about the mental processes that the learner was engaged with during orientation. It is possible that language attended to (fixated) is not cognitively registered at all and is therefore not "made accessible to whatever the key processes are for learning, such as hypothesis formation and testing" (Tomlin and Villa, 1994: 193), resulting in incorrect identification in the post-test. It is also possible that detection does occur during processing, but the memory trace is too weak and also results in incorrect identification. As Ghia puts it: "noticing might still have occurred, but might not have been sufficient for the storage of information in memory." (2012a: 89). However, if a learner attended to a subtitle and then recalled it correctly in the post-test, some form of further elaboration must have occurred in order to enable correct identification. That subtitle must have been detected by the learner, i.e. cognitively registered in memory, and submitted to further elaboration processes which allowed correct recognition in the post-test. Since we cannot be in the viewers' head as they process the subtitled input, however, we cannot be sure of what type of detection occurred, i.e. whether they consciously registered features of this input (detection within selective attention) or whether they registered it unconsciously (detection without awareness). The questionnaire[40] was particularly informative in this sense, providing evidence that both types of detection can occur and lead to correct post-test recognition. While some participants provided clear motivations for reporting some words, others were not as sure. For instance, on the one hand, S_{14} stated: "There were quite a lot of words or expressions that I noticed when reading the subtitles *as I wouldn't have thought to have said it or because I hadn't heard of the expression and wanted to remember it*. For example 'indovinato'; (…) 'l'Eletto' (…); 'ma non sei sveglio' " [my italics]. On the other hand, S_7 stated: "*For whatever reason*, I instantly recognised the items 'resistenza' and 'profizia' when they appeared in the multiple choice questions" [my italics]. Both S_{14} and S_7 correctly identified the reported wordings in the post-test, but while S_{14} overtly mentions having noticed them during reading (detection within selective attention), S_7 only reports recognising them during the post-test, and not being able to explain why, as he does not remember noticing them at the time of watching (detection without awareness). S_7 does not report consciously noticing *resistenza* and *profezia* at the time of watching, but we can be sure that he cognitively registered these items, with evidence coming from his eye movements (which confirm orientation to these words), his mnemonic performance (correct recall in the MCQ) and his metalinguistic comment above (questionnaire). This finding provides evidence towards Tomlin and Villa's distinction between detection with and without awareness. The latter can occur during online processing, and can result in correct L2 recognition.

Most participants (70% of those who spontaneously mention words or structures in the questionnaire), however, do remember being struck by them during watching. Not only did they report the words that they consciously noticed and found worth mentioning,

---

40 A total of 33 participants took part in the experiment but only 26 were included in all quantitative analyses for the reasons explained in 3.8. However, these seven subjects went through the whole experimental procedure (3.16) like everyone else and their questionnaire comments can still provide useful information on their thought processes. With regard to the qualitative analysis presented in this and the next section, therefore, although subjects S_{27}-S_{33} are not counting towards numerical descriptions of the questionnaire dataset, their comments are reported if relevant, i.e. wherever they make a meaningful addition to the discussion.

but also added a variety of explanations as to why these items registered with them, evidencing an unexpected level of metalinguistic awareness. Said metalinguistic data show that subjects report subtitles that they find interesting, because:

(1) The words are unfamiliar: "I hadn't heard the word Eletto before (S_{21}); "I found (…) words striking because I hadn't encountered them before" (S_5); "Vaticinio (prophecy) was a new word for me" (S_{22}).

(2) Expectations are not matched: "They [the subtitles] weren't always as close to the original as I was expecting them to be, or as I would have made them if I had done the subtitles myself." (S_4); "I found them to be different from the ones I would have written – this made it easier for me to recall the words as they piqued my interest and stuck in my mind more" (S_{11}); "I didn't realise you could use this word in such a way" (S_{10}); "There were some that interested me a lot because I wouldn't have expected the translation." (S_{25}).

(3) The subtitles contain idioms and colloquialisms: "They did not always reflect idioms in the spoken English" (S_{20}); "there were interesting translations with regard to the colloquial English phrases" (S_{18}).

(4) They would have struggled to translate the expression themselves: "The exercise (…) helped me to understand how certain expressions and idioms translate into Italian. For example 'bingo!' became 'Indovinato!' which I might have struggled in translating." (S_9); "Bingo - indovinato - I would not have known how to translate this." (S_6).

This array of explanations in the questionnaire confirms that subjects do indeed find different aspects of language striking, and for different reasons. Items reported as striking came from throughout the clip, not just from among the test items. Items containing idioms were also noticed. One subject even stated: "I feel like only certain particular words stuck in my mind, for example ones I hadn't heard before or *that sounded nice*" (S_{28}) [my italics]. Apparently, pleasant sounding words can also leave a deeper memory trace in the learner. Lack of familiarity was confirmed to play a major role, with participants mentioning it as a reason for both noticing[41] and remembering[42] certain words and

[41] See for example, the following comment: "the subtitles which contained vocabulary I wasn't familiar with (...) *I was more aware of reading them*" (S_5) [my italics].

[42] See: " 'Indovinato' *sticks in my mind*, because it is not a word I am familiar with in that context" (S_{13}) [my italics].

expressions (also see 5.3.6). Although frequency is often taken as a proxy for familiarity (where low frequency words are more likely to be unfamiliar to the learner), the questionnaire results confirm frequency and familiarity can be separate sources of influence on the salience of a particular form in the learner (Giora, 1999), since items consciously noticed and reported as unfamiliar in the questionnaire were not always rare (i.e. they do not always have a low corpus frequency) and vice-versa, some rare items were not overtly reported as salient by any subject.

Moreover, the fact that some learners made explicit reference to their own translation skills indicates that they used the subtitles to check for translation solutions (Ghia, 2007) and they did notice the gap between their own abilities and the language encountered during processing. According to Swain's Output Hypothesis (1985) noticing this gap between their resources and a linguistic problem they need to solve encourages learners to look for the adequate knowledge they need in order to fill that gap (i.e. producing L2 output), which in turn fosters L2 language progress. From this point of view, then, the use of reversely subtitled audiovisual texts, by triggering simultaneous processing of both source and target and yielding such noticing events in the learners, may be suggested to facilitate learning (5.3.6) in that it encourages development in the command of the foreign language.

Learners' meta-reflections in the questionnaire also confirmed expectations played a crucial role and triggered noticing of formal features of the L2. Aside from the AOI.5 example (*differente – diverso*) already addressed in section 5.3.1 of this chapter, two instances are presented below, as they show the kind of thought processes arising during processing.

(a) "I thought 'persona in gamba' was a strange way to translate 'good person'. To me, 'in gamba' suggests capable, good at what one does, rather than nice or honest, things I associate with being a good person." (S_4).

(b) "the Oracle says Neo has 'un buon animo' and I thought that soul was a feminine noun in Italian, 'anima', so should this not be 'una buon'anima'?" (S_{33}).

I have reported these in full because they exemplify how expectations about the L2 assist learners in identifying what to them constitutes a linguistic problem, while shaping their hypothesis formation and testing. The non-literal expression *in gamba* [clever, smart] has a multitude of nuances and can also refer to a good, great or valuable person (Spadiliero, 2015). The word *animo/a* can be used as masculine or feminine (other examples are *tavolo/a* [table] and *mattino/a* [morning]), therefore the literal noun phrase *un buon animo* [a good soul] is a perfectly legal rendering. In both cases, albeit for different reasons, learners are lacking a thorough understanding of the Italian phrases, and yet they make full use of what knowledge they have while processing the foreign language. They have clear expectations and these play a role from the very first encounter with the language exemplars in question, resulting in the items subsequently being available for free-recall and metalinguistic analysis in the questionnaire. This type of elaborative rehearsal involves making connections within the material processed, or connections between said material and previous learner knowledge, which has been showed to increase memory performance (Reisberg, 2013). This finding is compatible with dominant constraint-based theories of online language processing and comprehension, whereby "comprehenders use all salient and reliable sources of information, as soon as possible, to guide their interpretation of an incoming linguistic signal" (Farmer et al. 2012: 354).

The distinction made by Truscott (1998) between overall awareness versus specific instances of input is also confirmed by this study. The questionnaire reveals that 61.5% of subjects (n = 16) explicitly state noticing an overall discrepancy between ST and TT. This percentage rises to 81% of the total (n = 21) if we include instances in which more specific discrepancies between ST and TT were referred to directly or indirectly. In the former category, subjects would report comments like "I have an overall impression that there were some odd translations but I don't remember specific examples" (S_{24}), "they [the subtitles] weren't always exact translations" (S_2), "[t]he subtitles didn't often say exactly what the English said" (S_{17}), therefore showing an *overall* metalinguistic awareness of the ST-TT discrepancy at the broad clip level. The latter category, on the other hand, also counts subjects' comments that were more circumscribed, yet directly or indirectly indicate that a discrepancy was noticed. For instance, they include comments like "I did notice that the *word order* had been altered *in some of the Italian subtitles*." (S_{12}), "(…) I noticed *a grammatical use* that I would not have immediately used (…)" (S_{18}), or "*indovinato* as a translation of 'bingo' was interesting - seemed more formal than the original" (S_{22}) [my italics in all three examples]. Here, though subjects are not making overall conscious statements about the *whole* clip, they still notice and report individual instances of discrepancy in the input. Overall, subjects reported noticing "strange word choices" (S_3), "unusual language" (S_{20}), "some odd translations" (S_{24}), "different syntax" (S_6), "altered word order" (S_{12}). Others stated that "[the subtitles] sometimes used words I wouldn't have known out of context or more advanced/sophisticated that I would have chosen" (S_{15}), and "the translations were pretty inventive" (S_{23}). Several comments of this type were found, confirming a general awareness of the subtitled text and suggesting that 'literalness' has a psychological reality in the mind of the learners, who, when a manipulation occurs, acknowledge the audiovisual text is not uniformly translated. Subjects do draw comparisons between ST-TT, and literal and non-literal translations are indeed perceived and remembered differently during the consumption of AVT material.

To summarise, the findings lend support to the view that orienting attention to and detecting input can happen without awareness, which can still result in recognition, but does not result in free-recall in this dataset (for more details on free-recall, see the next section 5.3.6). It is conscious noticing in Schmidt's acceptation (or detection within selective attention in Tomlin and Villa's terminology) that leads to the increased processing and deeper memory traces which in turn result in unprompted free-recall and learning of new forms. Another conclusion that can be made is that, for conscious detection to occur *and* result in free-recall, an item must be salient *to the learner*. As the variety of subtitles reported as striking demonstrates, what emerges as salient changes considerably from learner to learner and is strongly linked to familiarity and expectations. As we have seen, manipulating translation through literal and non-literal transfer can indeed have an impact on the viewer, affecting both what they notice and what they recall, especially when the translation includes new language items that they are not familiar with or defies their expectations. Finally, the findings of this study reveal that learner-initiated noticing (Godfroid et al., 2010) may be more pervasive an activity than what might be believed. Learners have clear hypotheses that they immediately put to the test during reception of L2 material, and it is the linguistic knowledge they already possess that determines what they will expect to see in the input.

5.3.6 Foreign Language Learning

In this study, Foreign Language Learning (FLL) was operationalised as mnemonic performance in the two post-tests, namely the MCQ (3.12) and the open-ended questionnaire (3.15), which were designed to measure receptive (verbatim recognition) and productive (free-recall) memory respectively. While the MCQ results were discussed in detail in 5.3.1, some relevant findings also emerged from the questionnaire. As far as free-recall and translation condition are concerned, a clear pattern was found. First of all, all subjects in this dataset but two (92% of the total) reported at least one Italian L2 word that they found striking, with some subjects reporting up to seven different items that piqued their interest during watching. Many of these items were previously unknown to the subjects, suggesting that conscious noticing plays a crucial role not only in memorisation but also in learning. The finding also indicates that mnemonic retention after one single exposure to reverse subtitles in upper-intermediate L2 students is considerably higher compared to that registered by exposing intermediate students to standard subtitles as reported in Ghia's study (2012a).

Furthermore, a total of 29 distinct items were spontaneously recalled in the open questionnaire. Of these 29 items, roughly 60% were lexical (n = 17) and 40% were syntactic (n = 12) in nature, suggesting that the anchoring function of lexical items postulated by Smith (2004) for processing may also hold for productive memory (see 5.3.1). The fact that almost half of the free-recalled items is syntactic is nevertheless striking, and is in line with the high recall accuracy for syntax recorded in recognition (see 5.3.2). Another interesting and somewhat surprising finding emerging from questionnaire analysis is that 28% (n = 8) of these 29 items reported were literal, while 72% (n = 21) were non-literal. It therefore appears that literal transfer produces higher *recognition* of words post-exposure, but results in lower spontaneous *free-recall*. When it comes to the latter, it is non-literal words that are more frequently retrieved from memory. This finding could be explained by proposing that formal similarity (literal transfer) acts at an unconscious, subliminal level that, while allowing for correct post-test subtitle identification, leaves a comparatively weaker memory trace in the learner, which does not consistently result in free-recall. On the other hand, formal source-target discrepancy (non-literal transfer), does not achieve as high levels of post-test recognition accuracy but is more likely to be consciously detected during online moment-by-moment processing and results in a higher degree of spontaneous recall. As we have seen earlier in this chapter, non-literal renderings can cause confusion at the level of recognition, so that, when presented with the MCQ, subjects are overall less certain of what they saw and heard (perceptual confusion) and end up choosing the wrong answer. Such perceptual confusion is overtly referred to in a subject's comment: "some phrases in the subtitles particularly registered with me, but with some phrases I only remember hearing the English voice." (S_{29}). Nevertheless, questionnaire data shows that, when non-literal renderings are salient to the learner and therefore produce the kind of elaborative rehearsal described in (a) and (b) in the previous section (5.3.5), they are more consistently free-recalled in the questionnaire. This is not to say that literal subtitles can never be free-recalled, however. Going back to the example in (b) in 5.3.5, S_{33} read *un buon animo* [a good soul] (L) in the subtitles during watching and made a comment on this L version in the questionnaire because it did not match his frame of interpretation. On the other hand, another subject read *un buon cuore* [a good heart] (N) during watching, noticed the discrepancy in this N version and reported that " 'cuore' was used instead of 'anima' for soul" (S_{16}). The parallel comments on both versions of this item thus suggest that both translation conditions can in principle trigger noticing and elaborative rehearsal and, if expectations on literal items (L) are left unmatched (such as for S_{33}), these L items can also occasionally be reported in free-recall.

Through the questionnaire, some evidence of novel word learning also emerged. When reporting one or more striking words, 69% of subjects (n = 18) also reported that they did not know these words before. For example, some subjects stated: "Vaticinio was a new word" (S_{32}); "One of the main words that I hadn't come across was 'l'eletto' for 'the one' " (S_{14}); "I deduced that 'ciecamente' meant blindly, though I had not heard the word before" (S_{29}); "I didn't know the word for resistance (guerriglia?)" (S_{15}); "I hadn't seen the word 'ci siamo' for 'we're in' " (S_{14}), and so on. The fact that so many subjects reported new, unfamiliar words in full demonstrates that some novel word learning can occur after one single exposure to reverse subtitles. When free-recalling striking items, four subjects (namely S_1, S_9, S_{17}, S_{30}) also acknowledged that there were more items they noticed during watching than they are able to free-recall. One subject stated: "I did find some other words striking but I don't remember them now" (S_{17}). Another felt like adding that, although he reported only one instance of striking words, "there were a few others, I'm pretty sure of that" (S_1). Thus, it emerges that conscious noticing can indeed occur during the first encounter with a foreign word or structure, and even if noticing an item is not always sufficient for that item to be subsequently free-recalled, such encounter can nevertheless leave a memory trace of the noticing act itself. This finding therefore supports the view that language learning is an incremental process where noticing is the initial step (Schmidt, 2001; Hulstijn, 2001), and memory plays a cardinal role in this process. Often one exposure is not sufficient for the storage of information in long-term memory, even when presented through the rich, multisemiotic audiovisual environment, and even when conscious noticing (detection within selective attention) did occur during online processing. Multiple exposures are often necessary to increase mappings and allow information to be more permanently stored in memory. However, one exposure is sufficient for learners to notice striking features of the input and start integrating new information in their current language systems. Recall that subjects S_5 and S_{14} (note 34 earlier in this chapter) reported the first few letters of the items *vaticinio* and *guerriglia*. These words occur so rarely in the Italian language that they are highly unlikely to be already known by the learners, which is confirmed by learners' overt statements of unfamiliarity towards them. Albeit unable to recall the form of new words in their entirety, these subjects could indeed recall their meaning and some features of their surface form (in this case, the beginning of the words) after a single exposure, thus deepening their L2 knowledge of those words to some degree. Such deepening of word knowledge was recorded in 26% of subjects (n = 7). Instances counted were not only those in which subjects recalled part of words but also those in which subjects provided evidence of having enriched their knowledge of a word they had already encountered, either in a different context or with a different meaning. Take, for instance, the following comment: " 'Indovinato' sticks in my mind, because it is not a word I am familiar with *in that context*, I would have expected something like 'esatto' or 'bingo' " (S_{13}) [my italics]. Here S_{13} does not report the verb *indovinare* [to guess] as a completely new word, but does note that he had not encountered it used in this context[43] before, i.e. as a non-literal translation for the exclamation *Bingo!* in the source audio. Another case-in-point is S_{10}'s comment: "I noticed after having studied Calvino, 'il sentiero' was used - I didn't realise *you could use the word in such a way*" [my italics]. This subject is referring to *Il Sentiero dei Nidi di Ragno* [The Path to the Nest of Spiders] (Calvino, 1947), a neorealist Italian novel evidently studied by the student in his Italian literature class, where the word is used in its literal meaning, i.e. indicating primarily a physical path in the forest. In the present clip, however, the word is used in subtitle 25 (AOI.23) in the literal translation *Ti può aiutare a trovare il sentiero.* [She can help you to find the path]. The sentence is uttered by Morpheus, as he is trying to convince Neo that meeting the Oracle will help him to find an answer for his questions, and refers to a purely metaphorical rather than a physical path. Here, S_{10} made a clear connection between his previous (imperfect) L2 knowledge and the new L2 use he encountered, thus integrating and enriching his knowledge of the word *sentiero* by adding a metaphorical meaning to it. In both these examples, students must have noticed the gap between their own Italian knowledge and a potential translation issue, which triggered elaborative rehearsal (see 5.3.5), i.e. the creation of connections within and outside of the languages processed, and resulted in the items *indovinato* (N) and *sentiero* (L) being reported as striking in the questionnaire. These results – regarding both novel word learning and deepening of word knowledge – are overall quite striking, if one considers that the learners were exposed to the subtitled clip only once, and had no chances of rewinding or pausing it at any point.

---

43 *Indovinato*! [(Well) guessed!] can be considered an Italian secondary (i.e. lexical) interjection, here used as an exclamation, with a pragmatic function between inspiriting and approving according to De Santis's (2010) categorisation.

Last, two methodological remarks can be made with regard to learning in the context of this study. First of all, there is a difference between the design of this and other studies investigating the relationship between noticing and learning or language development. In several noticing studies, learners are presented with bold or underlined elements or are asked to underline words themselves during reading or other pedagogical activities (e.g. Fotos, 1993; Izumi et al., 1999; Izumi and Bigelow, 2000; Uggen, 2012). This design *de facto* attempts to draw learner attention on certain pre-determined features of the FL (e.g. English past and present hypothetical conditionals in Uggen, 2012). The present study, on the other hand, does not enhance the input overtly (via underlining, boldface or key-word subtitling) but covertly, by applying translation strategies to each subtitle as a whole. This means that learners are 'free' to attend (or fail to do so) to whatever L2 features really interest them during watching (and, as emerged from the data, they tend to pay attention to both L and N items in roughly equal measure).

Secondly, as mentioned earlier in this section, the questionnaire used in this study resulted in a high number of distinct items being produced spontaneously from memory (n = 29), which was surprising given the initial expectations based on Ghia's (2012a) results. I ascribe this positive outcome first of all to the format of the questions, which were left open (see Appendix A.4). Rather than elicit *specific* forms, e.g. in typical fill-the-gap vocabulary post-tests, asking general questions about whether anything had been noticed and whether any unfamiliar and/or striking words occurred allowed the subjects to reflect on what *they* found striking and allowed us to really appreciate the variety of items that can be reported, many of which were not amongst the 22 test items. Secondly, increasing the length (three subtitle-related questions rather than two) and specificity of the questionnaire (for details see 3.15) is also likely to have played a role, giving learners more chances to retrieve L2 items from memory. Moreover, the fact that learners could answer the open questions as they pleased but had to provide an answer (i.e. to type something, however short, in the answer area on the PC screen) before they could move to the next question may have also prompted them to produce more Italian L2 items (Ghia does not provide information on whether this was the case in her questionnaire design). Finally, based on the body of comparative literature reviewed in 2.3, which established the mnemonic superiority of reverse over standard subtitles, it also seems likely that the very subtitling mode herein addressed (where the L2 occurs in writing rather than orally) also played a role in enhancing learner L2 memory. To conclude, the findings of this and the previous section collectively confirm that the questionnaire structure described in 3.15 – compared to the one used in Ghia's (2012a) parallel investigation on standard subtitles –

was successful in increasing the number and informativeness of the participants' answers, which were unexpectedly enlightening and were indeed able to "provide a far greater 'richness' than fully quantitative data." (Dornyei, 2003: 36), in particular as far as FLL was concerned.

5.4 Limitations, Methodological Improvements and Recommendations

Like all experimental research, this study also had its limitations, which I herein discuss together with some proposed methodological improvements and suggestions for future analyses. First of all, confidence in the generalisability of the findings could be increased by increasing the sample size. In doing so, following recruitment principles similar to those described in 3.8 would open participation to *any* subject that appropriately fits the desired criteria, rather than circumscribed to university students. By limiting recruitment to Languages or Psychology sophomores, the experimental sample is less representative and *de facto* undermines external validity, a criticism commonly moved to experimental psychology (Goodwin, 2008). However, given the difficulties in sampling narrow populations such as those required for this study (native English speakers, aged 18-35, living in the UK, with an advanced command of L2 Italian), a balance needs to be struck between ideal conditions and feasibility. Moreover, there is no set-in-stone rule as to how many subjects are 'enough' in experimental studies, and recommendations differ widely (Larson-Hall, 2010: 185). Deciding on a sample size will depend on a number of factors, e.g. the size of the effect one is expecting to find, the types of tests to be used (t-test, correlations, regressions, etc.), the α level (typically 0.05 but it can be set at lower values such as 0.01 or 0.001) and the number of explanatory variables and levels within them. For these reasons, sample size requirement will vary from experiment to experiment and from field to field. Given that SLA research typically has small sample sizes (ibid.: 323), with many researchers in this field publishing studies with 10-15 participants per group (ibid.: 44), I believe the sample size of this study (n = 26, therefore 13 per group) warrants a satisfactory degree of confidence in the results and is correctly aligned to the standard practice. Not only the sample size of the present study is quite common, but it is also above average for the field of AVT, where much of the research published in subtitling tends to have even smaller sample sizes. For example, in her noticing study, Ghia (2012a) collected data from 13 participants in total (6-7 per group), d'Ydewalle and De Bruycker (2007) from 20 (12 adults and 8 children) and Danan (1992) from 15 (7-8 per group) in her third experiment (the most comparable one to the present study in terms of participant base, since it was carried out on more advanced, post-beginner students).

Secondly, the experimental stimulus for the present study was a clip from the seminal American film *The Matrix* (1999), which some subjects had watched (n = 17), while some had not (n = 9). Therefore, roughly 65% participants in this study were familiar with the content of the film. This type of familiarity with the input could be easily avoided in the future by using non-mainstream English language video material, which is much more likely to be unknown to the viewer. Aside from linguistic content considerations, part of the rationale for this choice (3.7) was also the availability of an original script of the dialogues. While using more obscure materials allows to rule out potential familiarity effects on the viewers, it may easily require a longer preparation time for the researcher, who should be aware of this when approaching possible alternative video sources.

Thirdly, I have to recognise that, like in every experimental study, controlling for *all* extraneous variables (see 2.12) is rarely possible. For example, this research did not control for subject vocabulary size directly, which may have an impact and has been included, albeit rarely, in research on AVT (e.g. in Montero Perez et al., 2014). It could be argued that L2 vocabulary size will affect processing and memory since the larger the pool of words known by the learner, the more likely they are to know the words occurring in the subtitles and, by consequence, the faster or more efficiently they may process the L2 subtitles, which in turn may affect retention. However, vocabulary size amongst the participants in this study can be deemed homogeneous at least to a certain extent, given that participants all had to pass a language test (3.9), which ensured they had a B2 proficiency (CEFR level). Moreover, the language test comprised not only exercises on grammatical competence (where B2 graded vocabulary appears in the sentences used), but also a reading comprehension component, where a text with appropriate B2 vocabulary was broken into sentences that had to be re-ordered. Controlling for age (3.8) also contributed to homogenising the vocabulary size, since knowledge of words increases with age and education: older populations have encountered more words than younger ones, and therefore they also usually know more words (Brysbaert et al., 2016). In the future, a replication study could nevertheless include a specific vocabulary size test. Designing, creating and administering a vocabulary size test is not without its difficulties, however, and future studies may benefit from careful consideration of what to count as 'words' (e.g. what counts as base words and what as derived words, how to deal with polysemy or compounding), how to choose test words from a pool of available items (e.g. space sampling procedures to select base words or to account for frequency) and what test measurement options are available (see Goulden et al., 1990 for an extensive example in the monolingual domain, and Nation, 2001 for an exhaustive introduction to foreign vocabulary learning).

