Dublin City University

A Reception Study of Machine Translated Subtitles for MOOCs

Ke Hu
B.A., M.E.

Thesis submitted for the degree of Doctor of Philosophy

School of Applied Language and Intercultural Studies

Jan 2020

Supervisors:

Prof Sharon O'Brien, Prof Dorothy Kenny

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____

(Candidate)

ID No.: _____

Date: _____

Acknowledgements

It is a genuine pleasure for me to express my sincere gratitude to my supervisors Prof Sharon O'Brien and Prof Dorothy Kenny for making my wish to study here come true. Their scholarly advice, prompt inspirations, meticulous scrutiny, and kind encouragement have helped me to a very great extent to accomplish this task.

I am extremely indebted to Dr Jinhua Du, Dr Longyue Wang, and Tianbo Ji for all their IT assistance with