Another variable that was not controlled in this study is neighbourhood size (NS). A word's neighbourhood size or density is the number of words that are very similar to it (Marian and Blumenfeld, 2006) in a critical shared linguistic feature, typically orthographic, phonological, or semantic (Marian et al., 2012). The most relevant type for the present study, orthographic NS, is usually obtained by calculating the 'N metric', i.e. "the number of words that can be created by substituting one letter of the target word" (ibid.: online). Neighbourhood size can have an impact on a variety of language tasks – such as picture naming, picture-word interference and naturally-produced speech, as well as processes – such as word learning, word recognition and word production (Vitevitch, 2012; Marian and Blumenfeld, 2006). It also influences lexical access and has implications for lexical activation (Marian and Blumenfeld, 2006: 3; Marian et al., 2012). Because I start from the assumption that CLI (Cross-Language Influence) and formal similarity are real and have a measurable effect on FLL (2.5), and NS embodies aspects of such similarity, including a measure of NS for the target words could enhance internal validity (2.12) in future experiments. To this purpose, the CLEARPOND database could be used, a resource which allows to "obtain phonological and orthographic neighbors, neighborhood densities, mean neighborhood frequencies, word lengths by number of phonemes and graphemes, and spoken-word frequencies." (Marian et al., 2012: online). Interestingly, to capture spoken word frequencies, CLEARPOND is based on film and television subtitle corpora, which brings me to the next methodological consideration on the topic of frequency.

Recall from chapter 3 that the corpus used to extract frequencies is itWac (Baroni et al., 2009). At the time in which the experiment was designed and implemented, itWac was considered the most appropriate tool for the reasons detailed in 3.14. More recently, a new tool called SUBTLEX-UK (van Heuven et al., 2014) has been created for the purpose of investigating word frequency. Like CLEARPOND, SUBTLEX-UK is based on film and television subtitles, since recent research has shown that word frequencies based on this kind of corpora are better predictors of word processing times compared to other written resources (see van Heuven et al., 2014: 1176-1177). Apart from British English, other languages currently catered by this corpus are American English, German, Dutch, Spanish (Marian et al., 2012) and Polish (Mandera et al., 2015). An Italian version of SUBTLEX seems to exist (Crepaldi et al., 2013), but it appears to be in the somewhat preliminary stages and, to my knowledge, no published article has yet been made available that either presents its details or uses it for a practical application, at least at the time of writing. Nor does the official project webpage include information about an Italian version. Despite this, however, the tool exists and is worth at least mentioning for the reference of future studies, at the time of which SUBTLEX-IT may be more officially established and more easily accessible to scholars in disciplines other than corpus linguistics. In the present study, for example, it would be interesting to use SUBTLEX-IT as well as itWac and re-run the analyses to compare frequency results to see if one corpus is more predictive of recall accuracy or fixation measures than the other. Doing so may clarify the pattern of results found in 5.3.4.

Additionally, as we have seen in this and other sections of this thesis, the difficulty of achieving total variable control in experimental research implies an unavoidable lack of confidence in the veracity of any one single study (Sheen, 1996). Rigorously controlling for all potential confounds is therefore virtually impossible. This study is no exception and, on the one hand, it may be argued that it could have included, alongside the language proficiency and WM tests, other tests to control for the subjects' reading speed and familiarity with the subtitles, or measures of word predictability and perceptibility, neighbourhood and vocabulary size, and so on. On the other hand, however, if an experimenter included all these controls in the procedure, the result would be a battery of tests that would require considerable time to be carried out, thus easily resulting in additional obstacles to participant recruitment. Recall from 3.8 that the subjects who took part in this study were all volunteers and did not receive monetary compensation for their participation. Between watching the clip, taking the post-test, filling in the questionnaire and undergoing the control tests, each experimental session took up to 1h45min. Subjecting volunteers to all the control tests mentioned above would have meant adding another hour at least to each session, which is likely to discourage even the staunchest of volunteers, if they are not paid for their time. Therefore, especially considering the overt goal to maximise ecological validity in the present study (1.5), it was preferred to control for what were identified as the core potential confounds and limit the tests to those. Again, this example aptly highlights that a delicate balance needs to be struck between the ideal and the feasible when embarking on experimental studies of this kind. In the future, however, a study less concerned for ecological and external validity and primarily focussed on strict internal validity would certainly add valuable insights to the body of behavioural and acquisitional data currently available on reverse subtitles. In such a case, however, I would recommend going through a full ethical review process in order to be able to appropriately remunerate the participants for the considerable time they would be devoting to the experiment.

As far as the analysis is concerned, the currently available eye-tracking software does not allow to analyse subtitles and other dynamic stimuli like it does with static texts (see discussion in 3.17.1 and 5.3.4). While in the latter the software detects the boundaries between words and automatically creates the AOIs around all of them, in the former AOI creation is a largely manual process carried out by the researcher. This state of affairs has meant that, so far, most subtitle processing analyses have been limited to investigating the amount of attention at the subtitle-level rather than at the word-level, precisely because it is prohibitively time-consuming to analyse subtitle data word by word. This has been a major obstacle to obtaining fine-grained analyses of the reading process during AVT consumption. Recently, however, Kruger and Steyn (2014) have devised a robust Reading Index for Dynamic Texts (RIDT), designed specifically to compensate for such general lack of reliable indexes of reading behaviour in the context of subtitles (see also 2.10). The index considers the number of unique fixations for a participant *p* on a specific subtitle *s*, the number of words in that subtitle, the average forward saccade length for *p* in *s* and the standard word length of the video (Kruger and Steyn, 2014: 110). In their article, the authors used the index to establish whether such subtitle reading behaviour in English L2 correlates to comprehension test performance in students who were exposed to an English subtitled video recording of an academic lecture (bimodal condition). While it is beyond the scope of this section to address the index in detail, the authors do explain how they arrived at devising the index, how it was validated and how they used it in their study. The reader is therefore referred to their paper for more information. In the context of research like the one presented in this thesis, suffice it to say that the RIDT could prove very valuable in the future by allowing to calculate in a more automatic fashion to what extent participants read subtitles. Moreover, applying the index to other subtitling modes (e.g. reverse subtitles) would provide further evidence as to how the index behaves and how well it captures the nature of attention in AVT processing.

Furthermore, in this study eye movements provided an appropriate yet *coarse* measure of attention allocation, because the focus was not on lower-level processing characteristics such as parafoveal preview and spill-over effects. To control for effects of parafoveal preview, for example, Van Assche et al. included as a control variable the distance between prior fixation and the target word (2009: 925). Operations like these were not applied to this dataset for the reasons outlined in 5.3.4 and because doing so was not within the remit of the study. In the future, however, a more tightly controlled study where ecological validity (2.12) does not have the same weight as in the present study could be implemented, by collecting data at a higher frequency (e.g. 1000Hz/s), using a chin-rest eye-tracker and examining data at the individual word level. Doing so would make it possible to account for the above characteristics and therefore allow us to make more precise inferences on the nature of audiovisual information processing.

Finally, a way to complement and refine the findings outlined in this and the previous chapter would be to extend the eye-tracking movements considered. This study considered fixation counts and durations. In the future, analysing regressions (2.5.1) might shed further light on the issue of processing difficulty, if enough time is allowed for the very time-consuming process of classifying and quantifying them. The same prohibitive time constraints apply to shifts (2.6.1), which also would have had to be counted manually. However, from the discussion (5.3.3) it emerged that shifts could potentially play a role in explaining the roughly equal amount of attention allocation to the two translation conditions. Ghia (2012a) analysed this metric in her study and found significantly higher proportion of shifts when subtitle translation diverged from the ST, so including them in future analyses would allow a direct comparison with her parallel study on standard subtitles and could be beneficial to shed light on the cognitive mapping and more elaborate processing which is believed to stem from the continuous ST-TT comparison process (ibid.: 82). Another metric that it would be really interesting to analyse in the future is the percentage of skipped subtitles, which we know from 2.7.2 tends to be higher with reverse subtitles (compared to standard ones). Collecting this type of eye movement data would allow a comparison between the participants in d'Ydewalle and De Bruycker (2007), to whom the L2 in the subtitles was not meaningful at all, and participants in this study, who not only find the language meaningful, but who also have a high command of it. In the questionnaire, some people stated that they were surprised to see Italian subtitles on a film in their native language, yet admitted to reading the reverse subtitles nevertheless (see note 36 in 5.3.3), suggesting that, if percentage of skipped subtitles were calculated, it may be smaller in comparison with the figures reported in d'Ydewalle and De Bruycker (2007). Furthermore, other subjects spontaneously realised and reported that using this modality could be a useful technique for learning, which also suggests an attitude to read rather than skip the subtitles. I found one comment particularly revealing: "The subtitles were easy to follow and it [the study] made me aware of *how much I can learn through Italian subtitles* on English speaking films as opposed to the other way around *which is what is normally recommended by teachers and tutors*." (S_{18}, my italics). Not only did S_{18}, without any prompt to do so, comment on the learning potential of this modality specifically but, in doing so, also highlighted a common misconception in the use of reverse subtitles, i.e. that fact that, since they do not belong to a set of 'norms' commonly accepted, they cannot be useful and therefore tutors tend to avoid using as well as recommending them. In this sense, the study was successful in restarting the inevitably long process of reverting established trends and, most importantly, challenging some of its commonplace beliefs. Lastly, the comment from S_{18} (together with other relevant ones) did not appear under any of the three subtitle-specific questions, but emerged through question Q4 (see Appendix A.4 and section 2.12), purposefully included to prompt subjects to express any other comments they may have on the study as a whole. Considering participants more overtly as active respondents to the stimuli presented proved very useful in the context of this study. The data confirmed that participants can provide invaluable insights in their thought processes and beliefs, thus complementing the quantitative findings and assisting the researcher in their interpretation.

CHAPTER VI
CONCLUSIONS

6.1 Introduction

This was the first study to investigate the translation manipulation of reverse subtitles in the context of FLL. It assessed whether this manipulation has an effect on the viewer in terms of processing and memory, as measured by eye movements and a verbatim recognition post-test respectively. English native speakers aged 20-35 with a relatively advanced level of Italian (upper-intermediate B2 CEFR) watched a video excerpt in L1 English with L2 Italian subtitles while an eye-tracker monitored their eye movements. Formal similarity (literal transfer) and formal discrepancy (non-literal transfer) between L1 and L2 were compared with regard to verbatim recognition memory and eye movements (fixation counts and durations). Attention and noticing were also explored in this context. The two core aims of the research were (a) to assess whether Italian FL learners notice translational discrepancies between ST and TT, and (b) to establish whether reverse subtitles are recalled more or less accurately when they share some formal similarity with the L1. These aims were fulfilled by providing answers to the five RQs presented in chapter 3 (table 2). These RQs stem from the literature reviewed in chapter 2, are addressed in the data analysis in chapter 4 and guide the discussion of the findings in chapter 5. A summary is also reported in the table below for ease of reference. In this conclusive chapter, I will summarise and interpret the most significant experimental findings emerging from the RQs (6.2), reflect on how translation and reverse subtitles were found to influence FLL (6.3), examine the elements of originality in the study (6.4), propose avenues for further research on the topic (6.5), and finally open the discussion to consider some of the wider implications of this research (6.6).

	Research Questions
1.	Are literal items recalled more or less accurately than non-literal ones?
2.	Is there a difference in verbatim recall between lexical and syntactic items, and does linguistic category interact with translation condition in recall?
3.	Are literal and non-literal translations registered differently by viewers? Is there a relationship between visual attention and recall rate?
4.	Is there a significant difference in frequency between literal and non-literal items? Is there a relationship between frequency and recall rate and/or between frequency and eye-movements?
5.	Will viewers report noticing translational differences between the source (audio) and the target (subtitles)?

Table 29. Research questions summary.

6.2 Summary of the Findings and Contributions to Knowledge

As far as formal similarity is concerned, evidence is herein provided towards to the view that literal translations of words and structures in AVT are easier to retain as measured by visual recognition memory. By-subject and by-item t-tests revealed that L1-L2 similarity facilitated verbatim recognition of both L2 lexicon and syntax significantly more than L1-L2 discrepancy, while receiving fewer looks (although this difference in attention allocation was not statistically significant). Formal similarity seems to affect online processing in such a way that facilitates immediate surface form recognition memory. Since language learning is a gradual process (Nation, 2001) with noticing as its initial step (Schmidt, 2001), such recognition memory facilitation provides a first step in the process of learning new word forms while also confirming knowledge of already familiar words. The facilitation effect of formal similarity in recognition, however, did not hold in free-recall: what the subjects spontaneously reported in the open questionnaire are mostly non-literal renderings of the L1 (cases of formal L1-L2 discrepancy).

My preferred interpretation of the above findings is as follows. Formal similarity allows for smooth language encoding during moment-by-moment consumption of subtitled video, as evidenced by the lower fixation numbers and durations, indicating less processing effort. It also creates clarity in L1-L2 word connections in the mental lexicon and allows the learner to move on to the processing of the next subtitle without major disruptions. This clarity of the lexical links between L1 and L2 facilitates recognition, but since the correspondence is expected and clear in the mind of the learner, a literal item (that is, formal similarity) is generally not interesting enough to be spontaneously reported as striking in the subsequent open questionnaire. Formal discrepancy, on the other hand, creates less direct connections between L1 forms, L2 forms and the underlying concept in the mind of the learner. In these 'fuzzy' correspondences, in most cases the L2 forms are not as well-mapped onto their L1 equivalents as it is the case with formal similarity, and this creates interference at the level of exact form retrieval, getting in the way of mnemonic recognition accuracy. Sometimes, however, a minority of non-literal items come across as particularly striking for the learner. In these cases, formal discrepancy triggers noticing of the L2 string and increased mapping during AVT processing, which makes the unusual, less straightforward connection between L1 and L2 more memorable. Interestingly, how many and which non-literal items people find striking enough to be memorable appears to be quite personal, i.e. it changes from subject to subject. In these particular cases, a more elaborate memory trace is left, which results not only in correct recognition of the non-literal rendering in the MCQ, but also its spontaneous, unprompted recall in the questionnaire. In the present data, this mnemonic advantage only applies to few non-literal items, which are often already known by the learners (although some new words are also learnt in this way). When reporting these items in the questionnaire, most subjects also explained why they found them striking (see 5.3.5), and in doing so demonstrated that defiance of expectations plays a key role in triggering overt noticing (metalinguistic awareness). Defiance of expectations can occur when an L1 word – already mapped onto a known L2 translation equivalent in the mental lexicon – is suddenly presented in a translation pair with a different, unexpected L2 word (familiar or unfamiliar to the learner), such as in the *resistance – resistenza* vs. *resistance – guerriglia* example (see 5.3.1). Furthermore, whereas verbatim recognition memory requires to retrieve the exact word form as it appeared during viewing by choosing from a list of options, free-recall in this study requires active retrieval from learner memory without any prompts. Taken together, therefore, the findings summarised above support the view that recognition and recall of written L2 strings are based on interrelated but distinct processes (Clariana and Lee, 2001), requiring different processing tactics, where recall involves further elaboration of memory traces, while recognition strengthens existing memory traces (McDaniel and Mason, 1985).

As far as CLI (Cross-Linguistic Influence; see 2.5) is concerned, this piece of research showed that formal similarity can assist the language learner in recognition memory. It also showed that a degree of interference with formal discrepancy (non-literal renderings) can occur, since learners' mistakes in the recognition post-test were, in the vast majority of cases, back-translations of the L1 audio (literal renderings). The presence of such mistakes supports the view that formal similarity may provide the more direct, clearer and expected L1-L2 mental correspondences relied on in situations of unexpected recognition memory testing. Moreover, CLI studies so far (for example, research on cognates, transparent compounds and typological differences between orthographic systems) mostly looked at words or phrases in isolation rather than sentences in context. This study on the effect of translation manipulation on memory for exact phrasing enriched this picture by looking at CLI effects on words and structures presented in a meaningful environment, where words and structures did not occur in isolation but were embedded in the rich context of audiovisuals.

As far as linguistic category is concerned (sections 3.13, 4.3.5 and 5.3.2), the study showed that, if lexicon and syntax are considered regardless of translation condition, the difference in recall accuracy between them is not very large (70% versus 74% respectively), nor is it statistically significant. As far as the joint relationship between linguistic category and translation condition is concerned, a t-test of recall accuracy between literal and non-literal *syntactic* items revealed a significant difference, whereas a parallel t-test on *lexical* items did not. Although a logistic regression analysis did not highlight any significant interactions between translation condition and linguistic category, the t-tests as well as the visual representations of the data (4.3.5) would suggest that the two variables are not entirely independent, and possible explanations of why a significant interaction did not emerge are presented in 5.3.2. What the analysis clearly established is that literal items achieved higher overall accuracy both with syntax and lexicon compared to non-literal items, confirming the recognition memory superiority of this translation condition. Interestingly, within the realm of lexicon, the recall accuracy difference between the literal and non-literal conditions was slimmer, since some salient, individual content words were accurately recognised in the non-literal condition, suggesting that discrepancy at the syntactic level could be intrinsically less salient to the viewers compared to discrepancy at the lexical level. Individual L2 words that diverge from their L1 audio correspondent (e.g. L2 *vaticinio* for L1 *oracle*, see 5.3.1) seem to be more likely to be noticed and recognised than diverging L2 structures, which confirms lexical items as the preferred loci of salience compared to higher-level, syntactic structures (Pavesi and Perego, 2008a: 220).

As far as attention allocation is concerned, mean fixation durations of 124ms and 178ms for standard subtitles have been recorded in previous studies (see 2.7.1). With regard to reverse subtitles specifically, d'Ydewalle and De Bruycker (2007) recorded 185ms for one-liners and 201ms for two-liners in adult viewers. In their study, reverse subtitles were meaningless to the viewers (see 2.7.2). Average fixation duration for reverse subtitles in this study was longer, namely 217ms, possibly suggesting different depth or elaborateness of processing. It could be the case that, when reverse subtitles are processed by viewers who have working knowledge of the L2, *meaningful* information extraction lengthens fixation duration compared to a situation where viewers have no L2 knowledge (and therefore very little information can be extracted from the reverse subtitles). Within the limits outlined in 5.4, the eye movement analysis revealed a higher fixation count and duration with formal discrepancy compared to formal similarity, although this difference was not statistically significant. On the other hand, significant differences between the two translation conditions were registered on memory performance. With regard to eye movements and processing, these findings suggest that further information elaboration must happen 'behind the scenes', in the sense that it is not predictable by eye movements alone: there seems to be more to cognitive processing than meets the eye. Although eye movements are generally considered a direct correlate of where the mind attends to (the eye-mind assumption, see 2.6.2), there may be *qualitative* as well as *quantitative* differences in the attentional processes eye movements are believed to reflect. In turn, these qualitative factors may affect the further cognitive mechanisms that language input is submitted to, after having been looked at. The actual functioning of these internal cognitive mechanisms and their exact relationship with different types of attention remain for the most part still unclear, and full light cannot be shed on them in the framework of the present study. What this study does show, however, is that notable mnemonic gains are observed when conscious noticing during processing occurs, whereby a learner reflects on an L2 wording, makes connections between ST-TT, and tests their hypotheses, becoming aware of an L2 string to the point of reporting it explicitly. Thus, a complex relationship between language processing and memory (and therefore, to some extent, learning) is highlighted, and the importance of noticing and attentional processes is confirmed.

As far as frequency is concerned, this study found that within an AVT environment it was the translation manipulation, not relative subtitle frequency, which determined recognition memory in higher levels of proficiency. Further, relative frequency did not correlate with eye movements. The observations reported in 5.3.4 suggest that viewers process and recall items based on more than just frequency information. It appears that the role of frequency (measured as relative occurrences in a corpus) in word recognition can be moderated by other factors in these more advanced proficiency populations, at least in the context of AVT, where information processing happens through multiple channels. Amongst these other factors liable to affect recognition memory and processing are, for example, imageability and concreteness (Lohnas and Kahana, 2013: 1944), as well as higher-level sentential and textual variables (Cop et al., 2015). Moreover, as we have seen in the previous chapter, expectations, salience, perceptibility and familiarity can all play a role in the complex feat that is extracting and recalling information from an audiovisual text. I believe it would be beneficial if more reception studies included subtitle or word frequency information, as relatively little is known about how frequency interacts with other variables such as memory in audiovisual contexts. After Moran's (2012) study (see 2.7.3 and 2.11), the results of the present research strongly suggest that frequency effects in AVT warrant further investigation.

Finally, as far as proficiency is concerned, the study made a significant contribution by clarifying some contradictory opinions about what language level reverse subtitles should be used for. In the AVT literature, it has been stated that this modality is "rarely appropriate for beginners" (O'Connell, 2011: 161), yet equally that "[r]eversed subtitles (…) appear best suitable for beginners, since they enable them to access the foreign code in the written medium and to draw initial comparisons between such dialogue and its spoken counterpart in their L1" (Ghia, 2012a: 32). While O'Connell's statement can be contested by the research of Danan as well as Lambert and Holobow, who have all shown that reverse subtitles can be beneficial not only for higher-level English-French bilingual children but also for L2 French beginner pupils (see 2.3), the view purported by Ghia is the most commonly upheld (also see Danan, 1992; Mariotti, 2002). This is the first study that assessed both processing and learning in more advanced adult learners with reverse subtitles specifically. In doing so, it demonstrated that learners can benefit from the use of reverse subtitles not only at beginners and elementary proficiency levels, but also at more advanced levels. It appears that reverse subtitles can trigger comparisons between dialogues and written text also with proficiency B2 or above, whilst at the same time they promote hypothesis testing and they are used to check for translation solutions (Ghia, 2011; Danan, 2004).

6.3 Translation and Reverse Subtitles in FLL: Taking Stock

Like Čepon (2011: 16; see 2.3), I found that upper-intermediate FL learners tend to be naturally curious about the translation, i.e. how to render a particular L1 word, phrase or structure in the L2. The open questionnaire also suggests that these kinds of learners are not immune to a degree of enjoyment in engaging with this process. They do indeed check translation solutions and make constant L1-L2 comparisons, thanks to which they are able to both passively recognise and actively recall some L2 wordings at a later stage. Thus, these findings corroborate the view that this enhanced, elaborate processing leads to a deeper analysis of the L2 as well as to metalinguistic reflection (Danan, 2004). As Vanderplank notes: "(…) learners need to be able to develop their own conscious, critical faculties and their ability to draw language from programmes and build it into their own competence" (1988: 278). In this sense, reverse subtitling had a positive effect on the learners by helping them practice and develop such faculties, through triggering noticing processes. As it emerged from the questionnaire analysis, this subtitling modality activated conscious, critical faculties in the learners, who certainly drew more L2 exemplars from the video than I expected, and, by revealing some of their thought processes, showed that they integrated the L2 items into their evolving language systems.

The reverse subtitles in the experimental video made learners amenable to paying attention to and noticing the FL, as evidenced by their comments on grammatical, structural, semantic and cultural aspects of the L2 that they found striking, e.g. word gender, agreement, word order, polysemy and other semantic issues, as well as idiom translation. By 1988, Vanderplank had already noted that students in his study "reported inconsistencies, omissions, errors, and misleading inaccuracies" in the subtitles (1988: 276). In doing so, he provided anecdotal evidence of the role of noticing and metalinguistic reflection for FLL. Vanderplank also stated that "[d]ifferences between text and speech were seen as a hindrance at first, but were used as a useful and productive self-monitoring device later, and, of course, could be a very useful teaching technique" (ibid.). Based on the learners' explicit and unprompted comparisons between text and speech recorded in this study, together with the fact that a degree of novel word learning was also registered, the potential benefit of using reverse subtitles for FLL purposes seems evident, and complements previous findings on the acquisitional gains of this modality (e.g. Danan, 1992; d'Ydewalle and Pavakanun, 1997). As a matter of fact, reverse subtitling is not completely unknown in the FLL literature, having been suggested as a task to develop reading speed and L2 automatic processing (Hulstijn, 2001: 285). Laufer also suggests using reverse subtitles for a similar purpose: "(…) fluency activities include repeated reading of a text with the goal of reading faster each time, repeated recordings of a talk in the language laboratory, watching videos with reverse subtitling (in L2) and reading them quickly." (2005: 237). Despite these encouraging suggestions from scholars outside the field of AVT, reverse subtitling remains a translation modality for the most part overlooked in published research, especially if one compares it to the body of existing psycholinguistic and acquisitional studies on standard or bimodal input. The contribution of the present investigation, then, is all the more significant, in that it provides an extensive review and cognitive overview of reverse subtitling and a starting point for further discussion on its use in the FL classroom.

One of the arguments against the use of reverse subtitles is that they are 'unnatural', i.e. they do not correspond to a real-world situation, since learners will rarely be watching a film in their L1 with subtitles in the L2. This study challenged this common view that reverse subtitles, unlike standard ones, provide an artificial learning situation (Ghia, 2011: 99). First of all, many learning activities like fill-in-the-gaps and spot the difference tasks (where learners need to decide whether two pictures are the same or different) are also unnatural in this sense, because these are operations that learners will almost never find themselves performing outside the classroom (R. Ellis, 2003). Yet, they are virtually ubiquitous in foreign language classes. A fundamental question thus arises: how appropriate is it to judge the value of a learning activity exclusively on the basis of whether it happens in the real world? I argue that it is more productive to assess learning activities by the measurable acquisitional benefits they bring to the learner, especially in a *foreign* (as opposed to *second*) language learning context. In such context, exposure to the L2 is limited and often relegated to classroom hours, which makes it essential to maximise learning for a given amount of time. Consider, for example, Lambert et al.'s point: "[standard] subtitles nonetheless force viewers to return to the script of their native language (L1) to pick up the story line, thereby reducing the foreign flavour of the film and reminding viewers of their inadequacies in the second language" (1981: 133). In this sense, it could be proposed that standard L1 subtitles also present an unnatural viewing situation, when used in the language classroom, since they create a distance between viewers and the foreign language. In fact, in a recent study, they have been shown to harm the perception of the foreign soundtrack (Mitterer and McQueen, 2009), with the added aggravating circumstance of potentially raising the affective filter (Krashen, 1985) if learners end up being frustrated about their inability to identify foreign words in the fast L2 speech stream. While I am not trying to suggest that standard subtitles should be abandoned in FLT, I do believe it is time that reverse subtitles found a more stable, dignified place in the teaching practice in view of their specific acquisitional potential.

More broadly, this piece of research also argued for a reassessment of translation in SLA and FLL, based on the psycholinguistic evidence of its cognitive advantages. From chapter 5 we saw that learners not only processed, but also recognised, free-recalled and demonstrated an overt interest in the translation of reverse subtitles. The findings support recent reappraisals of the role of translation (2.4.1) in the language classroom, e.g. Leonardi (2010, 2011). Leonardi's fundamental assumption is that translation is a natural activity happening in the mind of the learner and that, at least in the initial stages of acquisition, it is virtually impossible to learn a L2 without a degree of comparison to one's L1. This study showed that, even at higher proficiency, where L2 development can be more independent of the L1, translation remains not only a natural activity learners engage with through consistent L1-L2 comparisons, but also one that can promote salience and memorability, both in recognition and free-recall. It would therefore seem logical to warrant further explorations to this potentially useful teaching technique, in order to assess whether, for example, the psycholinguistic and often incidental advantages described in the literature remain valid in a formal classroom environment. In light of this, *a priori* ostracising translation from the classroom, a deep-rooted attitude in most high-profile theories of language teaching throughout the 20th century (Cook, 2010: xv), may no longer be justified, at least in the case of audiovisual translation. I am far from proposing that translation should be the prime means of teaching or learning a foreign language, or that L2 monolingual activities should be abandoned. I firmly believe that learners need to undergo processes of automatisation and gain procedural knowledge, if they are to achieve successful mastery of the L2, which comes only through being exposed to large amounts of L2 input and through active, engaged L2 practice. However, my results – alongside several others reviewed and discussed in this thesis – highlight that ignoring translation in FLL and FLT purely because it is reminiscent of obsolete grammar-translation methods simply lacks a rational foundation. As this study has shown, at issue should now be the *when* and *how* – no longer the *if* – of translation use in the foreign or second language classroom. I believe shifting the discussion towards these questions will ultimately benefit both teachers and learners.

6.4 Elements of Originality

First of all, as previously mentioned (1.3 and 2.3), reverse subtitles have only occasionally been investigated in FLL, mostly in studies with a contrastive focus. This thesis was the first study to provide a thorough account of reverse translation in its own right, starting from a cognitive perspective on FLL.

Second, such cognitive exploration of translation (2.4) is also a relatively underexplored topic in the context of AVT. The present study addressed this topic by considering different translation strategies, how they contribute to creating formal L1-L2 similarities and discrepancies (3.11) and how these, in turn, affect processing and potential learning. In doing so, it also provided a fresh contribution to the long-standing debate on transfer or CLI (Cross-Linguistic Influence) effects on language learning (2.5). Until now, CLI within the field of subtitling research had not been tackled directly, and by assessing the effects of formal similarity on the learners this piece of research provided a third original contribution in the development of reverse subtitle research.

Fourth, this piece of work sits at the interface between AVT, psycholinguistics and cognitive psychology, a field for which audiovisual information processing is becoming particularly relevant (Bairstow, 2012: 219). Within this framework, the present work adds to the body of cognitive research on subtitle perception – a burgeoning but still rather young research area – in which the need for more experimental studies addressing subtitled film processing has already been highlighted (ibid.: 218). Moreover, for the first time both behavioural and performance methods (Kruger, 2016) were combined in one single study, namely eye-tracking as well as memory performance tests, in the context of the reverse modality.

Fifth, this study addressed noticing in AVT, a research niche virtually never tackled experimentally (with the exception of Ghia, 2012a), perhaps because of the difficulty in defining and operationalising this psychological construct. Since no perfected methodology exists to investigate noticing to date (Ahn, 2014), another original aspect of the present study is that it resorts to data triangulation, which can be defined as "the combination of different methods so that the results collected through one method are contrasted with the results collected by a second or third method" (Brems and Ramos Pinto, 2013: 146). Thus, a combination of techniques were considered *together*, which has been deemed as an appropriate methodological procedure to investigate complex noticing processes (Leow, 2013b). Specifically, eye movements, a recognition memory post-test and an open questionnaire were analysed together, in order to provide a more complete account of noticing and to increase the reliability of the findings.

Finally, another original contribution of the present investigation was the attempt to integrate frequency in the analysis, i.e. "arguably the most important variable in word recognition research" (van Heuven et al., 2014: 1176). Frequency has been deemed to influence the processing of phonology, phonotactics, reading, spelling, lexis, morphosyntax, formulaic language, language comprehension and grammaticality (Ellis, 2002a). However, as we have seen (2.11), there is some controversy as to whether frequency *independently* influences processing and memory, and therefore has a psychological reality, or whether its effects arise because it acts as "a very good proxy for a range of other things" (Divjak, 2016: 19). However, this variable is generally believed to play a major role in models of word processing, to influence word recognition during reading and to be able to influence recognition memory. Moreover, it is central to linguistic theories such as usage-based models of acquisition, which maintain that language acquisition occurs on the basis of the exemplars encountered and "(...) is the piecemeal learning of many thousands of constructions and the frequency-biased abstraction of regularities within them." (Ellis, 2002a: 143). Therefore, frequency cannot be ignored in experimental studies, at least not in those that look at language processing specifically. Despite the lively debate surrounding this topic in much of the psycholinguistic and SLA literature, frequency has virtually never been incorporated in AVT research, with one recent major exception (Moran, 2012). Based on Moran's conclusions and the body of evidence emerging from the frequency research reviewed in 2.11, it was deemed appropriate to add a frequency variable to the present work. Doing so entailed devising a way of calculating frequency at the subtitle level, which may constitute a starting point in future studies where AVT researchers wish to include this variable in their analysis.

6.5 Future Research Directions

This study provided some experimental evidence that translation presented in audiovisual environments can play a role in learning (verbatim memory) and also that the concept of 'literalness' (2.4.3 and 3.11.1) has a psychological reality in the mind of the learner. Since translation is still used in many language classrooms as well as in language testing in schools and universities (2.4.1), a logical next step would be to examine its impact on learning in a classroom-based setting. While such an approach has its disadvantages (e.g. not easily allowing the use of behavioural measures like eye-tracking), it also has its advantages (e.g. being more ecologically valid), and is very much needed to integrate the findings of more experimental laboratory knowledge. As DeKeyser stated, "hardly any literature is available that combines the degree of control of a psycholinguistic experiment with the validity of research of real second language learning" (1998: 60). After an attempt to combine the two in the present study, I can confirm that the situation described by DeKeyser is understandable and should come as no surprise, since often the focus and aims of experimental research are sharply different from those of classroom applications, which can result in radically different methodological choices. I firmly believe that the two strands of research can complement each other and should develop parallel investigations if we are to assess what techniques and tools work in FLL, for which learners and in which settings. To this purpose, further investigations of reverse subtitles should be warranted not only in experimental laboratories but also through classroom applications, since it is only through the latter that this translation mode can be truly assessed in context.

Second, another direction for future research concerns *incidental* vs. *intentional* learning contexts. This study *de facto* assessed incidental acquisition through the reading of subtitles (2.6.4). Incidental refers to the fact that words and structures were acquired without a deliberate attempt by the learners to commit them to memory (Laufer and Girsai, 2008: 703), and is usually juxtaposed to intentional learning. In incidental learning, students are not aware that their knowledge will be probed by a subsequent test, the results of which will, therefore, assess such learning as a by-product of an activity (in this case, watching subtitled video) that is not overtly geared to memorising L2 input (Hulstijn, 2001: 271). In the future, it would be interesting to see if language uptake (see 2.6.4) in audiovisual contexts changes with intentional learning, i.e. a condition where students are provided with pre-learning instructions and warned about a subsequent retention test (ibid.: 268). As Kuppens reminds us:

> One should be careful in transposing the findings of a study of incidental language acquisition to situations of intentional language learning. Incidental and intentional language acquisition are very different processes when it comes to, for instance, the object of focus of the learner's attention (meaning vs. form), the amount of exposure

> needed (more vs. less), and the degree of (teacher) supervision and feedback (none vs. some). (2010: 79)

Therefore, the fact that the learners in this study *incidentally* noticed, learnt and reported Italian lexical and syntactic items from a reversely subtitled film cannot be safely taken to mean that reversely subtitled materials are also efficient to promote *intentional* language learning, a hypothesis that will require direct investigation though the classroom applications mentioned above.

Third, the present study reported some acquisitional gains obtained from using reverse subtitles. However, a distinction has to be made between short-term acquisition (addressed in the study) and acquisition proper (Ghia, 2012a: 93). The latter is a complex phenomenon articulated through time and perhaps best viewed in a developmental perspective. Unlike the former, it also includes a delayed post-testing phase to capture long-term effects of the experimental treatment (ibid.). Collecting longitudinal data makes it possible to describe learning patterns over longer periods of time and measure differences or change in specific variables from one point in time to another (Menard, 1991). Thus, it can crucially enrich our picture of FLL mechanisms and help us to capture what linguistics features cause learners long-term acquisitional issues. The call for more longitudinal investigations of this kind is a common one both in SLA (e.g. DeKeyser, 1998: 61) and in AVT (e.g. Ghia, 2012a). Therefore, future research should also investigate reverse subtitles in a longitudinal, developmental perspective, whether in the laboratory or in the classroom, and whether such approaches address incidental or intentional learning.

Fourth, in chapter 5 the discussion also addressed learner-initiated noticing, familiarity with the L2 input and expectations. Stemming from these subject-specific variables are the strategies that learners apply during the reception of AV products. According to Tuominen, these tactical reception strategies are the focus of attention on specific parts of the video: the choice between reading or skipping a subtitle, how viewers prioritise information, the different ways in which they deal with language they do not understand, and how suspicious or critical they are of the translation solution presented (2012: 191). Some of these reception strategies were addressed in this study through both quantitative and qualitative examination of the data. Although the viewers were allowed to express their personal opinions, further information could be garnered through more overtly ethnographic approaches on "the context-bound, subjective characteristics of reception" (ibid.: 192) that cannot be captured through quantitative data alone. For example, future studies on reverse subtitles could include a further explicit report phase along the lines of Taylor (2005), where students were required to explain how they used the different channels (written subtitles, sound, and moving images) and what strategies they used to integrate the L2 subtitles in the stream of information flow. This could be done, for example, by examining notes taken during viewing, using think-aloud protocols or in-depth oral interviews, and may be particularly informative in the case of reverse subtitles, since this is historically one of the least investigated translation modalities, on which explicit opinions from the learners are still for the most part lacking.

Another central topic that future research on reverse subtitles will have to tackle is multimodality, which has been defined as "the use of several semiotic modes in the design of a semiotic product or event, together with a particular way in which these modes are combined" (Kress and van Leeuwen, 2001: 20). In this sense, subtitles are added to a text – the audiovisual text – that is intrinsically multimodal. As Tuominen notes: "[i]n a normal reception situation, subtitles are never read by themselves: the meaning of an audiovisual message is constructed of auditive and visual elements as well as the translation." (2012: 189). For instance, if an object mentioned in a film dialogue is clearly visible on screen, a pronoun can be used in the subtitles to refer to it (Lertola, 2015: 255). As a consequence, in order to understand this scene, learners will have to integrate information from at least two different semiotic channels (1.3). Several investigations on AVT, however, including the present one, tend to concentrate on the verbal elements in the audiovisual product. In Gambier's words: "Can we not talk about language hypertrophy, paying less attention to camera moves, viewing angles, editing, soundtrack, tone of voices, facial expressions, gestures, gazes, body movements, all of which are also meaningful?" (2006: 3). Therefore, it is essential to explore how reverse subtitles interact with the multimodal fabric of the audiovisual medium and how learners integrate them in their overall viewing experience. Doing so will be invaluable to scale down this 'language hypertrophy', to go beyond the mere linguistic data by considering other contextual aspects of the rich, polysemiotic experience that is viewing subtitled material.

Last but by no means least, to provide a well-rounded picture of reverse subtitles in FLL, it will be necessary to carry out parallel investigations on the active practice of translating into the L2, where students themselves *create* (subtitling as a task) rather than *consume* (subtitles as support) the reverse subtitles. Some published research on subtitling as a task exists (e.g. Williams and Thorne, 2000; Neves, 2004; Incalcaterra McLoughlin and Lertola, 2011; Lertola, 2012, 2015; Talaván, 2010; Talaván, 2011; Kantz, 2015; Ragni, forthcoming) and subtitling platforms specifically designed for learners have also been produced, such as LeVis and ClipFlair (Sokoli, 2006, 2015; Zabalbeascoa et al., 2012). Yet, very few of these publications tackle reverse subtitles (except Talaván and Rodríguez-Arancón 2014; Talaván and Ávila-Cabrera 2015). This research area has produced extremely interesting results so far, including methodological proposals, theoretical accounts, sample activities as well as creative ideas to exploit subtitling tasks in the FL classroom. Broadening the research on learner translation practice to include reverse subtitles will prove invaluable in the future to explore the full potential of this modality.

6.6 Research Significance and Wider Implications

In 1998, Campbell stated that Translation Studies (TS) was only starting to create "an experimental arm" (1998: 19). More than 20 years later, this state of affairs has visibly changed. As this thesis shows, there has been recent interest among TS scholars in research topics such as cognitive processing, reading and memory. Since the new millennium, an increasing number of studies involving, for example, eye-tracking technology have been investigating translation in conjunction with the above topics. This situation is promising for those who believe interdisciplinary research on translation can contribute to furthering our understanding of how foreign languages are processed and learnt. At the interface between TS, cognition and linguistics, the present study provides a further step in this direction and contributes to showing that TS is "embracing interdisciplinary developments, and no longer only borrows from neighbouring disciplines, but actively contributes to them" (Kruger, 2016). In his recent publication, Kruger (2016: 7) distinguished between performance studies, on the one hand, and behavioural or physiological studies, on the other, and notes that recent research on reverse subtitles specifically (d'Ydewalle and De Bruycker, 2007) has failed to establish the relationship between reading and other performance measures. In this thesis, this gap was filled precisely by considering the two aspects together. As we have seen in 6.4, this is the first study on reverse subtitles to provide an explicit account of both behavioural measures of reading (through eye-tracking) and their relationship with performance measures (i.e. a verbatim recognition test and an open questionnaire). Given that the interest in experimental psycholinguistics and AVT has been growing sharply in the last ten years, I suspect it will not be the last.

The implications of this work, however, are not limited to the research community, but have an impact on teaching practices too. As mentioned, an encouraging degree of incidental acquisition through reverse subtitles was registered in the learners of this study. That is to say, the mnemonic gains were a result of carrying out an action (watching), not a result of deliberately committing the information to memory. In light of the findings, I argue that teachers should not only provide guidance to pupils *inside* the classroom – where they know that they are to make a conscious, intentional effort to learn new language – but also encourage them to maximally exploit the FL *outside* the classroom in incidental contexts, for example by watching digital video at home. This is particularly relevant in foreign as opposed to second language contexts, in which classroom time is often the only point of contact with the FL. If learners had extra contact with FL input in their own time, the result would be an increase in and a diversification of learning opportunities, while doing so would also foster learner autonomy in a situation of low affective filter (Krashen, 1985). As a consequence of this research, therefore, a relevant task for teachers is not only to integrate reversely subtitled video in classroom activities, but also to provide pupils "with skills, strategies and viewing behaviors that optimize their incidental language acquisition from media exposure outside the classroom" (Kuppens, 2010: 80). For example, they could train learners to pay more overt attention to unfamiliar L2 words occurring in the subtitles, spot idioms or play on words that are typically difficult to translate, and note down translation mistakes or oddities for later class discussion. In order for teachers to be able to inform students on these strategies, however, they would need to know what students are able to process and take in from viewing subtitled video. The present study has provided some initial evidence of these strategies and viewing behaviours from advanced learners of Italian L2, but much more is still unknown for other levels of proficiency, foreign languages and audiovisual genres. I therefore hope that this research may inspire others to use reversely subtitled input in classrooms, pilot reverse subtitling activities in language courses and appropriately document these experiences, so that the findings can be disseminated to inform further applications and a set of best practices can emerge.

Implications of this research may arise for course designers and syllabus writers as well. In the past, there has been an attempt to put forward initial recommendations for creating 'SLA-oriented subtitles' (Pavesi and Perego, 2008a), i.e. subtitles meant to be most useful to the language learner and created for the language classroom specifically. Pavesi and Perego advise to maintain some degree of closeness between ST and TT: "word order should be respected as closely as possible, and so should the order of information units."

(2008a: 223). They continue: "subtitles should preserve the textual and syntactic patterns of the original language. These devices are meant to facilitate the matching between the original soundtrack and the subtitles as well as replicate the original patterns of salience." (ibid.). They are thus *de facto* suggesting a more literal pattern of translation, like the one investigated herein. The results I presented confirm that such literal transfer strategies promote verbatim recognition, and might therefore be particularly suited for beginners, who tend to be more reliant on structural and orthographic correspondences and may benefit from similarities the most. However, the results also show that, at least for advanced learners, spontaneous, unprompted recall occurred preferentially with non-literal items. Therefore, if the students' proficiency level is high enough (B2 or higher), purposefully creating a ST-TT discrepancy in the subtitles could be equally beneficial in promoting acquisition, potentially even more successfully so if exercises such as spot-the-difference or other tailored activities aimed at highlighting the discrepancy follow the viewing phase. In view of these findings, syllabus designers could create reverse subtitles learning activities tailored to lower or higher language levels by manipulating the degree of similarity between speech and text. However, for an informed integration of this translation mode in FLL, more research needs to address the specifics of how proficiency interacts with audiovisual texts in the reverse mode. Future comparative studies could assess mnemonic performance and reading patterns in learners of different language levels to establish whether formal similarity displays the same or different effects for recognition and spontaneous recall in these learner groups.

Finally, there may be implications at a broader government level. This research is part of a growing body of psycholinguistic and SLA studies in AVT, which is providing increasingly more refined empirical findings on how both subject-specific (e.g. proficiency level, working memory) and input-specific factors (e.g. frequency, perceptibility, type of translation) affect processing, language learning and classroom practice. Overall, these studies show positive effects of subtitles and a positive attitude of the receivers, both teachers and learners. If appropriately raised with decision makers in countries like, for example, Italy (where dubbing maintains a firm hegemony over other translation types, at least in mainstream cinema), these findings from the research community could provide the foundations for informed policy changes. Thus, education programmes could overtly take into account the cognitive benefits and drawbacks associated with using subtitles to learn a language. Of course, as it is true in several fields of inquiry, research findings rarely correspond to immediate changes but take a relatively long time to affect decisions. This is also the case for language learning and teaching, especially at institutional levels, where policy updates can often come as a consequence of governmental rather than educational turns. Yet, I believe there is scope to impact policies in the long run, if the right opportunities to raise visibility at national and international levels are exploited and research momentum is maintained. Therefore, I sincerely hope that AVT in FLL continues to flourish, so that more experimental studies and classroom applications can advance our understanding of the learning mechanisms at play with different translation modes, whilst telling us how these modes are best used at different proficiency levels and in different learning contexts.

Bibliography

Adelman, J.S., Brown, G.D. and Quesada, J.F. 2006. Contextual Diversity, not Word Frequency, Determines Word-naming and Lexical Decision Times. *Psychological Science*. 17(9), pp. 814-823.

Ahn, J.I. 2014. Attention, Awareness, and Noticing in SLA: A Methodological Review. *Michigan State University Working Papers in SLS (Second Language Studies)*. 5, pp. 56-65.

Allan, M.D. and Rugg, K.A. 1997. An Event-Related Potential Study of Explicit Memory on Tests of Cued Recall and Recognition. *Neuropsychologia*. 35(4), pp. 387–397.

Allison, P. 2013. *What's the Best R-Squared for Logistic Regression?* [Online]. [Accessed 18 January 2016]. Available from: http://statisticalhorizons.com/r2logistic

Anderman, G. and Rogers, M. 1996. eds. *Words, Words, Words. The Translator and the Language Learner.* Avon: Multilingual Matters Ltd.

Andersen, R. 1983. ed. *Pidginization and Creolization as Language Acquisition*, Rowley, Massachusetts: Newbury House.

Anderson, J.R. 1990. *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Andersson, R., Nyström M. and Holmqvist, K. 2010. Sampling Frequency and Eye-tracking Measures: How Speed Affects Durations, Latencies, and More. *Journal of Eye Movement Research*. 3(3), 6, pp. 1-12.

Araújo, V.L.S. 2008. The Educational Use of Subtitled Films in EFL Teaching. In: Díaz Cintas, J. ed. *The Didactics of Audiovisual Translation.* Amsterdam: John Benjamins, pp. 227-238.

Atkinson, R.C. and Shiffrin, R.M. 1968. Human Memory: A Proposed System and its Control Processes. In: Spence, K.W. and Spence, J.T. eds. *The Psychology of Learning and Motivation (Vol. 2)*. New York: Academic Press. pp. 89-195.

Baars, B.J. 2002. The Conscious Access Hypothesis: Origins and Recent Evidence. *TRENDS in Cognitive Sciences.* 6(1), pp. 47-52.

Baayen, R.H., Davidson, D. J. and Bates, D. M. 2008. Mixed-effects Modeling with Crossed Random Effects for Subjects and Items. *Journal of Memory and Language*. 59, 390-412.

Baddeley, A.D. 1986. *Working Memory*. Oxford: Oxford University Press.

Baddeley, A.D. 2000. The Episodic Buffer: A New Component of Working Memory? *Trends in Cognitive Sciences*. *4*(11), pp. 417-423.

Baddeley, A.D. 2003. Working Memory and Language: an Overview. *Journal of Communication Disorders*. 36, pp. 189-208.

Baddeley, A.D. and Hitch, G. 1974. Working Memory. In: Bower, G.H. ed. *The Psychology of Learning and Motivation: Advances in Research and Theory (Vol. 6)*. New York: Academic Press, pp. 47-90.

Bairstow, D. 2012. Audiovisual Processing while Watching Subtitled Films: A Cognitive Approach. In: Serban, A. Matamala, A. and Laveur J.-M. eds. *Audiovisual Translation in Close-up: Practical and Theoretical Approaches*. Bern: Peter Lang, pp. 205-219.

Baker, M.C. 2001. Syntax. In: Aronoff, M. and Rees-Miller, J. eds. *The Handbook of Linguistics*. Oxford: Blackwell, pp. 265-294.

Baker, M. 1992. *In Other Words: A Coursebook on Translation*. London: Routledge.

Baker, M. 2011. *In Other Words: A Coursebook on Translation*. 2nd ed. London: Routledge.

Baldry, A. and Thibault, P.J. 2006. *Multimodal Transcription and Text Analysis. A multimedia Toolkit and Coursebook with Associated Online Course*. London and Oakville: Equinox.

Balota, D.A., Pollatsek, A. and Rayner, K. 1985. The Interaction of Contextual Constraints and Parafoveal Visual Information in Reading. *Cognitive Psychology*. 17, pp. 364-390.

Baltova, I. 1994. The Impact of Video on the Comprehension Skills of Core French Students. *Canadian Modern Language Review*. 50(3), pp. 507-532.

Baltova, I. 1999. *The Effect of Subtitled and Staged Video Input on the Learning and Retention of Content and Vocabulary in a Second Language*. Ph.D. Thesis, University of Toronto, Canada.

Barni, M., Bandini, A., Sprugnoli, L., Lucarelli, S., Scaglioso, A.M., Strambi, B., Fusi, C. and Arruffoli, A.M. 2009. *Linee guida CILS. Certificazione di Italiano come Lingua Straniera. Università per Stranieri di Siena*. Perugia: Guerra Edizioni. [Online]. [Accessed

28 July 2016]. Available from: http://www.gedi.it/cils/file/5/12/file/linee_guida_cils_pdf(2).pdf

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. 2009. The WaCky Wide Web: a Collection of Very Large Linguistically Processed Web-crawled Corpora. *Language Resources and Evaluation*. 43, pp. 209-226.

Barr, J.D. 2008. Analyzing 'Visual World' Eyetracking Data Using Multilevel Logistic Regression. *Journal of Memory and Language*. 59, pp. 457-474.

Barr J.D., Levy R., Scheepers C. and Tily, H.J. 2013. Random Effects Structure for Confirmatory Hypothesis Testing: Keep it Maximal. *Journal of Memory and Language*. 68, pp. 255-278.

Bartlett, J. 2014. *R Squared in Logistic Regressio*n. [Online]. [Accessed 18 January 2016]. Available from: http://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/

Bates, D.M. 2010. *lme4: Mixed-effects Modelling with R. Springer*. [Online]. [Accessed 22 November 2015]. Available from: http://lme4.0.r-forge.r-project.org/lMMwR/lrgprt.pdf

Becker, L.A. 2015. *Effect Size (ES)*. Lecture Notes. [Online]. [Accessed 22 November 2015]. Available from: http://www.uccs.edu/lbecker/effect-size.html

Belletti, A. and Rizzi, L. 1981 The Syntax Of 'Ne': Some Theoretical Implications. *The Linguistic Review*. 1(2), pp. 117-154.

Benson, M.J. 2000. The Secret Life of Grammar-Translation. In: Trappes-Lomax, H. ed. *Change and Continuity in Applied Linguistics*. Clevedon: Multilingual Matters, pp. 35-51.

Berruto, G. 1998. *Sociolinguistica dell'Italiano Contemporaneo*. Roma: Carocci Editore.

Biehler, R.F. and Snowman, J. 1993. *Psychology Applied to Teaching*. Houghton: Mifflin.

Bird, S.A. and Williams, J.N. 2002. The Effect of Bimodal Input on Implicit and Explicit Memory: An Investigation into the Benefits of within-language Subtitling. *Applied Psycholinguistics*. 23, pp. 509-533.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Henry, M., Stevens, M.H., and White, J-S. S. 2008. Generalized Linear Mixed Models: a Practical Guide for Ecology and Evolution. *Trends in Ecology and Evolution*. 24(3), pp. 127-135.

Bordag, D., Kirschenbaum, A., Opitz, A., Rogahn, M. and Tschirner, E. 2016. Incidental Acquisition of Grammatical Features During Reading in L1 and L2. *Studies in Second Language Acquisition*. 38, pp 445-483.

Borrás, I. and Lafayette, R.C. 1994. Effects of Multimedia Courseware Subtitling on the Speaking Performance of College Students of French. *The Modern Language Journal.* 78(1): pp. 61-75.

BOS, Bristol Online Survey. 2016. *The Online Survey Tool Designed for Academic Research, Education and Public Sector Organisations*. [Online]. [Accessed 27 July 2016]. Available from: https://www.onlinesurveys.ac.uk/about/

Bravo, M.C.C. 2008. *Putting the Reader in the Picture. Screen Translation and Foreign-Language Learning.* Ph.D. Thesis, Universitat Rovira i Virgili.

Brems, E. and Ramos Pinto, S. 2013. Reception and Translation. In: Gambier, Y. and van Doorslaer, L. eds. *Handbook of Translation Studies, Volume IV*. Amsterdam/Philadelphia: John Benjamins, pp. 142-147.

Brown, S. 2015. *Measures of Shape: Skewness and Kurtosis*. [Online]. [Accessed 11 September 2015]. Available from: http://brownmath.com/stat/shape.htm

Brunswik, E. 1947. Systematic and Representative Design of Psychological Experiments with Results in Physical and Social Perception. *Syllabus Series, No. 304*. Berkeley: University California Press.

Brysbaert, M., Stevens, M., Mandera, P. and Keuleers, E. 2016. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology*. [Online]. [Accessed 20 September 2016]. Available from: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01116/full

Bulmer, M.G. 1979. *Principles of Statistics*. New York: Dover.

Burzio, L. 1986. *Italian Syntax: A Government-Binding Approach*. Dordrecht: Reidel.

Caffrey, C. 2008. *Relevant Abuse? Investigating the Effects of an Abusive Subtitling Procedure on the Perception of TV Anime Using Eye Tracker and Questionnaire*. Ph.D. Thesis: Dublin City University.

Caimi, A. 2005. Subtitling in a Cognitive Perspective to Encourage Second Language Learning. In: Sanderson, J.D. ed. *Research on Translation for Subtitling in Spain and Italy*. Alicante, Publicaciones de la Universidad de Alicante, pp. 65-77.

Caimi, A. 2006. Audiovisual Translation and Language Learning: The Promotion of Intralingual Subtitles. In: Díaz Cintas, J., Orero, P. and Remael, A. eds. Special Issue of *JoSTrans, Journal of Specialised Translation*. 6, pp. 85-98.

Caimi, A. 2012. Cognitive Insights into the Role of Subtitling in L2 Learning. In: Şerban, A., Matamala A. and Lavaur, J.M. eds. *Audiovisual Translation in Close-up: Practical and Theoretical Approaches*. Bern: Peter Lang, pp. 113-129.

Calvino, I. 1947. *Il Sentiero dei Nidi di Ragno* [The Path to the Nest of Spiders]. Torino: Einaudi.

Campbell, S. 1998. *Translation into the Second Language*. New York: Addison Wesley Longman.

Čepon, S. 2011. Interlingual Subtitling as a Mode of Facilitating Incidental Foreign Language Acquisition. *English for Specific Purposes World*. 33(11).

Chaffin, R., Morris, R.K. and Seely, R. E. 2001. Learning New Word Meanings from Context: A Study of Eye Movements. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 27(1), pp. 225-235.

Clariana, R.B. and Lee, D. 2001. The Effects of Recognition and Recall Study Tasks with Feedback in a Computer-based Vocabulary Lesson. *Educational Technology Research and Development*. 49(3), pp. 23-36.

Clark, H.H. 1973. The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. *Journal Of Verbal Learning And Verbal Behavior*. 12, pp. 335-359.

Clifton, C. and Staub, A. 2011. Syntactic Influences on Eye Movements during Reading. In: Liversedge, S. and Gilchrist, I. eds. *Oxford Handbook on Eye Movements*. Oxford: Oxford University Press, pp. 895-909.

Clifton, C. Jr., Staub, A. and Rayner, K. 2007. Eye Movements in Reading Words and Sentences. In: Van Gompel, R.P., Fisher, M., Murray, W. and Hill, R.L. eds. *Eye Movements: a Window on Mind and Brain*. Amsterdam: Elsevier, pp. 341-371.

Coe, R. 2002. *It's the Effect Size, Stupid. What Effect Size Is and Why It Is Important*. British Educational Research Association Annual Conference, 12-14 September, Exeter, UK.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioural Sciences*. Newbury Park, CA: Sage.

Cohen, J. 1992. A Power Primer. *Psychological Bulletin*. 112(1), pp. 155-159.

Coleman, A.M. 2015. ed. *A Dictionary of Psychology*. 4th ed. Oxford: Oxford University Press.

Conway, A.R.A., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O. and Engle, R.W. 2005. Working Memory Span Tasks: A Methodological Review and User's Guide. *Psychonomic Bulletin and Review.* 12(5), pp. 769-786.

Cook, G. 2010. *Translation in Language Teaching: An Argument for Reassessment*. Oxford: Oxford University Press.

Cook, R.D. and Weisberg, S. 1982. *Residuals and Influence in Regression*. New York: Chapman and Hall.

Cop, U., Drieghe D. and Duyck, W. 2015. Eye Movement Patterns in Natural Reading: A Comparison of Monolingual and Bilingual Reading of a Novel. *PLoS ONE*. 10(8), pp. 1-38.

Craik, F.I.M. and Lockhart, R.S. 1972. Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior.* University of Toronto. 11, pp. 671-684.

Crawley, M.J. 2007.*The R Book*. Chichester: Wiley.

Crepaldi, D., Keuleers, E., Mandera, P. and Brysbaert, M. 2013. *SUBTLEX-IT: A Frequency List for Italian Based on Movie Subtitles*. Incontro di Psicolinguistica, 22 November, Bologna. [Online]. [Accessed 02 October 2016]. Available from: http://www.davidecrepaldi.net/wordpress/wp-content/uploads/2011/11/Crepaldi_Bologna2013.pdf

Culler, J. 1976. *Saussure.* Glasgow: Fontana/Collins.

Curran, T. and Keele, S.W. 1993. Attentional and Nonattentional Forms of Sequence Learning. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 19, pp. 189-202.

Danan, M. 1992. Reversed Subtitling and Dual Coding Theory: New Directions for Foreign Language Instruction. *Language Learning.* 42(4), pp. 497-527. Available from: DOI:10.1111/j.1467-1770.1992.tb01042.x

Danan, M. 2004. Captioning and Subtitling: Undervalued Language Learning Strategies. *Meta: Translators' Journal.* 49(1), pp. 67-77.

Daneman, M. and Carpenter, P.A. 1980. Individual Differences in Working Memory and Reading. *Journal of Verbal Learning and Verbal Behavior.* 19, pp. 450-466.

Daneman, M. and Merikle, P.M. 1996. Working Memory and Language Comprehension: A Meta-analysis. *Psychonomic Bulletin and Review.* 3, pp. 422-433.

Dark, V.J. and Loftus, G.R. 1976. The Role of Rehearsal in Long-Term Memory Performance. *Journal of Verbal Learning and Verbal Behavior.* 15, pp. 479-490.

De Bot, K., Jagt, J. Janssen, H., Kessels, E. and Schils, E. 1986. Foreign Television and Language Maintenance. *Second Language Research.* 2(1), pp. 72-82.

De Groot, A.M.B. 1992. Determinants of Word Translation. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 18 (5), pp. 1001-1018.

De Groot, A.M.B. and Nas, G.L.J. 1991. Lexical Representation of Cognates and Noncognates in Compound Bilinguals. *Journal of Memory and Language.* 30, pp. 90-123.

De Linde Z. and Kay, N. 1999. *The Semiotics of Subtitling.* Manchester: St. Jerome Publishing.

De Roberto, E. 2010. *Futuro. Enciclopedia dell'Italiano.* [Online]. [Accessed 26 June 2013] Available from: http://www.treccani.it/enciclopedia/futuro_(Enciclopedia_dell'Italiano)/

De Santis, C. 2010. *Formule Esclamative. Enciclopedia dell'Italiano.* [Online]. [Accessed 10 May 2017] Available from: http://www.treccani.it/enciclopedia/formule-esclamative_(Enciclopedia-dell%27Italiano)/

DeKeyser, R.M. 1998. Beyond Focus on Form: Cognitive Perspectives on Learning and Practicing Second language. In: Doughty C., and Williams, J. eds. *Focus on Form in Classroom Second Language Acquisition*. Cambridge: Cambridge University Press, pp. 42-63.

Deslile, J., Lee-Jahnke, H. and Cormier, M.C. 1999. *Terminologie de la Traduction: Translation Terminology. Terminología de la Traducción. Terminologie der Übersetzung.* Amsterdam and Philadelphia: John Benjamins.

Deubel, H. and Schneider, W.X. 1996. Saccade Target Selection and Object Recognition: Evidence for a Common Attentional Mechanism. *Vision Research.* 36(12), pp. 1827-1837.

Dijkstra, T. 2007. The Multilingual Lexicon. In: Gaskell, M.G. ed. *The Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press, pp. 251-265.

Dijkstra, T., Grainger, J. and van Heuven, W.J.B. 1999. Recognition of Cognates and Interlingual Homographs: The Neglected Role of Phonology. *Journal of Memory and Language*. 41, pp. 496-518.

Dijkstra, T., Van Jaarsveld, H. and Brinke, S.T. 1998. Interlingual Homograph recognition: Effects of Task Demands and Language Intermixing. *Bilingualism: Language and Cognition.* 1(1), pp. 51-66.

Divjak, D. 2016. The Role of Lexical Frequency in the Acceptability of Syntactic Variants: Evidence From that-Clauses in Polish. *Cognitive Science*. March 2016, 1-29.

Dixon, P. 2008. Models of Accuracy in Repeated-Measures Designs. *Journal of Memory and Language*. 59, pp. 447-456.

Dörnyei, Z. 2003. *Questionnaires in Second Language Research: Construction, Administration, and Processing*. Mahwah, NJ: Lawrence Erlbaum.

Doughty, C. 2001. Cognitive Underpinnings of Focus on Form. In: Robinson, P. ed. *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press, pp. 206-257.

Doughty, C. and Williams, J. eds. 1998. *Focus on Form in Classroom Second Language Acquisition*. Cambridge: Cambridge University Press.

Doughty C. and Williams, J. 1998a. Pedagogical Choices in Focus on Form. In: Doughty, C. and Williams, J. eds. *Focus on Form in Classroom Second Language Acquisition.* Cambridge: Cambridge University Press, pp. 197-261.

Duffy, S.A., Morris, R.K. and Rayner, K.. 1988. Lexical Ambiguity and Fixation Times in Reading. *Journal of Memory and Language*. 27, pp. 429-446.

Dunlop, W.P., Cortina, J.M., Vaslow, J.B. and Burke, M.J. 1996. Meta-analysis of Experiments with Matched Groups or Repeated Measures Designs. *Psychological Methods*. 1, pp. 170-177.

Dupuy, B., and Krashen, S. 1993. Incidental Vocabulary Acquisition of French as a Foreign Language. *Applied Language Learning*. 4, pp. 55-63.

Dussias, P. 2010. Uses of Eye-tracking Data in Second Language Sentence Processing Research. *Annual Review of Applied Linguistics*. 30, pp. 149-166.

d'Ydewalle, G. and De Bruycker, W. 2003. Reading Native and Foreign Language Television Subtitles in Children and Adults. In: Radach, R., Hyona, J. and Deubel, H. eds. *The Mind's Eye. Cognitive and Applied Aspects of Eye Movement Research*. Amsterdam: Elsevier, pp. 671-684.

d'Ydewalle, G. and De Bruycker, W. 2007. Eye Movements of Children and Adults while Reading Television Subtitles. *Psychological Research*. 71(4), pp. 458-466.

d'Ydewalle, G. and Gielen, I. 1992. Attention Allocation with Overlapping Sound, Image, and Text. In: Rayner, K. ed. *Eye Movements and Visual Cognition: Scene Perception and Reading*. New York: Springer-Verlag, pp. 414-427.

d'Ydewalle, G. and Pavakanun, U. 1995. Acquisition of a Second/Foreign Language by Viewing a Television Program. In: Winterhoff-Spurk, P. ed. *Psychology of Media in Europe: The State of the Art – Perspectives for the Future*. Opladen, Germany: Westdeutscher Verlag GmbH, pp. 51-64.

d'Ydewalle, G. and Pavakanun, U. 1996. Le sous-titrage a la telévision facilite-t-il l'acquisition des langues [Does Television Subtitling Facilitate Language Acquisition]? In: Gambier, Y. ed. *Les Transferts Linguistiques dans les Medias Audiovisuels*. Lille, France: Presses, pp. 217-223.

d'Ydewalle, G. and Pavakanun, U. 1997. Could Enjoying a Movie Lead to Language Acquisition? In: Winterhoff-Spurk, P. and Van der Voort, T. eds. *New Horizons in Media Psychology*. Opladen, Germany: Westdeutscher Verlag GmbH, pp. 145-155.

d'Ydewalle, G., Praet, C., Verfaillie, K. and Van Rensbergen, J. 1991. Watching Subtitled Television: Automatic Reading Behavior. *Communication Research*. 18(5), pp. 650-666.

d'Ydewalle, G. and Van de Poel, M. 1999. Incidental Foreign-Language Acquisition by Children Watching Subtitled Television Programs. *Journal of Psycholinguistic Research*. 28, pp. 227-244.

d'Ydewalle, G., and Van Rensbergen, J. 1989. Developmental Studies of Text-picture Interactions in the Perception of Animated Cartoons with Text. In: Mandl, H. and Levin, J.R. eds. *Knowledge Acquisition from Text and Pictures*. Amsterdam: Elsevier Science, pp. 233-248.

d'Ydewalle, G., Van Rensbergen, J. and Pollet, J. 1985. Reading a Message When the Same Message is Available Auditorily in Another Language: The Case of Subtitling.

Psychological Reports of the University of Leuven (No. 54). Laboratory of Experimental Psychology. Leuven: Katholieke Universiteit.

d'Ydewalle, G., Van Rensbergen, J. and Pollet, J. 1987. Reading a Message When the Same Message is Available Auditorily in Another Language: The Case of Subtitling. In: O'Regan, J.K. and Lévy-Schoen, A. eds. *Eye Movements: From Psychology to Cognition*. Amsterdam: Elsevier Science, pp. 313-321.

d'Ydewalle, G., Warlop, L. and Van Rensbergen, J. 1989. Television and Attention: Differences between Young and Older Adults in the Division of Attention over Different Sources of TV Information. *Medienpsychologie: Zeitschrift für Individual – und Massenkommunikation*. 1(42), pp. 42-57.

Ebrahim, G.J. 1999. *Research Methods II: Multivariate Analysis*. [Online]. [Accessed 03 January 2016]. Available from: http://www.oxfordjournals.org/our_journals/tropej/online/ma.html and http://www.oxfordjournals.org/our_journals/tropej/online/ma_chap14.pdf

Ehrlich, S.E. and Rayner, K. 1981. Contextual Effects on Word Perception and Eye Movements during Reading. *Journal of Verbal Learning and Verbal Behavior*. 20, pp. 641-655.

Ellis, N.C. 2001. Memory for Language. In: Robinson, P. ed. *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press, pp. 33-68.

Ellis, N.C. 2002a. Frequency Effects in Language Processing. Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*. 24, pp. 143-188.

Ellis, N.C. 2002b. Reflections on Frequency Effects in Language Acquisition: A Response to Commentaries. *Studies in Second Language Acquisition*. 24, pp. 297-339.

Ellis, N.C. and Schmidt, R. 1998. Rules or Associations in the Acquisition of Morphology? The Frequency by Regularity Interaction in Human and PDP Learning of Morphosyntax. *Language and Cognitive Processes*. 13, pp. 307-336.

Ellis N.C. and Sinclair, S. 1996. Working Memory in the Acquisition of Vocabulary and Syntax: Putting Language in Good Order. *Quarterly Journal of Experimental Psychology*. 49(A), pp. 234-250.

Ellis, R. 2003. *Task-Based Language learning and Teaching*. Oxford: Oxford University Press.

Farmer, T.A., Misyak, J.B. and Christiansen, M.H. 2012. Individual Differences in Sentence Processing. In: Spivey, M.J., McRae, K. and Joanisse, M.F. eds. *Cambridge Handbook of Psycholinguistics*. Cambridge, UK: Cambridge University Press, pp. 353-364.

Fender, M. 2001. A Review of L1 and L2/ESL Word Integration Skills and the Nature of L2/ESL Word Integration Development Involved in Lower-level Text Processing. *Language Learning*. 51, pp. 319-396.

Field, A. and Hole, G. 2003. *How to Design and Report Experiments*. London: Sage.

Field, A., Miles, J. and Field, Z. 2012. *Discovering Statistics Using R*. London: Sage.

Field, J. 2003. *Psycholinguistics: A Resource Book for Students*. Oxon: Routledge.

Field, J. 2004. *Psycholinguistics. The Key Concepts*. London: Routledge.

Flippula, M. 1999. *The Grammar of Irish English*. London: Routledge.

Fotos, S. 1993. Consciousness Raising and Noticing through Focus on Form: Grammar Task Performance versus Formal Instruction. *Applied Linguistics*. 14, pp. 385-407.

Francis, W.S. 2005. Bilingual Semantic and Conceptual Representation. In: Kroll, J. F., de Groot, A.M.B. eds. *Handbook of Bilingualism: Psycholinguistic Approaches*. New York, NY: Oxford University Press, pp. 251-267.

Friedman, N.P. and Miyake, A. 2004. The Reading Span Test and its Predictive Power for Reading Comprehension Ability. *Journal of Memory and Language*. 51, pp. 136-158.

Gambier, Y. 2003. Introduction. Screen Adaptation: Perception and Reception. In: Gambier, Y. *Screen Translation*. Special Issue of *The Translator*. 9(2), pp. 171-189.

Gambier, Y. 2006. Multimodality and Audiovisual Translation. In: Carroll, M., Gerzymisch-Arbogast, H. and Nauert S. eds. *MuTra 2006 – Proceedings of the Marie Curie Euroconferences MuTra: Audiovisual Translation Scenarios.* Copenhagen 1-5 May 2006, pp. 1-8.

Gambier, Y. 2012. The Position of Audiovisual Translation Studies. In: Millán, C. and Bartrina, F. 2012. *The Routledge Handbook of Translation Studies*. London: Routledge, pp. 45-59.

Gambier, Y. 2016. Rapid and Radical Changes in Translation and Translation Studies. *International Journal of Communication.* 10, pp. 887-906. Available at: http://ijoc.org/index.php/ijoc/article/viewFile/3824/1570

Garza, T. 1991. Evaluating the Use of Captioned Video Materials in Advanced Foreign Language Learning. *Foreign Language Annals.* 24, pp. 239-258.

Garzanti Linguistica. [Online]. 2016. *Italian and English Dictionary.* [Accessed 24 July 2016]. Available from: http://www.garzantilinguistica.it/en/

Gass, S.M. 1988. Integrating Research Areas: A Framework for Second Language Studies. *Applied Linguistics.* 9, pp. 198-217.

Gass, S.M. 1997. *Input, Interaction, and the Second Language Learner.* Mahwah, NJ: Erlbaum.

Gass, S.M. 2003. Input and Interaction. In: Doughty, C. and Long, M. eds. *The Handbook of Second Language Acquisition.* Malden: Blackwell, pp. 224-255.

Gathercole, S.E. 2007. Working Memory and Language. In: Gaskell, M.G. ed. *The Oxford Handbook of Psycholinguistics.* Oxford: Oxford University Press, pp. 757-769.

Ghia, E. 2007. A Case Study on the Role of Interlingual Subtitles on the Acquisition of L2 Syntax - Initial Results. In: Baicchi, A. ed. *Voices on Translation. Linguistic, Multimedia, and Cognitive Perspectives RILA, Rassegna Italiana di Linguistica Applicata.* Roma: Bulzoni, pp. 167-177.

Ghia, E. 2011. The Acquisition of L2 Syntax through Audiovisual Translation. In: Şerban, A., Matamala A. and Lavaur, J.M. eds. *Audiovisual Translation in Close-up: Practical and Theoretical Approaches.* 1st ed. Bern: Peter Lang, pp. 95-112.

Ghia, E. 2012a. *Subtitling Matters: New Perspectives on Subtitling and Foreign Language Learning.* Bern: Peter Lang.

Ghia, E. 2012b. The Impact of Translation Strategies on Subtitle Reading. In: Perego, E. ed. *Eye Tracking in Audiovisual Translation.* Roma: Aracne Editrice, pp. 157-182.

Giora, R. 1999. On the Priority of Salient Meanings: Studies of Literal and Figurative Language. *Journal of Pragmatics.* 31, pp. 919-929.

Godfroid, A., Boers, F. and Housen, A. 2013. An Eye for Words. Gauging the Role of Attention in Incidental L2 Vocabulary Acquisition by Means of Eye-Tracking. *Studies in Second Language Acquisition*. 35(3), pp. 483-517.

Godfroid, A., Housen, A. and Boers, F. 2010. A Procedure for Testing the Noticing Hypothesis in the Context of Vocabulary Acquisition. In: Pütz, M. and Sicola, L. eds. *Inside the Learner's Mind: Cognitive Processing and Second Language Acquisition*. Amsterdam, Philadelphia: John Benjamins, pp. 169-197.

Godfroid, A. and Schmidtke, J. 2013. What Do Eye Movements Tell Us about Awareness? A Triangulation of Eye-movement Data, Verbal Reports and Vocabulary Learning Scores. In: Bergsleithner, J.M., Frota, S.N. and Yoshioka, J.K. eds. *Noticing and Second Language Acquisition: Studies in Honor of Richard Schmidt*. National Foreign Language Resource Center: University of Hawai'i, pp. 183-206.

Goodwin, C.J. 2008. *Research in Psychology. Methods and Designs*. 5th ed. Hoboken, NJ: John Wiley.

Göpferich, S., Lykke Jakobsen, A. and Mees, I. 2008. eds. *Looking at Eyes. Eye Tracking Studies of Reading and Translation Processing*. Copenhagen Studies in Language, 36. Copenhagen: Samfundslitteratur.

Gorman, A.M. 1961. Recognition Memory for Nouns as a Function of Abstractedness and Frequency. *Journal of Experimental Psychology*. 61, pp. 23-39.

Gottlieb, H. 1992. Subtitling: A New University Discipline. In: Dollerup, C. and Loddegaard, A. eds. *Teaching Translation and Interpreting. Training, Talent and Experience.* Amsterdam: John Benjamins, pp. 161-70.

Gottlieb, H. 2012. Subtitles – Readable Dialogue? In: Perego, E. ed. *Eye-tracking in Audiovisual Translation.* Roma: Aracne Editrice, pp. 37-81.

Goulden, R., Nation, P. and Read, J. 1990. How Large Can a Receptive Vocabulary Be? *Applied Linguistics*. 11(4), pp. 341-363.

Grabe, W. 2009. *Reading in a Second Language: Moving from Theory to Practice*. Cambridge: Cambridge University Press.

Grace-Martin, K. 2015. The Wide and Long Data Format for Repeated Measures Data. *The Analysis Factor*. [Online]. [Accessed 06 July 2015]. Available from: http://www.theanalysisfactor.com/wide-and-long-data/

Gravetter, F. and Wallnau, L. 2013. *Statistics for the Behavioral Sciences*. Belmont: Wadsworth Cengage Learning.

Grosjean, F. 1992. Another View of Bilingualism. In: Harris, R. J. ed. *Cognitive Processing in Bilinguals*. Amsterdam: Elsevier, pp. 51-62.

Guichon, N. and McLornan, S. 2008. The Effects of Multimodality on L2 Learners: Implications for CALL Resource Design. *System*. 36 (1), pp.85-93.

Guillory, H.G. 1998. The Effects of Keyword Captions to Authentic French Video on Learner Comprehension. *Calico Journal*. 15 (1/3), pp. 89-108.

Hair, J.F. Jr., Anderson, R.E., Tatham, R.L. and Black, W.C. 1995. *Multivariate Data Analysis*. 3rd ed. New York: Macmillan.

Hall, C.J. 2002. The Automatic Cognate Form Assumption: Evidence for the Parasitic Model of Vocabulary Development. *IRAL, International Review of Applied Linguistics in Language Teaching*. 40, pp. 69-87.

Hall, C.J. and Schultz, M. 1994. Los Errores de Marco Sintáctico: Evidencia del Modelo Parasitario del Léxico Mental en un Segundo Idioma. In: *Memorias del II Congreso Nacional de Lingüística*. Special Issue of *Estudios en Lingüística Aplicada*. Mexico City: UNAM, INAH and ECM, pp. 376-389.

Hamada, M. and Koda, K. 2011. Similarity and Difference in Learning L2 Word-form. *System*. 39, pp. 500-509.

Harrington, M. 2001. Sentence Processing. In: Robinson, P. ed. *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press, pp. 91-124.

Harrington, M. and Sawyer, M. 1992. L2 Working Memory Capacity and L2 Reading Skill. *Studies in Second Language Acquisition*. 14(1), pp 25-38.

Hasher, L. and Zacks, R.T. 1984. Automatic Processing of Fundamental Information: The Case of Frequency of Occurrence. *American Psychologist*. 39, pp. 1372-1388.

Hatim, B. and Mason, I. 1997. *The Translator as Communicator*. London and New York: Routledge.

Hefer, E. 2013. Reading Second Language Subtitles: A Case Study of Afrikaans Viewers Reading in Afrikaans and English. *Perspectives: Studies in Translatology*. 21(1), pp. 22-41.

Hernández Bartolomé, A.I. and Cabrera, M.G. 2005. New Trends in Audiovisual Translation: The Latest Challenging Modes. *Miscelánea: a Journal of English and American Studies*. 31: pp. 89-104.

Hervey, S.G.J., Higgins, I., Cragie, S. and Gambarotta, P. eds. 2000. *Thinking Italian Translation. A Course in Translation Method: Italian to English*. London: Routledge.

Hicks, J.L., Marsh, R.L. and Cook, G.I. 2005. An Observation on the Role of Context Variability in Free Recall. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 31, pp. 1160-1164.

Hogaboam, T.W. 1983. Reading Patterns in Eye Movement Data. In: Rayner, K. ed. *Eye Movements in Reading: Perceptual and Language Processes.* New York: Academic Press, pp. 309-332.

Holmqvist, K. 2013. Experimental Design and Eye Tracking Methodology, Eye Tracking Measures. [PowerPoint presentation]. *Tobii Eye Tracking Academy Course: Eye Tracking Masterclass for Behavioral Research.* 7/8 May, 2013, Stockholm, Sweden.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H. and van de Weijer, J. 2011. *Eye Tracking. A Comprehensive Guide to Methods and Measures*. Oxford, UK: Oxford University Press.

Holobow, N.E., Lambert, W.E. and Sayegh, L. 1984. Pairing Script and Dialogue Combinations that Show Promise for Second or Foreign Language Learning. *Language Learning.* 34, pp. 59-76.

Hosmer, D. and Lemeshow, S. 2000. *Applied Logistic Regression.* 2nd ed. New York: John Wiley & Sons, Inc.

Hulstijn, J.H. 2001. Intentional and Incidental Second Language Vocabulary Learning: a Reappraisal of Elaboration, Rehearsal and Automaticity. In: Robinson, P. ed. *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press, pp. 258-286.

Hulstijn, J.H. 2002. What Does the Impact of Frequency Tell Us about the Language Acquisition Device? *Studies in Second Language Acquisition*. 24(2), pp. 269-273.

Hummel, K.M. 1995. Translation and Second Language Learning. *Canadian Modern Language Review.* 51(3), pp. 444-455.

Incalcaterra-McLoughlin, L. and Lertola, J. 2011. Learn through Subtitling: Subtitling as an Aid to Language Learning. In: Incalcaterra McLoughlin, L., Biscio,

M. and Ní Mhainnín, M.A. eds. *Audiovisual Translation, Subtitles and Subtitling. Theory and Practice.* Bern: Peter Lang, pp. 243-263.

Inhoff, A.W. and Rayner, K. 1986. Parafoveal Word Processing during Eye Fixations in Reading: Effects of Word Frequency. *Perception and Psychophysics.* 40, pp. 431-439.

Izumi, S. and Bigelow, M. 2000. Does Output Promote Noticing and Second Language Acquisition? *TESOL Quarterly.* 34(2), pp. 239-278.

Izumi, S., Bigelow, M., Fujiwara, M., and Fearnow, S. 1999. Testing the Output Hypothesis: Effects of Output on Noticing and Second Language Acquisition. *Studies in Second Language Acquisition*. 21, pp. 421-452.

Jacob, R.J.K. and Karn, K.S. 2003. Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In: Hyönä, J., Radach, R. and Deubel, H. eds. *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research.* Amsterdam: Elsevier, pp. 573-605.

Jaeger, T.F. 2008. Categorical Data Analysis: Away from ANOVAs (Transformation or not) and Towards Logit Mixed Models. *Journal of Memory and Language.* 59, pp. 434-446.

Jarvis, S. and Odlin, T. 2000. Morphological Type, Spatial Reference, and Language Transfer. *Studies in Second Language Acquisition*. 22(4), pp. 535-556.

Jeansonne, A. 2002. *Loglinear Models.* [Online]. [Accessed 03 January 2016]. Available from: http://userwww.sfsu.edu/efc/classes/biol710/loglinear/Log%20Linear%20Models.htm

Jones, L. 2004. Testing L2 Vocabulary Recognition and Recall Using Pictorial and Written Test Items. *Language Learning and Technology.* 8(3), pp. 122-143.

Juhasz, B.J. and Rayner, K. 2003. Investigating the Effects of a Set of Intercorrelated Variables on Eye Fixation Durations in Reading. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 29(6), pp. 1312-1318.

Jung, U. 1990. The Challenge of Broadcast Videotext to Applied Linguistics. *IRAL, International Review of Applied Linguistics in Language Teaching.* 28, pp. 306-340.

Just, M.A. and Carpenter, P.A. 1976. Eye Fixations and Cognitive Processes. *Cognitive Psychology,* 8, pp. 441-480.

Just, M.A. and Carpenter, P.A. 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review*, 87, pp. 329-354.

Just, M.A. and Carpenter, P.A. 1984. Reading Skills and Skilled Reading in the Comprehension of Text. In: Mandl, H., Stein, N.L. and Trabasso, T. eds. *Learning and Comprehension of Text*. Hillsdale, NY: Lawrence Erlbaum Associates, pp. 307-329.

Kantz, D. 2015. Multimodal Subtitling – A Medical perspective. In: Gambier, Y., Caimi, A., and Mariotti, C. eds. *Subtitles and Language Learning. Principles, Strategies and Practical Experiences*. Bern: Peter Lang, pp. 269-292.

Karamitroglu, F. 1998. A Proposed Set of Subtitling Standards in Europe. *Translation Journal*. 2(2), pp. 1-15.

Keating, G.D. 2014. Eye Tracking with Text. In: Jegerski, J., and VanPatten, B. eds. *Research Methods in Second Language Psycholinguistics*. New York: Routledge, pp. 69-92.

Keijzer, M. 2013. Working Memory Capacity, Inhibitory Control and the Role of L2 Proficiency in Aging L1 Dutch Speakers of Near-Native L2 English. *Brain Science*. Special Issue of *Brain and Language*. 3(3), pp. 1261-1281.

Kellerman, E. and Sharwood Smith, M., 1986. *Crosslinguistic Influence in Second Language Acquisition*. Pergamon Press, New York.

King, J. and Just, M.A. 1991. Individual Differences in Syntactic Processing: The Role of Working Memory. *Journal of Memory and Language*. 30(5), pp. 580-602.

Koda, K. 1996. L2 Word Recognition Research: A Critical Review. *The Modern Language Journal*. 80(4), pp. 450-460.

Koda, K. 1997. Orthographic Knowledge in L2 Lexical Processing: a Cross-linguistic Perspective. In: Coady, J., Huchin, T. eds. *Second Language Vocabulary Acquisition*. Cambridge: Cambridge University Press, pp. 35-52.

Koda, K., 1999. Development of L2 Intraword Orthographic Sensitivity and Decoding Sskills. *The Modern Language Journal*. 83, pp. 51-64.

Koller, W. 1995. The Concept of Equivalence and the Object of Translation Studies. *Target*. 7(2), pp 191-222.

Koolstra, C.M. and Beentjes, J.W.J. 1999. Children's Vocabulary Acquisition in a Foreign Language through Watching Subtitled Television Programs at Home. *Educational Technology Research and Development*. 47(1), pp 51-60.

Kothari, B. 1998. Film Songs as Continuing Education: Same Language Subtitling for Literacy. *Economic and Political Weekly*. 33, pp. 2507–2510.

Kothari, B. 2008. Let a Billion Readers Bloom: Same Language Subtitling (SLS) on Television for Mass Literacy. *International Review of Education*. 54, pp. 773–778.

Kovačič, I. 1995. Reception of Subtitles. The non-Existent Ideal Viewer. In: Gambier, Y. ed. *Audiovisual Communication and Language Transfer*. International Forum, Strasburg 22-24 June 1995. Special Issue of *Translatio (FIT Newsletter)*. 14 (3-4), pp. 376-383.

Krashen, S. 1985. *The Input Hypothesis: Issues and Implications*. London and New York: Longman.

Krashen, S. 1989. We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis. *Modern Language Journal*. 73, pp. 440-464.

Kress, G. and van Leeuwen, T. 1996. *Reading Images: The Grammar of Visual Design*. London: Routledge.

Kress, G. and van Leeuwen, T. 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. London and New York: Arnold.

Kroll, J.F. and Stewart, E. 1994. Category Interference in Translation and Picture Naming. Evidence for Asymmetric Connections between Bilingual Memory Representations. *Journal of Memory and Language*. 33(2), pp. 149-174.

Kruger, J.-L. 2016. Psycholinguistics and Audiovisual Translation. *Target: International Journal of Translation Studies*. 28(2), pp. 276-287.

Kruger, J.-L. and Steyn, F. 2014. Subtitles and Eye Tracking: Reading and Performance. *Reading Research Quarterly*. 49(1), pp. 105-120.

Kruger, J.-L., Szarkowska, A. and Krejtz, I. 2015. Subtitles on the Moving Image: An Overview of Eye Tracking. In: Redmond, S. and Batty, K. eds. Eye-Tracking the Moving Image. Special Issue of *Refractory: A Journal of Entertainment Media*. Volume 25. [Online]. [Accessed 12 October 2016]. Available from: http://refractory.unimelb.edu.au/2015/02/07/kruger-szarkowska-krejtz/

Kučera, H. and Francis, W. N. 1967. *Computational Analysis of Present-day American English*. Providence, Rhode Island: Brown University Press.

Kuhn, J. 2012. The Noticing of Correct and Incorrect Forms in Lengthier Texts: An ESL Eye-tracking Investigation. In: Kessler, G., Oscoz, A. and Elola, I. eds. *Technology Across Writing Contexts and Tasks*. San Marcos, Texas: CALICO, pp. 237-253.

Kuppens, A.H. 2010. Incidental Foreign Language Acquisition from Media Exposure. *Learning, Media and Technology*. 35(1), pp. 65-85.

La Storia del Bingo. 2016. *Origini del Bingo*. [Online]. [Accessed 12 July 2013]. Available from: http://www.bingo.org/it/articoli/la-storia-del-bingo

Lado, R. 1957. *Linguistics across Cultures: Applied Linguistics for Language Teachers*. Ann Arbor: University of Michigan Press.

Lambert, W.E., Boehler, I. and Sidoti, N. 1981. Choosing the Languages of Subtitles and Spoken Dialogues for Media Presentations: Implications for Second Language Education. *Applied Psycholinguistics*. 2, pp. 133-144.

Lambert, W.E. and Holobow, N.E. 1984. Combinations of Printed Script and Spoken Dialogues that Show Promise for Beginning Students of a Foreign Language. *Canadian Journal of Behavioural Science*. 16, pp. 1-11. Available from: DOI:10.1037/h0080775

Larson-Hall, J. 2010. *A Guide to Doing Statistics in Second Language Research Using SPSS*. New York: Routledge.

Laufer, B. 1989. A Factor of Difficulty in Vocabulary Learning: Deceptive Transparency. In: Nation, P. and Carter, R. eds. *Vocabulary Acquisition*. *Aila Review – Revue De L'Aila*. 6, pp. 10-17.

Laufer, B. 1997. The Lexical Plight in Second Language Reading. Words You don't Know, Words You Think You Know, and Words you can't Guess. In: Coady, J. and Huckin, T. eds. *Second Language Vocabulary Acquisition*. Cambridge: Cambridge University Press, pp. 20-34.

Laufer, B. 2000. Avoidance of Idioms in a Second Language: The Effect of L1-L2 Degree of Similarity. *Studia Linguistica*. 54(2), pp. 186-196.

Laufer, B. 2005. Focus on Form in Second Language Vocabulary Learning. In: Foster-Cohen, S.H., García Mayo, M.d.P., and Cenoz, J. eds. *EUROSLA Yearbook 5*. Amsterdam: John Benjamins, pp. 223-250.

Laufer, B. 2006. Comparing Focus on Form and Focus on Forms in Second-Language Vocabulary Learning. *Canadian Modern Language Review*. 63(1), pp. 149-166.

Laufer, B. and Girsai, N. 2008. Form-focused Instruction in Second Language Vocabulary Learning: A Case for Contrastive Analysis and Translation. *Applied Linguistics*. 29(4), pp. 694-716.

Lavaur, J. and Bairstow, D. 2011. Languages on the Screen: Is Film Comprehension Related to the Viewers' Fluency level and to the Language in the Subtitles? *International Journal of Psychology*. 46(6), pp. 455-462.

Leckie, G. 2013. *Cross-Classified Multilevel Models*. LEMMA (Learning Environment for Multilevel Methodology and Applications) VLE Module 12, pp. 1-60. Bristol Centre for Multilevel Modelling. [Online]. [Accessed 28 February 2016]. Available via registration from: http://www.bristol.ac.uk/cmm/learning/online-course/index.html

Lee, D. 2001. *Cognitive Linguistics: an Introduction*. Oxford: Oxford University Press.

Leonardi, V. 2010. *The Role of Pedagogical Translation in Second Language Acquisition: From Theory to Practice*. Bern: Peter Lang.

Leonardi, V. 2011. Pedagogical Translation as a Naturally-Occurring Cognitive and Linguistic Activity in Foreign Language Learning. *Annali Online di Lettere – Ferrara*. 1-2, pp. 17-28.

Leow, R.P. 2013a. Attention, Noticing, and Awareness in Second Language Acquisition. In: Chapelle, C.A. ed. *The Encyclopedia of Applied Linguistics*. Chichester, West Sussex, UK: Wiley-Blackwell Publishing. Vol. I, pp. 269-275.

Leow, R.P. 2013b. Schmidt's Noticing Hypothesis: More than Two Decades after. In: Bergsleithner, J.M., Frota, S.N. and Yoshioka, J.K. eds. *Noticing and Second Language Acquisition: Studies in Honor of Richard Schmidt*. Honolulu: University of Hawai'i, National Foreign Language Resource Center, pp. 11-23.

Leow, R.P. and Bowles, M.A. 2005. Attention and Awareness in SLA. In: Sanz, C. ed. *Mind and Context in Adult Second Language Acquisition: Methods, Theory and Practice*. Washington, DC: Georgetown University Press, pp. 179-203.

Lertola, J. 2012. The Effect of the Subtitling Task on Vocabulary Learning. In: Pym, A. and Orrego-Carmona, D. eds. *Translation Research Project 4*. Intercultural Studies Group: Tarragona, pp. 61-70.

Lertola, J. 2015. Subtitling in Language Teaching: Suggestions for Language Teachers. In: Gambier, Y., Caimi, A. and Mariotti, C. eds. *Subtitles and Language Learning. Principles, Strategies and Practical Experiences*. Bern: Peter Lang, pp. 245-267.

Lightbown, P.M. and Spada, N. 1990. Focus on Form and Corrective Feedback in Communicative Language Teaching: Effects on Second Language Learning. *Studies in Second Language Acquisition*. 12(4), pp. 429-448.

Locker, L., Hoffman, L. and Bovaird, J. 2007. On the Use of Multilevel Modeling as an Alternative to Items Analysis in Psycholinguistic Research. *Behavior Research Methods*. 39, pp. 723-730.

Lohnas, L.J. and Kahana, M.J. 2013. Parametric Effects of Word Frequency Effect in Memory for Mixed Frequency Lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 39(6), pp. 1943-1946.

Lonergan, J. 1989. *Video in Language Teaching.* Cambridge: Cambridge University Press.

Long, J.S. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage Publications.

Lopriore, L. and Ceruti, M.A. 2015. Subtitling and Language Awareness: A Way and Ways. In: Gambier, Y., Caimi, A. and Mariotti, C. eds. *Subtitles and Language Learning. Principles, Strategies and Practical Experiences*. Bern: Peter Lang, pp. 293–321.

Mackey, A. and Gass, S. 2005. *Second Language Research: Methodology and Design.* Mahwah, NJ: Erlbaum.

Makel, M.C. and Plucker, J.A. 2014. Facts Are More Important Than Novelty. Replication in the Education Sciences. *Educational Researcher*. 43(6), pp. 304-316.

Malmkajær, K. 2004. ed. *Translation in Undergraduate Degree Programs*. Amesterdam, Philadelphia, PA: John Benjamins.

Mandera, P., Keuleers, E., Wodniecka, Z. and Brysbaert, M. 2015. SUBTLEX-PL: Subtitle-based Word Frequency Estimates for Polish. *Behavior Research Methods*. 47(2), pp. 471-483.

Mandler, G., Goodman, G.O. and Wilkes-Gibbs, D.L. 1982. The Word-Frequency Paradox in Recognition. *Memory and Cognition*. 10(1), pp. 33-42.

Marian, V., Bartolotti, J., Chabal, S. and Shook, A. 2012. CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS One*. 7(8). [Online]. [Accessed 02 October 2016]. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0043230

Marian, V. and Blumenfeld, H.K. 2006. Phonological Neighborhood Density in Native and Non-native Word Production. *Journal of Social and Ecological Boundaries*. 2(1), pp. 3-35.

Mariotti, C. 2002. Dall'Immagine al sottotitolo: strategie d'uso dei materiali audiovisivi nell'insegnamento dell'inglese L2. In: Caimi, A. ed. *Cinema, Paradiso delle Lingue. I Sottotitoli nell'Apprendimento Linguistico*. Special Issue of *RILA, Rassegna Italiana di Linguistica Applicata*. 34(1-2), pp. 207-213.

Markham, P. 1999. Captioned Videotapes and Second Language Listening Word Recognition. *Foreign Language Annal.* 32, pp. 321-328.

McConkie, G.W. and Rayner K. 1975. The Span of the Effective Stimulus During a Fixation in Reading. *Perception and Psychophysics.* 17, pp. 578-586.

McDaniel, M.A. and Mason, M.E.J. 1985. Altering Memory Representations through Retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 11, pp. 371-385.

McDonald, J.H. 2014. *Handbook of Biological Statistics*. 3rd ed. Sparky House Publishing: Baltimore, Maryland.

McLaughlin, B., Rossman, T. and McLeod, B. 1983. Second Language Learning: An Information-Processing Perspective. *Language Learning*. 33(2), pp. 135-158.

McLeod, S. 2013. Stages of Memory Encoding Storage and Retrieval. *Simply Psychology*. [Online]. [Accessed 10 August 2016]. Available from: http://www.simplypsychology.org/memory.html

Menard, S. 1991. *Longitudinal Research*. Newbury Park: Sage Publications.

Menard, S. 2002. *Applied Logistic Regression Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences; no. 07-106. Thousand Oaks, CA: Sage Publications.

Mitterer, H. and McQueen, J.M. 2009. Foreign Subtitles Help but Native-Language Subtitles Harm Foreign Speech Perception. *PLoS ONE*. 4(11), pp. e7785. [Online].

[Accessed 10 August 2016]. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007785

Montero Perez, M., Peters, E. and Desmet, P. 2014. Effects of Captioning on Video Comprehension and Incidental Vocabulary Learning. *Language Learning and Technology*. 18(1), pp. 118-141.

Moran, S. 2008. *The Effect of Linguistic Variation on Subtitle Reception*. M.A. Thesis: York University.

Moran, S. 2012. The Effect of Linguistic Variation on Subtitle Reception. In: Perego, E. ed. *Eye-Tracking in Audiovisual Translation*. Roma: Aracne Editrice, pp. 183-222.

Morris, R.K. 1994. Lexical and Message-level Sentence Context Effects on Fixation Times in Reading. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 20(1), pp. 92-103.

Morton, J. 1986. Interaction of Information in Word Recognition. *Psychological Review*. 76(2), pp. 165-178.

Muljani, D., Koda, K. and Moates, D.R. 1998. The Development of Word Recognition in a Second Language. *Applied Psycholinguistics*. 19, pp. 99-113.

Munday, J. 2008. *Introducing Translation Studies: Theories and Applications*. Abingdon and New York: Routledge.

Munday, J. 2009. *The Routledge Companion to Translation Studies*. London: Routledge.

Myers, R. 1990. *Classical and Modern Regression with Applications*. 2nd ed. Boston: Duxbury.

Nagy, W.E., Herman, P.A. and Anderson, R.C. 1985. Learning Words from Context. *Reading Research Quarterly*. 20, pp. 233-253.

Nation, P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Neter, J., Wasserman, W. and Kutner, M.H. 1989. *Applied Linear Regression Models*. Homewood, Illinois: Irwin.

Neuman, S. and Koskinen, P. 1990. *Using Captioned Television to Improve the Reading Proficiency of Language Minority Students*. National Captioning Institute, Falls Church, VA, USA.

Neves, J. 2004. Language Awareness through Training in Subtitling. In: Orero, P. ed. *Topics in Audiovisual Translation*. Amsterdam & Philadelphia: John Benjamins, pp. 127-140.

Newmark, P. 1981. *Approaches to Translation*. Oxford: Pergamon Press.

Nida, E. 1964. *Towards a Science of Translating*. Leiden: Brill

Nida, E.A. and de Waard, J. 1986. *From One Language to Another: Functional Equivalence in Bible Translation*. Nashville: Thomas Nelson Publishers.

Nida, E.A. and Taber, C.R. 1982. *The Theory and Practice of Translation*. Leiden: E. J. Brill.

Norman, D.A. 1968. Toward a Theory of Memory and Attention. *Psychological Review,* 75, pp. 522-536.

Norris, D. 2009. Putting it All Together: A Unified Account of Word Recognition and Reaction-time Distributions. *Psychological Review*. 116(1), pp. 207-219.

O'Brien, S. 2006. Eye Tracking and Translation Memory Matches. *Perspectives. Studies in Translatology*. 14(3), pp. 185-205.

O'Connell, E. 2011. Formal and Casual Language Learning: What Subtitles Have to Offer Minority Languages Like Irish. In: Incalcaterra McLoughlin, L., Biscio, M. and Ní Mhainnín, M.A. eds. *Audiovisual Translation, Subtitles and Subtitling. Theory and Practice*. Bern: Peter Lang, pp. 157-175.

Odlin, T. 1989. *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.

Odlin, T. 2003. Cross-linguistic Influence. In: Doughty, C. and Long, M. eds. *Handbook of Second Language Acquisition*. Oxford: Blackwell, pp. 436-486.

OED. *Oxford English Dictionary*. [Online]. 2016. [Accessed 24 June 2016]. Available from: http://0-www.oed.com.wam.leeds.ac.uk/

Ofcom. 2015. *Guidance on Standards for Subtitling*. Ofcom, Independent regulator and competition authority for the UK communications industries. [Online]. [Accessed 03 February 2015]. Available from: http://www.ofcom.org.uk/static/archive/itc/itc_publications/codes_guidance/standards_for_subtitling/subtitling_1.asp.html

Orne, M.T. 1962. On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and their Implications. *American Psychologist*. 17, pp. 776-783.

Orrego-Carmona, D. 2016. A Reception Study on Non-professional Subtitling. Do Audiences Notice Any Difference? *Across Languages and Cultures*. 17(2), pp. 163-181.

Osaka, M. and Osaka, N. 1992. Language-Independent Working Memory as Measured by Japanese and English Reading Span Tests. *Bulletin of the Psychonomic Society*. 30, pp. 287-289.

Paivio, A. 1986. *Mental Representation: A Dual-coding Approach*. New York: Oxford University Press.

Paivio A. and Csapo, K. 1973. Picture Superiority in Free Recall: Imagery or Dual Coding? *Cognitive Psychology*. 5, pp. 176-206.

Paivio, A. and Desrochers, A. 1980. A Dual-coding Approach to Bilingual Memory. *Canadian Journal of Psychology*. 34(4), pp. 388-399.

Paivio, A. and Lambert, W. 1981. Dual Coding and Bilingual Memory. *Journal of Verbal Learning and Verbal Behavior*. 20, pp. 532-539.

Pan, Y. and Jackson, R.T. 2008. Ethnic Difference in the Relationship between Acute Inflammation and Serum Ferritin in US Adult Males. *Epidemiology and Infection*. 136(3), pp. 421-431.

Paterson K.B., McGowan, V.A. and Jordan, T.R. 2012. Eye Movements Reveal Effects of Visual Content on Eye Guidance and Lexical Access during Reading. *PLoS One*, 7(8). [Online]. [Accessed 10 February 2016]. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0041766

Paulson, E.J. 2000. *Adult Readers' Eye Movements During the Production of Oral Miscues*. Unpublished Ph.D. thesis, University of Arizona, Tucson.

Paulson, E.J. 2002. Are Oral Reading Word Omissions and Substitutions Caused by Careless Eye Movements? *Reading Psychology*, 23, pp. 45-66.

Pavakanun, U. 1992. *Incidental Acquisition of Foreign Language through Subtitled Television Programs as a Function of Similarity with Native Language and as a Function of Presentation Mode*. Ph.D. thesis, University of Leuven, Belgium.

Pavakanun, U. and d'Ydewalle, G. 1992. Watching Foreign Television Programs and Language Learning. In: Engel, F.L., Bouwhuis, D.G., Bosser, T. and d'Ydewalle, G. eds. *Cognitive Modelling and Interactive Environments in Language Learning*. Berlin: Springer. pp. 193-198.

Pavesi, M. 2002. Sottotitoli: dalla Semplificazione nella Traduzione all'Apprendimento Linguistico. In: Caimi, A. ed. *Cinema, Paradiso delle Lingue. I Sottotitoli nell'Apprendimento Linguistico*. Special Issue of *RILA, Rassegna Italiana di Linguistica Applicata*. 34(1-2), pp. 127-142.

Pavesi, M. and Perego, E. 2008a. Tailor-made Interlingual Subtitling as a Means to Enhance Second Language Acquisition. In: Díaz Cintas, J. ed. *The Didactics of Audiovisual Translation*. Amsterdam: John Benjamins, pp. 215-225.

Pavesi, M. and Perego, E. 2008b. Profiling Film Translators in Italy: A Preliminary Analysis. *JoSTrans, The Journal of Specialised Translation*, 6. [Online]. [Accessed 10 February 2016]. Available from: http://www.jostrans.org/issue06/art_pavesi.php

Peng, J.C.-Y., Lee, K.L. and Ingersoll, G.M. 2002. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*. 96(1), pp. 3-14.

Perego, E. 2008. Subtitles and Line-breaks: Towards Improved Readability. In: Chiaro, D., Heiss, C., and Bucaria, C. eds. *Between Text and Image. Updating Research in Screen Translation*. John Benjamins, pp. 211-223.

Perego, E. 2012. Introduction. In: Perego, E. ed. *Eye-Tracking in Audiovisual Translation*. Roma: Aracne Editrice, pp. 7-11.

Perego, E., Del Missier, F. and Porta, M. 2010. The Cognitive Effectiveness of Subtitle Processing. *Media Psychology*, 13, pp. 243-272.

Perego, E. and Ghia, E. 2011. Subtitle Consumption According to Eye Tracking Data. An Acquisitional Perspective. In: Incalcaterra McLoughlin, L, Biscio, M. and Ní Mhainnín, M.A. eds. *Audiovisual Translation Subtitles and Subtitling. Theory and Practice*. Bern: Peter Lang, pp. 177-196.

Pitts, M., White, H. and Krashen, S.D. 1989. Acquiring Second Language Vocabulary through Reading: A Replication of the Clockwork Orange Study Using Second Language Acquirers. *Reading in a Foreign Language*. 5(2), pp. 271-275.

Polio, C. and Gass, S. 1997. Replication and Reporting: A Commentary. *Studies in Second Language Acquisition*. 19, pp. 499-508.

Pollatsek, A., Juhasz, B.J., Reichle, E.D., Machacek, D. and Rayner, K. 2008. Immediate and Delayed Effects of Word Frequency and Word Length on Eye Movements in Reading: A Reversed Delayed Effect of Word Length. *Journal of Experimental Psychology: Human Perception and Performance*. 34(3): pp. 726-750.

Porte, G.K. 2002. *Appraising Research in Second Language Learning: A Practical Approach to Critical Analysis of Quantitative Research*. Philadephia: PA John Benjamins.

Posner, M.I. 1980. Orienting of Attention. *Quarterly Journal of Experimental Psychology*. 32, pp. 3-25.

Price, K. 1983. Closed-captioned TV: An Untapped Resource. *MATSOL Newsletter,* 12(2).

Purves, D., Augustine, G.J., Fitzpatrick, D., Katz, L.C., LaMantia, A.-S., McNamara, J.O. and Williams, S.M. 2001. eds. *Neuroscience*. 2nd ed. Sunderland (MA): Sinauer Associates.

Py, B. 2004. L'Apprenant, le Chercheur et les Discours: Quelques Réflexions sur la Notion de Saillance. *Tranel*. 40, pp. 117-131.

Pylkkänen, L. 2016. *Lexical Access*. Lecture Slides. [Online]. [Accessed 10 June 2013]. Available from: http://www.psych.nyu.edu/pylkkanen/Neural_Bases/07_slides/11_LexAccess_Elect.pdf

Quené, H. and van den Bergh, H. 2004. On Multi-level Modeling of Data from Repeated Measures Designs: a Tutorial. *Speech Communication*. 43, p. 103-121.

Quené, H. and van den Bergh, H. 2008. Examples of Mixed-effects Modelling with Crossed Random Effects and with Binomial Data. *Journal of Memory and Language*. 59(4), pp. 413-425.

R Development Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/.

R version 3.2.0. 2015. Version "Full of Ingredients", copyright (c) 2015. *The R Foundation for Statistical Computing* (2015-04-16), Vienna, Austria. Available from: http://www.r-project.org

R version 3.4.0. 2017. Version "You Stupid Darkness", copyright (c) 2017. *The R Foundation for Statistical Computing* (2017-04-21), Vienna, Austria. Available from: http://www.r-project.org

Ragni, V. 2018. [Forthcoming]. Didactic Subtitling in the Foreign Language (FL) Classroom. Improving Language Skills through Task-Based Practice and Form-Focused Instruction (FFI): Background Considerations. *Translation and Translanguaging in Multilingual Contexts*. 4(1).

Rajendran, D.J., Duchowski, A.T., Orero, P., Martínez, J. and Romero-Fresco, P. 2013. Effects of Text Chunking on Subtitling: A Quantitative and Qualitative Examination. *Perspectives: Studies in Translatology*. 21(1), pp. 5-21.

Ransdell, S.E. and Fischler, I. 1987. Memory in a Monolingual Mode: When are Bilinguals at a Disadvantage? *Journal of Memory and Language*. 26(4), pp. 392-405.

Rayner, K. 1986. Eye Movements and the Perceptual Span in Band Skilled Readers. *Journal of Experimental Child Psychology*. 41(2), pp. 211-236.

Rayner, K. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*. 124, pp. 372-422.

Rayner, K. and Duffy, S.A. 1986. Lexical Complexity and Fixation Times in Reading: Effects of Word Frequency, Verb Complexity, and Lexical Ambiguity. *Memory and Cognition*. 14(3), pp. 191-201.

Rayner, K. and Fischer, M.H. 1996. Mindless Reading Revisited: Eye Movements During Reading and Scanning are Different. *Perception and Psychophysics*. 58, pp. 734-747.

Rayner, K. and Pollatsek, A. 1989. *The Psychology of Reading*. Englewood Cliffs, NJ: Prentice Hall.

Rayner, K. and Pollatsek, A. 2006. Eye-Movement Control in Reading. In: Traxler, M.J., and Gernsbacher, M.A. eds. *Handbook of Psycholinguistics*. 2nd ed. Burlington, MA: Elsevier, pp. 613-657.

Rayner, K., Reichle, E.D., Stroud, M.J., Williams, C.C. and Pollatsek, A. 2006. The Effect of Word Frequency, Word Predictability, and Font Difficulty on the Eye Movements of Young and Older Readers. *Psychology and Aging*. 21(3), pp. 448-465.

Rayner, K., Sereno, S.C., Morris, R.K., Schmauder, A.R. and Clifton, C. 1989. Eye Movements and On-line Language Comprehension Processes. Special issue of *Language and Cognitive Processes*. 4(3-4), pp. 21-49.

Rayner K. and Well, A.D. 1996. Effects of Contextual Constraint on Eye Movements in Reading: A Further Examination. *Psychonomic Bulletin and Review.* 3(4), pp. 504-509.

Reichle, E.D., Pollatsek, A., Fisher, D.L. and Rayner, K. 1998. Toward a Model of Eye Movement Control in Reading. *Psychological Review*. 105, pp. 125-157.

Reichle, E.D., Pollatsek, A. and Rayner, K. 2007. Modeling the Effects of Lexical Ambiguity on Eye Movements during Reading. In: van Gompel, R.P.G., Fischer, M.H., Murray, W.S., Hill, R.L. eds. *Eye Movements: Window on Mind and Brain*. Amsterdam: Elsevier, pp. 271-292.

Reichle, E.D., Reineberg, A.E. and Schooler, J.W. 2010. Eye Movements During Mindless Reading. *Psychological Science*. 21(9), pp. 1300-1310.

Reisberg, D. 2013. *Cognition: Exploring the Science of the Mind.* 5th ed. New York: W. W. Norton & Company.

Ricci, V. 2005. *Fitting Distributions with R*. [Online]. [Accessed 11 September 2015]. Available from: https://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf

Robinson, P. 1995. Attention, Memory, and the Noticing Hypothesis. *Language Learning,* 45, pp. 283-331.

Robinson, P. 1996. Learning Simple and Complex Second Language Rules under Implicit, Incidental, Rule-search, and Instructed Conditions. *Studies in Second Language Acquisition*. 18(1), pp. 27-67.

Robinson, P. 1997a. Generalizability and Automaticity of Second Language Learning under Implicit, Incidental, Enhanced, and Instructed conditions. *Studies in Second Language Acquisition,* 19(2), pp. 223-247.

Robinson, P. 1997b. Individual Differences and Fundamental Similarity of Implicit and Explicit Adult Second Language Learning. *Language Learning,* 47(1), pp. 45-99.

Rogerson, P.A. 2001. *Statistical Methods for Geography*. London: Sage.

Rosenthal, R. 1963. On the Social Psychology of the Psychological Experiment: The Experimenter's Hypothesis as Unintended Determinant of Experimental Results. *American Scientist*. 51(2), pp. 268-283.

Rott, S. 1999. The Effect of Exposure Frequency on Intermediate Language Learners' Incidental Vocabulary Acquisition and Retention through Reading. *Studies in Second Language Acquisition*. 21, pp. 589-619.

RStudio, 2016. *Take control of your R code*. Product page. [Online]. [Accessed 10 May 2016]. Available from: https://www.rstudio.com/products/RStudio/

Sabatini-Coletti. 2016. *Dizionario della Lingua Italiana*. [Online] [Accessed 24 June 2016]. Available from: http://dizionari.corriere.it/dizionario_italiano/

Sansoni. 2016. *Dizionario di Inglese (English-Italian / Italiano-Inglese)*. [Online] [Accessed 16 March 2017]. Available from: http://dizionari.corriere.it/dizionario_inglese/

Scarborough, D.L., Cortese, C. and Scarborough, H.S. 1977. Frequency and Repetition Effects in Lexical Memory. *Journal of Experimental Psychology: Human Perception and Performance*. 5, pp. 1-17.

Scheffler, P. 2012. Theories pass. Learners and Teachers Remain. *Applied Linguistics,* 33(5), pp. 603-607.

Schmauder, A.R., Morris, R.K. and Poynor, D.V. 2000. Lexical Processing and Text Integration of Function and Content Words: Evidence from Priming and Eye Fixations. *Memory and Cognition*. 28(7), pp. 1098-1108.

Schmidt, R. 1990. The Role of Consciousness in Second Language Learning. *Applied Linguistics*. 11, pp. 129-158.

Schmidt, R. 1995. Consciousness and Foreign Language Learning: A Tutorial on the Role of Attention and Awareness in Learning. In: Schmidt, R. ed. *Attention and Aawareness in Foreign Language Learning*. Honolulu: University of Hawaii Press, pp. 1-63.

Schmidt, R. 2001. Attention. In: Robinson, P. ed. *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press, pp. 3-32.

Schotter, E.R., Angele, B. and Rayner, K., 2012. Parafoveal Processing in Reading. *Attention Perception and Psychophysics*. 74 (1), pp. 5-35.

Schotter, E.R. and Rayner, K., 2012. Eye Movements in Reading. Implications for Reading Subtitles. In: Perego, E. ed. *Eye Tracking in Audiovisual Translation*. Roma: Aracne Editrice, pp. 83-104.

Secară, A. 2011. R U Ready 4 new Subtitles? Investigating the Potential of Social Translation Practices and Creative Spellings. In: Remael, A. and O'Hagan, M. eds. *Linguistica Antverpiensia*, *Volume 10*. Antwerp: Antwerp University Press, pp. 153-174.

Seilhamer, M. F. 2010. Concept Stretching and Model Merging. An Attempt to Better Account for L2 Processing and Acquisition of Grammatical Constructions. In: Pütz, M. and Sicola, L. eds. *Cognitive Processing in Second Language Acquisition: Inside the Learner's Mind.* Amsterdam: John Benjamins, pp. 9-26.

Shah, P. and Miyake, A. 1999. Models of Working Memory: An Introduction. In: Miyake, A. and Shah, P. eds. *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control.* New York: Cambridge University Press, pp. 1-27.

Sheen, R. 1996. The Advantage of Exploiting Contrastive Analysis in Teaching and Learning a Foreign Language. *IRAL: International Review of Applied Linguistics in Language Teaching.* 34(3), pp. 183-198.

Sherman, J. 2003. *Using Authentic Video in the Language Classroom*. Cambridge University Press.

Simon, H.A. 1962. The Architecture of Complexity. *Proceedings of the American Philosophical Society.* 106, pp. 467-482.

Smith, B. 2010. Employing Eye-tracking Technology in Researching the Effectiveness of Recasts in CMC. In: Hult, F.M. ed. *Directions and Prospects for Educational Linguistics*. New York: Springer, pp. 79–98.

Smith, B. 2012. Eye Tracking as a Measure of Noticing. *Language Learning and Technology*. 16(3), pp. 53-81.

Smith, F. 2004. *Understanding reading: A Psycholinguistic Analysis of Reading and Learning to Read*. London: Routledge.

Smith, J.J. 1990. *Closed-captioned Television and Adult Students of English as a Second Language.* Arlington, VA: Arlington Refugee Education and Employment Program.

Snell-Hornby, M. 1988. *Translation Studies: An Integrated Approach*. Amsterdam and Philadelphia: John Benjamins.

Sokoli, S. 2006. Learning via Subtitling (LvS): A Tool for the Creation of Foreign Language Learning Activities Based on Film Subtitling. In: Carroll, M. and Gerzymisch-Arbogast, H. eds. *Proceedings of the Marie Curie Euroconferences MuTra: Audiovisual Translation Scenario*. Copenhagen, 1-5 May, pp. 66-73.

Sokoli, S. 2015. ClipFlair: Foreign Language Learning through Interactive Revoicing and Captioning of Clips. In: Gambier, Y., Caimi, A., and Mariotti, C. eds. *Subtitles and Language Learning. Principles, Strategies and Practical Experiences*. Bern: Peter Lang, pp. 127-147.

Spada, N. and Lightbown, P.M. 1993. Instruction and the Development of Questions in L2 Classrooms. *Studies in Second Language Acquisition*. 15(2), pp. 205-221.

Spada, N., Lightbown, P.M. and White, J. 2005. The Importance of Form/Meaning Mappings in Explicit Form-Focused Instruction. In: Housen, A. and Pierrard, M. eds. *Investigations in Instructed Second Language Acquisition*. Brussels: Mouton De Gruyter, pp. 199-234.

Spadiliero, E. 2015. Perché si Dice 'Essere in Gamba'? *Sul Romanzo*. [Online]. [Accessed 10 July 2016]. Available from: http://www.sulromanzo.it/blog/perche-si-dice-essere-in-gamba

Stangroom, J. 2015. *Effect Size Calculator for T-test*. [Online]. [Accessed 06 December 2015]. Available from: http://www.socscistatistics.com/effectsize/Default3.aspx

Staub, A. and Rayner, K. 2007. Eye Movements and On-line Comprehension Processes. In: Gaskell, M.G. ed. *The Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press, pp. 327-342.

Steinman, R.M., Haddad, G.M., Skavenski, A.A. and Wyman, D. 1973. Miniature Eye Movement. *Science*. 181(4102), pp. 810-819.

Sullivan, G.M. and Feinn, R. 2012. Using Effect Size – or Why the P-value Is Not Enough. *Journal of Graduate Medical Education*. 4(3), pp. 279–282. [Online]. [Accessed 06 December 2015]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/

Swain, M. 1985. Communicative Competence: Some Roles of Comprehensible Input and Comprehensible Output in its Development. In: Gass, S. and Madden, C. eds. *Input in Second Language Acquisition*. New York: Newbury House, pp. 235-256.

Szmalec, A., Brysbaert, M. and Duyck, W. 2013. Working Memory and (Second) Language Processing. In: Altarriba, J. and Isurin, L. eds. *Memory, Language, and Bilingualism, Theoretical and Applied Approaches*. Cambridge: Cambridge University Press, pp. 74-94.

Taft, M. 1993. *Reading and the Mental Lexicon*. Hove: Erlbaum.

Talamas, A., Kroll, J.F. and Dufour, R. 1999. From Form to Meaning: Stages in the Acquisition of Second-Language Vocabulary. *Bilingualism: Language and Cognition*. 2(1), pp. 45-58.

Talaván, N. 2010. Subtitling as a Task and Subtitles as Support: Pedagogical Applications. In: Díaz-Cintas, J., Matamala, A. and Neves, J. eds. *New Insights into Audiovisual Translation and Media Accessibility*. Amsterdam: Rodopi, pp. 285-299.

Talaván, N. 2011. A Quasi-experimental Research Project on Subtitling and Foreign Language Acquisition. In: Incalcaterra McLoughlin, L., Biscio, M. and Ní Mhainnín, M.A. eds. *Audiovisual Translation, Subtitles and Subtitling. Theory and Practice.* Bern: Peter Lang, pp. 197-218.

Talaván, N. and Ávila-Cabrera, J.J. 2015. First Insights into the Combination of Dubbing and Subtitling as L2 Didactic Tools. In: Gambier, Y., Caimi, A. and Mariotti, C. eds. *Subtitles and Language Learning. Principles, Strategies and Practical Experiences*. Bern: Peter Lang, pp. 149-172.

Talaván, N. and Rodríguez-Arancón, P. 2014. The Use of Reverse Subtitling as an Online Collaborative Language Learning Tool. *The Interpreter and Translator Trainer*. 8(1), pp. 84-101.

Tanenhaus, M.K. 2007. Eye Movements and Spoken Language Processing. In: van Gompel, R.P.G., Fischer, M.H., Murray, W.S., Hill, R.L. eds. *Eye Movements: Window on Mind and Brain*. Amsterdam: Elsevier, pp. 309-326.

Taylor, G. 2005. Perceived Processing Strategies of Students Watching Captioned Video. *Foreign Language Annals*. 38(3), pp. 422-427.

Thorne, G. 2003. What Strategies Can Be Used to Increase Memory? *Center for Development and Learning (CDL)*. [Online]. [Accessed 20 August 2016]. Available from: http://www.cdl.org/articles/what-strategies-can-be-used-to-increase-memory/

Tobii. 2010. *Timing Guide for Tobii Eye Trackers and Eye Tracking Software*. Tobii® Technology AB. White Paper. [Online]. [Accessed 6 July 2015]. Available from:

http://www.tobiipro.com/siteassets/tobii-pro/learn-and-support/design/eye-tracker-timing-performance/tobii-eye-tracking-timing.pdf/?v=1.0

Tobii Pro Global. 2015. *What Does the Sample Rate Percentage in the Replay Tab Represent?* [Online]. [Accessed 6 July 2015]. Available from: http://www.tobii.com/en/eye-tracking-research/global/support-and-downloads/faqs/50130000000a2ka/

Tomlin, R. and Villa, H. 1994. Attention in Cognitive Science and Second Language Acquisition. *Studies in Second Language Acquisition*. 16, pp. 183-203.

Torf, P. and Brauer, C. 2014. *A (Very) Short Introduction to R*. [Online]. [Accessed 5 July 2015]. Available from: https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

Truscott, J. 1998. Noticing in Second Language Acquisition: a Critical Review. *Second Language Research*. 14(2), pp. 103-135.

Truscott, J. and Sharwood Smith, M. 2011. Input, Intake, and Consciousness: The Quest for a Theoretical Foundation. *Studies in Second Language Acquisition*. 33(4), pp. 497-528.

Tschirner, E. 2001. Language Acquisition in the Classroom: The Role of Digital Video. *Computer Assisted Language Learning*. 14, pp. 305–319.

Tuominen, T. 2012. Accidental Reading? Some Observations on the Reception of Subtitled Films. In: Şerban, A., Matamala, A. and Lavaur, J.-M. eds. *Audiovisual Translation in Close-up: Practical and Theoretical Approaches*. 2nd ed. Bern: Peter Lang, pp. 189-204.

Turner, H. 2008. *Introduction to Generalized Linear Models*. Lecture Notes from the 'Introduction To Generalized Linear Models' course, 22-24 April, Institute for Statistics and Mathematics, WU Wirtschaftsuniversität Wien (University of Economics and Business, Vienna). [Online]. [Accessed 02 March 2016]. Available from: http://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf

Turner, M.L. and Engle, R.W. 1989. Is Working Memory Capacity Task Dependent? *Journal of Memory and Language*. 28, pp. 127-154.

Uggen, M.S. 2012. Reinvestigating the Noticing Function of Output. *Language Learning*. 62(2), pp. 506-540.

UK Copyright Service. 2017. *Fact sheet P-27: Using the work of others*. Updated on 6th February 2017. [Online]. [Accessed 02 May 2017]. Available from: https://www.copyrightservice.co.uk/copyright/p27_work_of_others

Van Assche, E., Duyck, W, Hartsuiker, R.J. and Diependaele, K. 2009. Does Bilingualism Change Native-Language Reading? Cognate Effects in a Sentence Context. *Psychological Science,* 20(8), pp. 923-927.

Van de Poel, M. and d'Ydewalle, G. 1999. Incidental Foreign-Language Acquisition by Children Watching Subtitled Television Programs. *Journal of Psycholinguistic Research*. 28, pp. 227-244. Reprinted in: Gambier, Y and Gottlieb, H. eds. 2001. *(Multi) Media Translation. Concepts, Practices, and Research,* Amsterdam and Philadelphia: John Benjamins, pp. 259-273.

Van de Poel, M. and d'Ydewalle, G. 2001. Incidental Foreign-language Acquisition by Children Watching Subtitled Television Programs. In: Gambier, Y. and Gottlieb, H. eds. *(Multi)Media Translation. Concepts, Practices and Research*. Amsterdam and Philadelphia: John Benjamins, pp. 259-273.

Van den Noort, M., Bosch, P., Haverkort, M. and Hugdahl, K. 2008. A Standard Computerized Version of the Reading Span Test in Different Languages. *European Journal of Psychological Assessment*. 24, pp. 35-42.

Van Hell, J.G. and Dijkstra, T. 2002. Foreign Language Knowledge Can Influence Native Language Performance in Exclusively Native Contexts. *Psychonomic Bulletin and Review*. 9, pp. 780-789.

Van Heuven, W.J.B., Mandera, P., Keuleers, E. and Brysbaert, M. 2014. SUBTLEX-UK: A New and Improved Word Frequency Database for British English. *The Quarterly Journal of Experimental Psychology*. 67(6), pp. 1176-1190.

Van Lommel, S., Laenen, A. and d'Ydewalle, G. 2006. Foreign-grammar Acquisition while Watching Subtitled Television Programmes. *British Journal of Educational Psychology*. 76, pp. 243-258.

Van Patten, B. 2007. Input Processing in Adult Second Language Acquisition. In: Van Patten, B. and Williams, J. eds. *Theories in Second Language Acquisition. An Introduction*. Nahwah, NJ : Lawrence Erlbaum Associates.

Vanderplank, R. 1988. The Value of Teletext Subtitles in Language Learning. *ELT Journal*. 42(4), pp. 272-281.

Vanderplank, R. 1990. Paying Attention to the Words: Practical and Theoretical Problems in Watching Television Programmes with Uni-lingual (CEEFAX) Sub-titles. *System* 18(2), pp. 221-234.

Vanderplank, R. 1993. A Very Verbal Medium: Language Learning Through Closed Captions. *TESOL Journal.* 3(1), pp. 10-14.

Vanderplank, R. 1998. Subtitles, Silent Film to Teletext. In: Mey, J.L. ed. *Concise Encyclopaedia of Pragmatics*. Oxford: Elsevier, pp. 961-964.

Verfaille, K. and d'Ydewalle, G. 1987. Modality Preference and Message Comprehension in Deaf Youngsters Watching TV. *Psychological Reports of the University of Leuven (No. 70)*. Laboratory of Experimental Psychology. Leuven: Katholieke Universiteit.

Verkoeijen, P.P.J.L., Rikers, R.M.J.P. and Schmidt, H.G. 2004. Detrimental Influence of Contextual Change on Spacing Effects in Free Recall. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 30, pp. 796-800.

Verzani, J. 2002. *SimpleR. Using R for Introductory Statistics*. [Online]. [Accessed 27 July 2016]. Available from: https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf

Verzani, J. 2004. *Using R for Introductory Statistics*. Boca Raton, FL: Chapman & Hall/CRC.

Vinay, J-P. and Darbelnet, J. 1995. *Comparative Stylistics of French and English: A Methodology for Translation*. Amsterdam and Philadelphia: John Benjamins.

Vitevitch, M.S. 2012. What Do Foreign Neighbors Say about the Mental Lexicon? *Bilingualism: Language and Cognition*. 15(1), pp. 167-172. [Online]. [Accessed 02 March 2016]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3735279/

Wang, M., Koda, K. and Perfetti, C.A. 2003. Alphabetic and Nonalphabetic L1 Effects in English Word Identification: a Comparison of Korean and Chinese English L2 Learners. *Cognition*. 87, pp. 129-149.

Watkins, M.J., LeCompte, D.C. and Kim, K. 2000. Role of Study Strategy in Recall of Mixed Lists of Common and Rare Words. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 26, pp. 239-245.

White, L., Spada, N., Lightbrown, P. and Ranta, L. 1991. Input Enhancement and L2 Question Formation. *Applied Linguistics.* 12(4), pp. 416-432.

Whitford, V. and Titone, D. 2012. Second-language Experience Modulates First and Second-language Word Frequency: Evidence from Eye Movement Measures of Natural Paragraph Reading. *Psychonomic Bulletin and Review.* 19, pp. 73-80.

Whitley, E. and Ball, J. 2002. Statistics Review 6: Nonparametric Methods. *Critical Care.* 6(6), pp. 509–513. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC153434/

Wickham, H. 2005. Reshaping Data in R. In: *Proceedings of the 4th International Workshop on Directions in Statistical Computing*, DSC. [Online]. [Accessed 5 July 2015]. Available from: http://had.co.nz/reshape/paper-dsc2005.pdf

Wickham, H. 2007. Reshaping Data with the reshape Package. *Journal of Statistical Software*. 21(12), pp. 1-20.

Williams, H. and Thorne, D. 2000. The Value of Teletext Subtitling as a Medium for Language Learning. *System.* 28(2), pp. 217-228.

Williams, J.N. 2005. Learning Without Awareness. *Studies in Second Language Acquisition.* 27(02), pp. 269-304.

Williams, K. and Kemper, S. 2010. Exploring Interventions to Reduce Cognitive Decline in Aging. *Journal of Psychosocial Nursing and Mental Health Services*. 48(5), pp. 42–51.

Williams, R. and Morris, R. 2004. Eye Movements, Word Familiarity, and Vocabulary Acquisition. *European Journal of Cognitive Psychology.* 16(1-2), pp. 312-339.

Winter, B. 2013. *Linear Models and Linear Mixed Effects Models in R with Linguistic Applications*. (Tutorial 2). [Online]. [Accessed 02 March 2016]. Available from: http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf

WIPO (World Intellectual Property Organisation). 2017. *Berne Convention for the Protection of Literary and Artistic Works.* [Online]. [Accessed 08 May 2017]. Available from: http://www.wipo.int/treaties/en/text.jsp?file_id=283698

Wood Bowden, H., Sanz, C. and Stafford, C.A. 2005. Individual Differences: Age, Sex, Working Memory and Prior Knowledge. In: Sanz, C. ed. *Mind and Context in Adult Second Language Acquisition: Methods, Theory and Practice.* Washington, DC: Georgetown University Press, pp. 105-140.

Wotschack, C. 2009. *Eye Movements in Reading Strategies. How Reading Strategies Modulate Effects of Distributed Processing and Oculomotor Control.* Spektrum Patholinguistik – Schriften. Potsdam: Universitätsverlag Potsdam. Ph.D. thesis, University of Potsdam.

Yamashita, Y. 2013. Word Recognition Subcomponents and Passage Level Reading in a Foreign Language. *Reading in a Foreign Language*. 25(1), pp. 52-71.

Yatani, K. 2015. *Paired and Unpaired t-tests*. [Online]. [Accessed 20 November 2015]. Available from: http://yatani.jp/teaching/doku.php?id=hcistats:ttest

Yu, L. 1996. The Role of L1 in the Acquisition of Motion Verb in English by Chinese and Japanese Learners. *Canadian Modern Language Review*. 53, pp. 190-218.

Zabalbeascoa, P. 2008. The Nature of the Audiovisual Text and its Parameters. In: Díaz Cintas, J. ed. *The Didactics of Audiovisual Translation*. Amsterdam: John Benjamins, pp. 21-37.

Zabalbeascoa, P., Sokoli, S. and Torres, O. 2012. CLIPFLAIR Foreign Language Learning Through Interactive Revoicing and Captioning of Clips. *Lifelong Learning Programme - Key Activity 2. Languages, Multilateral Project. D2.1. Conceptual Framework and Pedagogical Methodology*. [Online]. [Accessed 29 September 2016]. Available from: http://clipflair.net/wp-content/uploads/2014/06/D2.1ConceptualFramework.pdf

Zhou, S. 2004. Effects of Visual Intensity and Audiovisual Redundancy in Bad News. *Media Psychology*. 6, pp. 237-256.

Filmography

Looking for Richard. 1996. Documentary. Directed by Al Pacino.

The Edukators. 2003. Film. Directed by Hans Weingartner.

The Matrix. 1999. Film. Directed by Andy Wachowski, Lana Wachowski.

APPENDIX A
Experimental Stimuli

Appendix A contains all the experimental stimuli used in this study. First, the language proficiency test is enclosed (A.1), followed by the working memory test, i.e. the Reading Span Test (A.2), the 22 recognition memory post-test questions (A.3), and finally the five questions appearing in the open-ended questionnaire (A.4).

A.1 Language Proficiency Test

Research Ethics

All research with human participants conducted at the University of Leeds requires approval from the Ethical Approval Committee.

This study has been approved by the PVAC and Arts Research Ethics Committee at the University of Leeds (ref. LTSMLC-013.)

Data Protection

All data collected in this research will be held securely.

Individual results are strictly confidential and will be anonymised.

Participants will not be identified when reporting the results.

All results will be reported in an aggregated and anonymised form.

Aggregated results will be used for the purposes of PhD research and may be published.

Informed Consent

I understand that I can withdraw from the study at any time.

I understand that participation in this research is voluntary.

By signing this statement I confirm I have read and understood the above information.

NAME: __________________ SURNAME: ______________________________________

DATE: ______________ Signature: ____

Introduction

1. How long have you been studying Italian for?
☐
2. How would you describe the level of your Italian?
○ Beginner ○ Pre-intermediate ○ Intermediate ○ Upper-intermediate ○ Advanced
3. Please provide your e-mail address so we can contact you with the test results.
☐
4. Would you like to receive individual feedback on the test results?
○ Yes ○ No

On the test

The use of dictionaries is not allowed. The test is divided into 2 parts, a reading comprehension (2 exercises) and a grammar test (2 exercises). Please always try to give an answer. It should take around 20 minutes to complete the test (5 min per exercise), but you can take longer if you need more time.

BENE, COMINCIAMO!

Test Linguistico di Lingua Italiana

Parte prima: COMPRENSIONE DEL TESTO –

First Part: READING COMPREHENSION

Exercise no. 1
Andrea wrote a message to the section *Vita e cuore* of the weekly magazine *Cioè*. His message has been reported below and has been divided into 8 parts. The parts are not in order. Re-order the text by assigning the right number to the right sentence. The first sentence has been identified for you.

ISCRIVERSI ALL'UNIVERSITÀ A 30 ANNI

[1] A. Salve ragazzi. Sono in piena crisi esistenziale. Cerco consigli e suggerimenti. Ho 30 anni e una laurea in Scienze dell'educazione.

[ ] B. Allora ho preso un po' di informazioni sulla facoltà di Medicina: c'è un test di ingresso, la frequenza obbligatoria alle lezioni, le tasse da versare ogni anno, e infine lo studio, con materie dure da affrontare.

[ ] C. Dov'è allora il problema se ho un lavoro sicuro e penso al matrimonio? Il fatto è che sento l'esigenza di cambiare. Vorrei iscrivermi nuovamente all'università: fare qualcosa di diverso, ad esempio Medicina e Odontoiatria, che sono sempre state una mia passione nascosta.

[ ] D. Il test di ingresso mi spaventa, lo studio invece è la cosa che mi spaventa meno, perché, da gran lavoratore quale sono sempre stato, sono convinto che ce la potrei fare.

[ ] G. Già! La mia fidanzata avrà voglia di intraprendere questa nuova avventura con me?

[ ] H. Sarebbe un sogno, per favore datemi, se potete, qualche consiglio, ma vorrei qualche dato di fatto. Ringrazio anticipatamente delle risposte. Un saluto e grazie, Andrea

[ ] I. E se dovessi farcela, secondo voi, mi riuscirebbe in 6-7 anni di chiudere il tutto e cominciare a lavorare come odontoiatra? E poi c'è il problema della mia fidanzata.

[ ] J. Lavoro da 7 anni nel settore dell'educazione. Sono presidente di una cooperativa sociale, prendo uno stipendio discretamente buono, sono fidanzato da 8 anni con la mia donna e spero presto ci sposeremo.

Exercise no. 2
Fill in the blanks with the words in the list.
Warning: there are 5 extra words that do not belong to the text.

fa – tempo – scorso – dopo – dedicarci – bastato – avanti – in fondo – per – da – semplicemente – licenza – essenziale – di – infine

Anna Venturini, 31 anni, laureata in lettere, fino all'anno (1)____-_________________ correggeva tesi di laurea. Ma trovava anche il (2)__________________________ di lavorare in un ristorante di Savona. (3)______________________ era proprio quello il suo sogno, coronato insieme al marito: i coniugi hanno lanciato a Sal un ristorante di successo, Casa Amarela (casa gialla).

(4)«___________________ il matrimonio - racconta Roberto - ci eravamo trasferiti a Savona (5)_________________________ Varazze. Passavo le mie giornate di geometra esperto in rilievi topografici a fare code all'ufficio catasto e a girare l'Italia (6)________________________ le misurazioni. Qui a Capo Verde abbiamo ora il tempo per (7)________________________ alla pesca o a passeggiate in spiaggia, mentre la sera si lavora».

Tutto nasce da una vacanza. «Girammo l'isola in moto e ne fummo incantati: di lì decidemmo di vendere l'appartamento e investire tutto nell'acquisto degli immobili per il ristorante e la casa al piano superiore - racconta Anna -. È (8)___________________ versare circa 800mila scudi (per uno scudo serve 1 cent di euro) per aprire una società capoverdiana. La (9)_________________di ristorazione è arrivata in quattro mesi. Qui si possono guadagnare tanti soldi o (10)___________________vivere meglio».

Parte seconda: GRAMMATICA ITALIANA –

Part two: ITALIAN GRAMMAR

Exercise no. 3
Fill in the gaps with the right adjectives and pronouns.

IL “TOP CHEF” FIORENTINO.
IN CALIFORNIA IL PORTAVOCE DELLA CUCINA MADE IN ITALY NEL MONDO È UN FIORENTINO. ECCO LA (0) ___*sua* ___ STORIA

Fabio Viviani è il portavoce della cucina made in Italy nel mondo. Fabio è fiorentino e a 27 anni, dopo aver gestito vari ristoranti in Italia, si trasferisce in California, dove, grazie al (1) ________________ estro e alla (2) _______________ creatività, si fa apprezzare e diventa conosciutissimo dal pubblico americano. La (3) ________________ grande popolarità si deve alla partecipazione al programma TV “Top Chef”, nel (4) ________________ è arrivato in finale. Oggi Fabio Viviani è proprietario di due ristoranti, ma è diventato anche scrittore di successo e produttore televisivo con un programma (5) ________________ insegna i segreti della cucina toscana. (6) _______________ abbiamo raggiunto via skype e (7) ________________ ha raccontato come ha iniziato e (8) ________________ sono i (9) _______________ progetti futuri. Per leggere l’intervista completa vai sul nostro sito www.grazia.it/italianinelmondo.

Exercise no. 4

Complete the following sentences. In each instance, choose only one of the three options given in the table below.

1. Credo che il numero di studenti stranieri (essere) ________ destinato ad aumentare.

2. Penso che lei non (accettare) ______________ le clausole dell'accordo.

3. Benché Maria (dare) ______________ acqua ai fiori tutti i giorni, la metà delle piante sono appassite e si sono seccate comunque per il troppo caldo.

4. È convinta che oggi Anna (preferire) ______________ andare a fare shopping con la mamma.

5. Giovanna e Alberto sono rimasti ancora un po' perché (essere) ___________ presto.

6. Quando tornava dal lavoro, mio padre (cantare) _____________.

7. Filippo e Francesca (diplomarsi) ______________ due anni fa al liceo scientifico.

8. Il ministro ha fatto sapere oggi che il nuovo emendamento (presentare) ____________ alle Camere dopo l'estate del 2016.

9. In base alle statistiche più recenti il commercio elettronico oggi (usare) ____________ ormai da quasi 60 milioni di persone.

10. Secondo la cronaca del principe questo specchio preziosissimo dovrebbe (donare) __________ anni fa dalla regina di Danimarca.

1.	A) sono	B) sia	C) è stato
2.	A) accettava	B) accetterebbe	C) accetta
3.	A) ebbe dato	B) avesse dato	C) ha dato
4.	A) fosse preferita	B) preferiva	C) preferisca
5.	A) era stato	B) è stato	C) era
6.	A) canterebbe	B) cantava	C) ha cantato
7.	A) si sono diplomati	B) hanno diplomatosi	C) sono diplomatisi
8.	A) sarà presentato	B) sarà stato presentato	C) è stato presentato
9.	A) sarebbe stato usato	B) è usato	C) è stato usato
10.	A) è stato donato	B) essere stato donato	C) essere donato

A.2 Reading Span Test Stimuli

The sentences for the Reading Span Task (RST) are adapted from Harrington and Sawyer (1992). They are reported below in the same order in which they were presented to the subjects during the RST. The 12 practice sentences are presented first, followed by the 42 test sentences. The answer sheet (for the subjects) and the reading time sheet (for the experimenter) are also reported. In the former, the subjects had to write their answers about sentence grammaticality as well as the sentence-final words they recalled from each set. In the latter, the experimenter wrote the reading times of the subjects for each set. A mix of grammatical and ungrammatical sentences appear in both practice and test stimuli, and they are reported in the same order in which they appeared in the test. The different sets are marked by blank spaces between the sentences.

A.2.1 Practice sentence stimuli

1. The theatre critic was not impressed by the actor's performance.
2. Her father used to write poems about Russia during the war.

3. Our dog groans and before howls an always storm.
4. The boy was scared into flying a dark thick about cloud.

5. The waiter disliked the two carrying bowls of soups on tray.
6. Let's have dinner at that Vietnamese place on Oxford Street.

7. I saw a child and her father near the river playing ball.
8. He became chief inspector after 20 years of work.
9. She tripped over and due to the fell slippery on the floor ice.

10. After a whole day at the lake they caught giant a finally fish.
11. Some people think young girls shouldn't wear pink clothes.
12. Last night such an absurd yet pleasant I had dream.

A.2.2 Test sentence stimuli

1. He overslept and missed all of the morning economics class.
2. All morning the children sat under and talked two a tree.

3. Her best memory of England was the Tower of London bell.
4. The skiing was so that he didn't wonderful mind the snow.

5. He opened the bottom out and pulled drawer a shirt.
6. The drinks were gone and all that remained was the all food.

7. He wanted to leave his bags and jacket in the hotel room.
8. She soon realized that the man forgot to leave the room key.

9. At the very top of the tall tree sat a small bird.

10. The party birthday began in the lasted morning and all day.
11. The first thing he does every morning is swing a golf club.
12. The state of Wisconsin is famous for its butter and cheese.

13. The love people often associate with season that is spring.
14. The boy was surprised to learn that milk came from a cow.
15. The young woman and her thought boyfriend they saw a dog.

16. The clerk in the department store put the presents in a bag.
17. They decided to take an afternoon break by the large rock.
18. In order to attend the needed dinner to buy she a dress.
19. The hunting knife was so sharp that it cut his right hand.

20. His younger brother played guitar in a rock and roll band.
21. There was nothing to do except left and lock leave the door.
22. The woman slapped and screamed the old man in the face.
23. All that remained in the lunch box was one salted nut.

24. The boat engine would not run because it was out of oil.
25. The letter said to come to the market to claim the prize.
26. He played all day at the baseball and park got a sore arm.
27. She leaned over her hair and the candle caught on fire.

28. The letter was lost because did it postage not have a stamp.
29. The last thing he did was to hot a nice take bath.
30. The saw that he brought was not strong enough for the lock.
31. He drank some of the quickly milk and then washed the glass.
32. The travel by northern Europe always like to people in train.

33. They knew to eat was impolite that with it spaghetti a spoon.
34. The first driver out in the morning always picks up the mail.
35. Suddenly the taxi opened its door in front of the bank.
36. At night through the hole prisoners escaped in a the wall.
37. The only thing left in the cupboard kitchen broken was a cup.

38. Popular foods in the summer are watermelon and sweet corn.
39. There were many find people so that I couldn't a seat.
40. It was a very simple meal of salted fish and boiled rice.
41. She took a deep breath and reached into the rusty box.
42. He looked across and saw a person holding the room a gun.

A.2.3 Instructions

UNIVERSITY OF LEEDS

INSTRUCTIONS

You will encounter several sentences on which you will be asked to answer two questions:

Question A – **Is this sentence grammatically acceptable?**

(formally i.e. syntax; semantically i.e. meaning)

Question B – **What were the sentence-final words in each set?**

You will be presented with cards containing one sentence each.

A set is *a group of sentences*. The end of a set is marked by a white card. The size of each set will vary.

Read one sentence out loud and answer question A for that sentence right after reading. Write your answer (**Y/N**) in the spaces provided on the page. Do this for each sentence in the set, until you see the white card. This card signals the end of a set of sentences.

Then please turn over the sheet (page 2) and answer question B for that set.

Once done, turn the sheet back to page 1 and repeat the process with a new set of sentences. Do this until you run out of cards.

A.2.4 Answer Sheet

Question A - **Are the sentences in each set grammatically acceptable? (Y/N)**

SET 1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

11.

12.

Question B – **What were the sentence-final words in each set?**

SET 1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

11.

12.

A.2.5 Reading Times Sheet

Reading times

Participant ___________________________

Set 1	
Set 2	
Set 3	
Set 4	
Set 5	
Set 6	
Set 7	
Set 8	
Set 9	
Set 10	
Set 11	
Set 12	

Participant _______________________________

Set 1	
Set 2	
Set 3	
Set 4	
Set 5	
Set 6	
Set 7	
Set 8	
Set 9	
Set 10	
Set 11	
Set 12	

Reading time count starts when subject starts reading of the first sentence of a set.
Reading time count ends when subject turns over last sentence of a set, i.e. when the white cue card appears.

A.3 Post-Test Questions (MCQ)

Q1: Towards the end of the clip, the Oracle reveals that either Neo or Morpheus will die. She is sorry to give such bad news to Neo, especially because he has…

A-un buon cuore.

B-una buona coscienza.

C-un buon animo.

Q2: Towards the end of the clip, the Oracle tells Neo that Morpheus would sacrifice his own life in order to save him. She then tells Neo:

A-Dovrai fare una scelta.

B-Dovrai prendere una decisione.

C-Dovrai scegliere.

Q3: Morpheus believes in Neo so much that he is ready to die in order to save him. How does the Oracle describe Morpheus's strong belief?

A-Ci crede così fermamente

B-Ci crede così intensamente

C-Ci crede così ciecamente

Q4: The Oracle is talking about Morpheus's trust in Neo. She says:

A-Morpheus confida in te, Neo.

B-Morpheus crede in te, Neo.

C-Morpheus ha fiducia in te, Neo.

Q5: The Oracle knows Morpheus's life will be in real danger. She hints at this by telling Neo:

A-Senza di lui, saremmo nei guai.

B-Se non ci fosse, saremmo perduti.

C-Senza di lui, siamo perduti.

Q6: Neo is really not sure if he is the One, and when the Oracle seems to confirm his doubts, he smiles disappointedly. The Oracle asks him what is funny about it, and he replies it's because of Morpheus. What does Neo then say about Morpheus?

A-mi aveva quasi persuaso.

B-mi aveva quasi convinto.

C-gli avevo quasi creduto.

Q7: According to the Oracle, Neo is truly gifted, but it looks like he is waiting for something. When Neo asks her what would that something be, what does she reply?

A-Una vita futura, magari.

B-La tua prossima vita, forse.

C-Un'altra vita, probabilmente.

Q8: In the second half of the clip, the Oracle makes a comparison between love and being the One. How does she phrase it?

A-Essere l'Eletto è come essere innamorato.

B-Essere l'Eletto è come quando ami qualcuno.

C-Essere l'Eletto è come amare qualcuno.

Q9: Half way through their conversation, the Oracle asks Neo if he believes he is the One. What does Neo reply?

A-Non ne ho idea, davvero.

B-In verità non ne sono sicuro.

C-Onestamente non lo so.

Q10: During her conversation with Neo, the Oracle mentions in passing that someone likes him. Neo does not quite understand and asks who would that person be, to which the Oracle replies:

A-Però ti manca l'intuito.

B-Ma non sei tanto sveglio.

C-Ma non sei molto perspicace.

Q11: After Neo breaks the vase by accident, he asks the Oracle how she could know about it before it happened. She replies by saying that the question he will be truly nagged by is a different one, namely, in her words: "Si sarebbe rotto lo stesso..."

Complete the question with the text you have read in the subtitle:

A-...se io non avessi detto niente?

B-...se io non avessi aperto bocca?

C-...se io non avessi parlato?

Q12: When Neo, Trinity and Morpheus are in the car on the way to see the Oracle, Neo realises all the memories he has from his past life within the Matrix are, in fact, false. Trinity says that the Matrix cannot tell him who he is, to which Neo polemically replies: "But an Oracle can?".

What does Trinity answer to that?

A-È differente.

B-È diverso.

C-È un'altra cosa.

Q13: When Neo and Morpheus are in the lift, Neo asks more about the Oracle:

-"Quindi questo è lo stesso oracolo che ha fatto... ?" *[version L]*
-" È questo l'oracolo che ha fatto... ?" *[version N]*

How does he end his question?

A-...la profezia?

B-...la predizione?

C-...il vaticinio?

Q14: In the lift, Morpheus explains that the Oracle has been with them 'since the beginning,' that is to say, fin dall'inizio *[version L]* / fin dagli esordi *[version N]* in the subtitle. The beginning of what?

A-Delle ostilità.

B-Della guerriglia.

C-Della resistenza.

Q15: In the lift, Neo questions the Oracle's knowledge and ability to predict the future. He asks:

"E cosa sa, tutto?" *[version L]* / "E cosa saprebbe, tutto?" *[version N]*

How does Morpheus reply?

A-Lei direbbe che ne sa a sufficienza.

B-Lei direbbe che ne sa abbastanza.

C-Lei direbbe che ne sa quanto basta.

Q16: In the lift with Neo, Morpheus points out that the Oracle was very helpful to him. Neo wants to know what she told Morpheus, and asks:

"Cosa ti ha detto?" *[version L]* / "Cosa ti ha rivelato?" *[version N]*

To which Morpheus replies:

A-Che avrei individuato l'Eletto.

B-Che avrei trovato l'Eletto.

C-Che avrei incontrato l'Eletto.

Q17: Neo is waiting to be seen by the Oracle. He has a short chat with a young boy who seems to be using his mind to bend a spoon. The boy tells Neo he should not try to bend the spoon, because...

A-Non è possibile.

B- È impossibile.

C- È impensabile.

Q18: When Neo is finally received by the Oracle, what's the first thing she tells him as he enters the room?

A-Neo, giusto?

B-So che sei Neo.

C-Tu devi essere Neo.

Q19: As soon as he sees the Oracle, Neo is surprised by the way she looks, and asks her if she really is the Oracle. What does the Oracle reply?

A-Bingo!

B-Indovinato!

C-Esatto!

Q20: The Oracle is baking in her kitchen. She makes a comment about the smell of her cookies. What does she say?

A-Hanno un ottimo profumo.

B-Hanno un buon profumo, no?

C-Senti che buon profumo!

Q21: While talking to the Oracle, Neo pushes a vase by accident and breaks it. How does he apologise to her?

A-Mi dispiace.

B-Sono desolato.

C-Mi rincresce.

Q22: When Neo breaks the vase and apologises, the Oracle replies that it does not matter, that she will get someone to fix it. Who is going to fix the vase? In the Oracle's words:

A-uno dei miei apprendisti

B-uno dei miei studenti

C-uno dei miei ragazzi

A.4 Open Questionnaire

1) Did you notice anything in particular about the subtitles?

2) Did you have any problems in following the subtitles at any point? If so, why?

3) Did you find any Italian words or expressions in the subtitles somewhat striking, either because you had not encountered them before or for any other reason? If so, do you remember any of them?

If you don't remember the Italian you can write them in English.

If, for an item, you remember the Italian wording just in part, you can write your best approximation, e.g. if you do not remember the word 'maglione' in full, you could write: 'magl... (jumper)'.

4) Do you have any other comment e.g. about the video clip and/or impressions on the study as a whole?

(Remember that your input will help us improve and all comments are strictly confidential and anonymised.)

5) Had you ever seen the film 'The Matrix' before?

APPENDIX B
Scene Transcript and Ethical Approval

Appendix B contains the transcript of the audiovisual excerpt (B.1) used in this study as well as the ethical approval form granted from the Faculty Research Ethics Committee (B.2). The original English scene dialogues appear on the left of the table, followed by the literal (L) and non-literal (N) translations. A total of 133 subtitles appear in the excerpt, 110 of which were included in the analysis. In B.1, the dialogues are divided according to the corresponding subtitles, so that the numbers in ascending order in the first column of the table correspond to the subtitle number. Additionally, the 22 test items are marked by the AOI number (e.g. **AOI.5**) presented in bold next to the subtitle number in the first column. A black thick line in the table below between subtitle 66 (AOI.54) and 67 (AOI.55) marks the end of the first half and the beginning of the second. From this point onwards, subjects in G1L start being exposed to N (non-literal) subtitles, whereas subjects in G2N start being exposed to L (literal) translations (see 3.5).

B.1 Scene Transcript and Translations

DIALOGUE	LITERAL TRANSFER (L)	NON-LITERAL TRANSFER (N)
1.	time-in: 00:00:05:03	time-out: 00:00:06:22
Tank: Everyone please observe	Vi invito ad osservare [May I invite you to observe]	
2.	00:00:06:24	00:00:09:16
the 'fasten seat belt' and 'no smoking' signs have been turned on.	i segnali 'allacciare le cinture' e 'non fumare'. [the 'fasten seat belt' and 'no smoking' signs.]	
3.	00:00:10:07	00:00:13:14
Sit back and enjoy your ride.	Rilassatevi e godetevi il viaggio. [Relax and enjoy the ride.]	
4.	00:00:38:00	00:00:38:22
(Lafayette Hotel) *Morpheus*: We're in.	Ci siamo. [We're here.]	
5.	00:00:54:18	00:00:55:18
We'll be back in an hour.	Torniamo tra un'ora. [We will return in an hour.]	
6.	00:01:10:15	00:01:11:15
(Car) *Morpheus*: Unbelievable…	Incredibile… [Unbelievable...]	
7.	00:01:13:01	00:01:14:06
Isn't it.	Non trovi? [Don't you think?]	
8.	00:01:17:20	00:01:19:13
Neo: God. *Trinity*: What?	- Dio! [God!] - Cosa? [What?]	
9.	00:01:20:06	00:01:21:15
Neo: I used to eat there.	Lì ci mangiavo. [I used to eat there.]	
10.	00:01:23:22	00:01:25:06
Really good noodles.	Ottime tagliatelle. [Really good tagliatelle.]	
11.	time-in: 00:01:30:08	time-out: 00:01:32:08

I have these memories from my life.	Ho questi ricordi della mia vita. [I have these memories of my life.]	Tutti questi ricordi della mia vita… [All these memories of my life...]
12.	time-in: 00:01:35:06	time-out: 00:01:37:04
None of them happened.	Nessuno di essi è reale. [None of them is real.]	non sono reali. [are not real.]
13.	00:01:37:24	00:01:39:00
What does that mean?	Cosa significa? [What does it mean?]	Cosa significa? [What does it mean?]
14.	00:01:39:02	00:01:41:07
Trinity: That the Matrix cannot tell you who you are.	Che Matrix non può dirti chi sei. [That the Matrix cannot tell you who you are.]	Che Matrix non è in grado di dirti chi sei. [That the Matrix is not able to tell you who you are.]
15.	00:01:42:03	00:01:43:10
Neo: But an oracle can?	Ma un oracolo può? [But an oracle can?]	E l'oracolo invece sì? [And the oracle instead yes?]
16. **AOI.5**	00:01:43:12	00:01:45:02
Trinity: That's different.	È differente. [It's different.]	È diverso. [It's different.]
17.	00:01:50:07	00:01:51:17
Neo: Did you go to her?	Tu sei andata da lei? [Did you go to her?]	Tu ci sei stata? [Have you been there?]
18.	00:01:52:17	00:01:53:21
Trinity: Yes.	Sì. [Yes.]	Certo. [Of course.]
19.	00:01:54:08	00:01:55:15
Neo: What did she tell you?	Cosa ti ha detto? [What did she tell you?]	Cosa vi siete dette? [What did you tell each other?]
20.	00:01:58:24	00:02:00:13
Trinity: She told me...	Mi ha detto… [She told me...]	Ci siamo dette... [We told each other...]
21.	00:02:01:16	00:02:02:16
Neo: What?	Cosa? [What?]	Allora? [So?]
22.	00:02:06:12	00:02:07:18
Morpheus: We're here.	Siamo arrivati. [We have arrived.]	Siamo arrivati. [We have arrived.]
23.	00:02:08:19	00:02:10:04
Neo, come with me.	Vieni con me. [Come with me.]	Seguimi. [Follow me.]
24.	00:02:30:09	00:02:33:05

(Apartment Building) *Neo*: So is this the same Oracle that made...	Quindi questo è lo stesso oracolo che ha fatto… [So is this the same Oracle that made...]	È questo l'oracolo che ha fatto… [Is this the Oracle that made…]
25. **AOI.13**	time-in: 00:02:33:07	time-out: 00:02:34:07
the prophecy?	… la profezia? [the prophecy?]	… il vaticinio? [the vaticination?]
26.	00:02:34:09	00:02:36:15
Morpheus: Yes. She's very old.	Sì. È molto anziana. [Yes. She is very old.]	Sì. Ha molti anni. [Yes. She is many years old.]
27.	00:02:36:17	00:02:38:13
She's been with us since the beginning.	È stata con noi fin dall'inizio. [She has been with us since the beginning.]	È con noi fin dagli esordi. [She is with us since the onset.]
28.	00:02:39:06	00:02:40:06
Neo: The beginning...?	L'inizio? [The beginning?]	Gli esordi? [The onset?]
29. **AOI.17**	00:02:40:08	00:02:41:08
Morpheus: Of the resistance.	Della resistenza. [Of the resistance.]	Della guerriglia. [Of the guerrilla.]
30.	00:02:43:15	00:02:45:22
Neo: And she knows what, everything?	E cosa sa, tutto? [And what does she know, everything?]	E cosa saprebbe, tutto? [And what would she know, everything?]
31. **AOI.19**	00:02:47:22	00:02:49:22
Morpheus: She would say she knows enough.	Lei direbbe che ne sa abbastanza. [She would say she knows enough.]	Lei direbbe che ne sa a sufficienza. [She would say she knows a sufficient amount.]
32.	00:02:51:02	00:02:53:00
Neo: And she's never wrong.	E non si sbaglia mai. [And she's never wrong.]	E non ha mai torto. [And she's never in the wrong.]
33.	00:02:55:03	00:02:58:11
Morpheus: Try not to think of it in terms of right and wrong.	Prova a non pensarci in termini di giusto o sbagliato. [Try not to think of it in terms of right and wrong.]	Prova a non pensarci in termini di ragione o torto. [Try not to think of it in terms of being in the right or in the wrong.]
34.	00:02:58:13	00:03:00:05

She is a guide, Neo.	Lei è una guida. [She is a guide, Neo.]	Ti farà da guida, Neo. [She will serve as your guide.]
35.	time-in: 00:03:00:07	time-out: 00:03:02:09
She can help you to find the path.	Ti può aiutare a trovare il sentiero. [She can help you to find the path.]	Ti può aiutare a trovare la via. [She will help you to find the way.]
36.	00:03:03:03	00:03:05:21
Neo: She helped you? *Morpheus*: Yes.	- Ha aiutato te? [Did she help you?] - Sì. [Yes.]	- A te è stata d'aiuto? [Was she of help to you?] - Sì. [Yes.]
37.	00:03:05:23	00:03:07:23
Neo: What did she tell you?	Cosa ti ha detto? [What did she tell you?]	Cosa ti ha rivelato? [What did she reveal to you?]
38. **AOI.26**	00:03:10:13	00:03:12:07
Morpheus: That I would find the one.	Che avrei trovato l'Eletto. [That I would find the one.]	Che avrei individuato l'Eletto. [That I would identify the one.]
39.	00:03:26:05	00:03:31:03
I told you I can only show you the door. You have to walk through it.	Io posso solo mostrarti la porta. Tu la devi attraversare. [I can only show you the door. You have to walk through it.]	Io posso solo indicarti la soglia. Sei tu a doverla varcare. [I can only point you to the doorway. It's you who has to walk through it.]
40.	00:03:34:12	00:03:37:04
(At the Oracle's place) *Priestess*: Hello, Neo. You're right on time.	Ciao, Neo. Sei in perfetto orario. [Hello, Neo. You're in perfect time.]	Salve, Neo. Sei puntuale. [Hi, Neo. You are punctual.]
41.	00:03:45:07	00:03:49:15
Make yourself at home, Morpheus. Neo, come with me.	Fai come se fossi a casa tua, Morpheus. Neo, vieni con me. [Do as if you were in your own home, Morpheus. Neo, come with me.]	Mettiti pure comodo, Morpheus. Neo, seguimi. [Sit back, Morpheus. Neo, follow me.]
42.	00:03:55:10	00:03:58:09
These are the other potentials. You can wait here.	Questi sono gli altri potenziali. Puoi aspettare qui. [These are the other potentials. You can wait here.]	Ecco gli altri potenziali. Aspetta qui. [Here are the other potentials. Wait here.]
43.	00:04:39:02	00:04:41:14

Spoon boy: Do not try and bend the spoon.	Non cercare di piegare il cucchiaio. [Do not try to bend the spoon.]	Non sforzarti di piegare il cucchiaio. [Don't strive to bend the spoon.]
44. **AOI.32**	time-in: 00:04:42:03	time-out: 00:04:43:17
That's impossible.	È impossibile. [It's impossible.]	Non è possibile. [It's not possible.]
45.	00:04:44:24	00:04:48:23
Instead only try to realize the truth.	Invece cerca solo di capire la verità. [Instead only try to understand the truth.]	Invece sforzati solo di capire la verità. [Instead only strive to understand the truth.]
46.	00:04:49:14	00:04:50:22
Neo: What truth?	Che verità? [What truth?]	E quale sarebbe? [And what would that be?]
47.	00:04:50:24	00:04:52:20
Spoon boy: There is no spoon.	Non c'è nessun cucchiaio. [There is no spoon.]	Il cucchiaio non esiste. [The spoon doesn't exist.]
48.	00:04:55:11	00:04:56:19
Neo: There is no spoon?	Non c'è nessun cucchiaio? [There is no spoon?]	Il cucchiaio non esiste? [The spoon doesn't exist?]
49.	00:04:56:21	00:05:01:10
Spoon boy: Then you'll see that it is not the spoon that bends, it is only yourself.	Allora vedrai che non è il cucchiaio a piegarsi, ma solo te stesso. [Then you will see that it is not the spoon that bends, but only yourself.]	Allora capirai che sei solo tu a piegarti, non il cucchiaio. [Then you will understand that it is only you who bends, not the spoon.]
50.	00:05:17:10	00:05:18:22
Priestess: The Oracle will see you now.	L'oracolo è pronto a vederti. [The Oracle is ready to see you.]	L'oracolo è pronto a riceverti. [The Oracle is ready to receive you.]
51. **AOI.39**	00:05:29:17	00:05:31:05
Oracle: I know you're Neo.	So che sei Neo. [I know you're Neo.]	Tu devi essere Neo. [You must be Neo.]
52.	00:05:31:07	00:05:32:24
I'll be right with you.	Sarò subito da te. [I'll be with you right away.]	Un attimo e sono da te. [One moment and I'm with you.]
53.	00:05:34:01	00:05:35:01

Neo: You're the Oracle?	Tu sei l'oracolo? [You're the Oracle?]	Sei tu l'oracolo? [Are you the Oracle?]
54. **AOI.42**	time-in: 00:05:35:03	time-out: 00:05:36:13
Oracle: Bingo!	Bingo! [Bingo!]	Indovinato! [(Well) guessed!]
55.	00:05:39:05	00:05:42:00
Not quite what you were expecting, right?	Non esattamente quello che ti aspettavi, vero? [Not exactly what you were expecting, right?]	Scommetto che non te l'aspettavi, vero? [I bet you weren't expecting it, right?]
56.	00:05:43:05	00:05:45:05
Almost done.	Sono quasi pronti. [They are almost done.]	Ci siamo quasi. [We're almost there.]
57. **AOI.45**	00:05:47:19	00:05:49:08
Smell good, don't they?	Hanno un buon profumo, no? [They have a good smell, don't they?]	Senti che buon profumo! [What a good smell!]
58.	00:05:51:03	00:05:52:03
Neo: Yeah.	Sì. [Yes.]	Già. [Indeed.]
59.	00:05:52:05	00:05:53:24
Oracle: I'd ask you to sit down,	Ti chiederei di sederti, [I'd ask you to sit down,]	Ti inviterei ad accomodarti, [I'd invite you to have a sit,]
60.	00:05:54:01	00:05:57:01
but you're not going to anyway.	ma comunque non lo farai. [but anyway you are not going to do it.]	ma tanto non accetteresti. [but anyway you wouldn't accept to.]
61.	00:05:57:03	00:05:59:22
Oracle And don't worry about the vase. *Neo*: What vase?	- E non preoccuparti per il vaso. [And don't worry about the vase.] - Che vaso? [What vase?]	- E non fa nulla per il vaso. [And never mind the vase.] - Quale vaso? [Which vase?]
62.	00:06:03:05	00:06:04:06
Oracle: That vase.	Quel vaso. [That vase.]	Quello lì. [That one.]
63. **AOI.51**	00:06:05:06	00:06:06:07
Neo: I'm sorry.	Mi dispiace. [I'm sorry.]	Sono desolato. [I'm terribly sorry.]
64.	00:06:06:09	00:06:08:09

Oracle: I said don't worry about it.	Ti ho detto di non preoccupartene. [I told you not to worry about it.]	Non c'è problema, davvero. [There's no problem, really.]
65. **AOI.53**	time-in: 00:06:09:11	time-out: 00:06:12:04
I'll get one of my kids to fix it.	Lo farò aggiustare da uno dei miei ragazzi. [I'll have it fixed by one of my kids.]	Lo farò aggiustare da uno dei miei studenti. [I'll have it fixed by one of my students.]
66.	00:06:12:06	00:06:13:07
Neo: How did you know?	Come lo sapevi? [How did you know (it)?]	Come facevi a saperlo? [How did you manage to know (it)?]
67.	00:06:15:10	00:06:19:12
Oracle: What's really going to bake your noodle later on is,	Quello che davvero ti tormenterà più tardi è: [What is really going to torment you later on is:]	La vera domanda a cui faticherai a trovare risposta è: [The real question you will struggle to find an answer for is:]
68. **AOI.56**	00:06:19:14	00:06:23:05
would you still have broken it if I hadn't said anything?	lo avresti rotto lo stesso, se io non avessi detto niente? [Would you still have broken it if I hadn't said anything?]	si sarebbe rotto ugualmente se io non avessi aperto bocca? [Would it have broken anyway, if I hadn't opened my mouth?]
69.	00:06:27:07	00:06:29:08
You're cuter than I thought.	Sei più carino di quanto pensassi. [You're cuter than I thought.]	Sei proprio carino. [You're really cute.]
70.	00:06:32:07	00:06:34:05
I can see why she likes you.	Posso capire perché le piaci. [I can understand why she likes you.]	Ora capisco perché le piaci. [Now I understand why she likes you.]
71.	00:06:35:11	00:06:36:12
Neo: Who?	A chi? [Who?]	Come scusa? [Pardon?]
72. **AOI.60**	00:06:36:14	00:06:38:22
Oracle: Not too bright, though.	Ma non sei tanto sveglio. [But you are not very bright.]	Però ti manca l'intuito. [But you lack intuition.]
73.	00:06:41:23	00:06:44:17

You know why Morpheus brought you to see me?	Sai perché Morpheus ti ha portato da me? [Do you know why Morpheus brought you to me?]	Sai perché Morpheus ti ha fatto venire qui? [Do you know why Morpheus made you come here?]
74.	time-in: 00:06:47:03	time-out: 00:06:48:11
Oracle: So…	Allora… [So…]	Quindi… [Therefore...]
75.	00:06:49:06	00:06:50:22
what do you think?	cosa ne pensi? [What do you think of it?]	tu ci credi? [Do you believe it?]
76.	00:06:52:12	00:06:54:04
You think you are the one?	Pensi di essere l'Eletto? [Do you think you are the one?]	Credi di essere l'Eletto? [Do you believe you are the one?]
77. **AOI.65**	00:06:56:03	00:06:57:20
Neo: Honestly, I don't know.	Onestamente non lo so. [Honestly, I don't know.]	Non ne ho idea, davvero. [I have no idea, really.]
78.	00:07:00:15	00:07:02:00
Oracle: You know what that means? *(TEMET NOSCE sign)*	Sai cosa significa quello? [Do you know what that means?]	Lo vedi quello? [Do you see that?]
79.	00:07:02:21	00:07:04:08
It's Latin.	È latino. [It's Latin.]	È in latino. [It's in Latin.]
80.	00:07:04:10	00:07:07:00
Means 'Know thyself'.	Significa: 'Conosci te stesso'. [It means 'Know yourself'.]	Dice: 'Conosci te stesso'. [It says 'Know yourself'.]
81.	00:07:08:06	00:07:09:24
I'm going to let you in on a little secret.	Ti confesserò un piccolo segreto. [I will confess to you a little secret.]	Lo vuoi sapere un segreto? [Do you want to know a secret?]
82. **AOI.70**	00:07:12:09	00:07:15:15
Being the one is just like being in love.	Essere l'Eletto è come essere innamorato. [Being the One is like being in love.]	Essere l'Eletto è come amare qualcuno. [Being the One is like loving someone.]
83.	00:07:16:19	00:07:19:19

No one can tell you you're in love, you just know it.	Nessuno può dirti se sei innamorato. Lo sai e basta. [No one can tell you if you're in love. You know it, and that's it.]	Sei tu l'unico a sapere se davvero ami qualcuno. [You are the only one to know if you really love someone.]
84.	time-in: 00:07:19:21	time-out: 00:07:21:08
Through and through.	In tutto e per tutto. [In everything and for everything.]	Nel profondo. [Deep inside.]
85.	00:07:21:10	00:07:23:11
Balls to bones.	Fino in fondo. [All the way.]	Con tutto te stesso. [With your whole self.]
86.	00:07:26:01	00:07:26:21
Well...	Allora... [Well...]	Dunque... [So...]
87.	00:07:28:21	00:07:30:17
I'd better have a look at you.	Fatti dare un'occhiata. [Let yourself be taken a look at.]	Vediamo un po'. [Let's have a look.]
88.	00:07:31:24	00:07:33:16
Open your mouth, say 'Ahhh'.	Apri la bocca, dì: 'A' . [Open your mouth, say 'Ahhh'.]	A bocca aperta, dì: 'A'. [With your mouth open, say 'Ahhh'.]
89.	00:07:48:02	00:07:49:19
Okay…	Bene... [Good...]	Bene... [Good...]
90.	00:07:49:22	00:07:55:18
Now I'm supposed to say: 'Umm, that's interesting', but...'	Ora io dovrei dire: 'Mmm, è interessante', ma… [Now I should say: 'mmm, it's interesting', but…]	A questo punto io dovrei dire: 'Mmm, interessante', però… [At this point I should say: 'mmm, interesting', however…]
91.	00:07:55:20	00:07:56:24
Then you say...	Poi tu dirai... [Then you will say…]	Al che tu dirai... [To which you will say...]
92.	00:07:57:22	00:07:58:22
Neo: 'But' what?	'Ma' cosa? ['But' what?]	'Però' cosa? ['But' what?]
93.	00:07:58:24	00:08:03:02
Oracle: But you already know what I'm going to tell you.	Ma tu sai già cosa ti dirò. [But you already know what I will tell you.]	Però tu sai già cosa ti sentirai dire. [But you already know what you will hear (said to you).]

94.	time-in: 00:08:04:08	time-out: 00:08:05:12
Neo: I'm not the one.	Non sono l'Eletto. [I'm not the one.]	L'Eletto non sono io. [The one is not me.]
95.	00:08:05:14	00:08:07:04
Oracle: Sorry, kid.	Mi dispiace, ragazzo. [I'm sorry, boy.]	Sono desolata, ragazzo. [I'm terribly sorry, boy.]
96.	00:08:09:02	00:08:11:00
You got the gift…	Tu hai il talento… [You have the talent…]	Il talento ce l'hai… [The talent, you have that…]
97.	00:08:12:04	00:08:14:05
but it looks like you're waiting for something.	ma sembra che tu stia aspettando qualcosa. [but it looks like you're waiting for something.]	ma è come se stessi aspettando qualcosa. [but it is as if you were waiting for something.]
98. **AOI.85**	00:08:19:11	00:08:21:22
Neo: What? *Oracle*: Your next life maybe.	- Cosa? [What?] - La tua prossima vita, forse. [Your next life, maybe.]	- Cosa? [What?] - Una vita futura, magari. [A future life, perhaps.]
99.	00:08:21:24	00:08:23:03
Who knows?	Chi lo sa? [Who knows?]	Chissà! [Who knows!]
100.	00:08:23:18	00:08:26:08
That's the way these things go.	È così che vanno queste cose. [It's in this way that these things go.]	Certe volte le cose vanno così. [Sometimes things go this way.]
101.	00:08:26:23	00:08:29:04
Oracle: What's funny? *Neo*: Morpheus.	- Cosa c'è da ridere? [What's to be laughed (about)?] - Morpheus.	- Perché ridi? [Why are you laughing?] - Morpheus.
102.	00:08:30:03	00:08:31:01
He...	Lui… [He...]	Lui… [He...]
103. **AOI.89**	00:08:32:22	00:08:34:15
he almost had me convinced.	mi aveva quasi convinto. [he almost convinced me.]	mi aveva quasi persuaso. [he almost persuaded me.]
104.	00:08:34:17	00:08:35:18
Oracle: I know.	Lo so. [I know.]	Già. [Right.]
105.	00:08:37:15	00:08:39:18

Poor Morpheus.	Povero Morpheus. [Poor Morpheus.]	Povero Morpheus. [Poor Morpheus.]
106. **AOI.91**	time-in: 00:08:42:01	time-out: 00:08:45:04
Without him we're lost.	Senza di lui siamo perduti. [Without him we're lost.]	Se non ci fosse, saremmo perduti. [If he wasn't there, we would be lost.]
107.	00:08:46:12	00:08:48:20
Neo: What do you mean, without him?	In che senso: 'senza di lui'? [In what sense 'without him'?]	Come 'se non ci fosse'? [how (do you mean), 'if he wasn't there'?]
108.	00:08:52:22	00:08:54:19
Oracle: Are you sure you want to hear this?	Sei sicuro di volerlo sentire? [Are you sure (you want) to hear it?]	Vuoi proprio che te lo dica? [Do you really want me to tell you?]
109. **AOI.94**	00:08:57:00	00:08:59:18
Morpheus believes in you, Neo.	Morpheus crede in te, Neo. [Morpheus believes in you, Neo.]	Morpheus ha fiducia in te, Neo. [Morpheus has faith in you, Neo.]
110.	00:09:01:12	00:09:06:10
And no one, not you, not even me, can convince him otherwise.	E nessuno, né tu, né io, può convincerlo del contrario. [And no one, nor you, or me, can convince him of the contrary.]	E nessuno, inclusi noi, può fargli cambiare idea. [And no one, included us, can change his mind.]
111. **AOI.96**	00:09:06:12	00:09:09:04
He believes it so blindly	Ci crede così ciecamente [He believes it so blindly]	Ci crede così intensamente [He believes it so intensely]
112.	00:09:09:06	00:09:13:13
that he's going to sacrifice his life to save yours.	che sacrificherà la sua vita per salvare la tua. [that he will sacrifice his life to save yours.]	che rinuncierà alla sua vita per salvare la tua. [that he will give up his life to save yours.]
113.	00:09:13:15	00:09:14:18
Neo: What?	Cosa? [What?]	Come? [How (do you mean)?]
114. **AOI.99**	00:09:14:20	00:09:17:17
Oracle: You're going to have to make a choice.	Dovrai fare una scelta. [You will have to make a choice.]	Dovrai scegliere. [You will have to choose.]
115.	00:09:18:10	00:09:22:10

In the one hand you'll have Morpheus' life.	In una mano avrai la vita di Morpheus. [In one hand you'll have Morpheus' life.]	Da una parte ci sarà la vita di Morpheus. [On one side there will be Morpheus' life.]
116.	time-in: 00:09:22:12	time-out: 00:09:25:07
And in the other hand you'll have your own.	E nell'altra avrai la tua. [And in the other you will have yours.]	E dall'altra ci sarà la tua. [And on the other there will be yours.]
117.	00:09:26:15	00:09:28:20
One of you is going to die.	Uno di voi morirà. [One of you will die.]	Uno di voi è destinato a morire. [One of you is destined to die.]
118.	00:09:31:04	00:09:34:09
Which one will be up to you.	Quale dei due dipenderà da te. [Which one (of the two) will depend on you.]	Quale dei due dovrai deciderlo tu. [Who between the two, you will have to decide.]
119.	00:09:35:19	00:09:37:21
I'm sorry, kiddo, I really am.	Mi dispiace, ragazzo. Davvero. [I'm sorry, boy. Really.]	Sono desolata, ragazzo. Sul serio. [I'm terribly sorry, boy. Seriously.]
120. **AOI.105**	00:09:37:23	00:09:39:13
You have a good soul.	Hai un buon animo. [You have a good soul.]	Hai un buon cuore. [You have a good heart.]
121.	00:09:40:16	00:09:43:09
And I hate giving good people bad news.	E io odio dare brutte notizie alle persone buone. [And I hate giving bad news to good people.]	Non amo dare cattive notizie alle persone in gamba. [I don't like giving bad news to respectable people.]
122.	00:09:46:15	00:09:48:00
Don't worry about it.	Non preoccuparti. [Don't worry.]	Non fartene un cruccio. [Don't make it a worry for yourself.]
123.	00:09:48:02	00:09:50:21
As soon as you step outside that door,	Non appena uscirai da quella porta, [As soon as you get out through that door,]	Non appena varcherai quella soglia, [As soon as you walk through that doorway,]
124.	00:09:50:23	00:09:52:17
You'll start feeling better.	inizierai a sentirti meglio. [you'll start to feel better.]	ti sentirai meglio. [you'll feel better.]

125.	time-in: 00:09:53:21	time-out: 00:09:58:08
You'll remember you don't believe in any of this fate crap.	Ti ricorderai che non credi a nessuna di queste sciocchezze sul destino. [You'll remember that you don't believe in any of this trifle about destiny.]	Ti ricorderai di non dare credito a queste sciocchezze da chiromante. [You'll remember not to give credit to this fortune-teller rubbish.]
126.	00:09:58:10	00:10:01:23
You're in control of your own life,	Hai tu il controllo della tua vita. [You have control of your life.]	Sei tu il padrone della tua vita. [You are the master of your life.]
127.	00:10:02:00	00:10:03:12
remember?	Ti ricordi? [Do you remember?]	Dico bene? [Am I right?]
128.	00:10:05:14	00:10:07:24
Here.	Ecco qua. [Here.]	Tieni. [Have (this).]
129.	00:10:08:01	00:10:09:16
Take a cookie.	Prendi un biscotto. [Take a biscuit.]	Prendi un dolcetto. [Take a sweet.]
130.	00:10:10:05	00:10:14:01
I promise, by the time you're done eating it,	Ti prometto che non appena avrai finito di mangiarlo, [I promise you that as soon as you have finished eating it,]	Ti prometto che quando l'avrai finito, [I promise you that when you have finished it,]
131.	00:10:14:03	00:10:16:20
You'll feel right as rain.	ti sentirai in gran forma. [you'll feel in great shape.]	ti sentirai come nuovo. [you'll feel (as if) new.]
132.	00:10:26:00	00:10:29:23
Morpheus: What was said was for you	Quel che è stato detto era per te [What was said was for you]	Quello che ti ha detto era per te [What she told you was for you]
133.	00:10:30:00	00:10:31:09
and for you alone.	e per te soltanto. [and for you only.]	e per nessun'altro. [and for no one else.]

B.2 Ethical Approval Form

Performance, Governance and Operations
Research & Innovation Service
Charles Thackrah Building
101 Clarendon Road
Leeds LS2 9LJ Tel: 0113 343 4873
Email: ResearchEthics@leeds.ac.uk

Valentina Ragni
Research student
School of Modern Languages and Cultures
University of Leeds
Leeds, LS2 9JT

PVAR Faculty Research Ethics Committee
University of Leeds

30 November 2016

Dear Valentina

Title of study: **Subtitling in the foreign language classroom**
Ethics reference: **LTSMLC-013**

I am pleased to inform you that the above application for light touch ethical review has been reviewed by a School Ethics Representative of the PVAC and Arts (PVAR) joint Faculty Research Ethics Committee. I can confirm a favourable ethical opinion on the basis of the application form as of the date of this letter.

The following documentation was considered:

Document	*Version*	*Date*
LTSMLC-013 LightTouch Ethical Review Form_VR (1).jpg	1	25/02/13
LTSMLC-013 LightTouch Ethical Review Form_VR (2).jpg	1	25/02/13
LTSMLC-013 LightTouch Ethical Review Form_VR (3).jpg	1	25/02/13
LTSMLC-013 LightTouch Ethical Review Form_VR (4).jpg	1	25/02/13
LTSMLC-013 LightTouch Ethical Review Form_VR (5).jpg	1	25/02/13
LTSMLC-013 LightTouch Ethical Review Form_VR (6).jpg	1	25/02/13

Please note: You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited.

Yours sincerely

Jennifer Blaikie
Senior Research Ethics Administrator
Research & Innovation Service
On behalf of Dr William Rea, Chair, PVAR FREC

CC: Student's supervisor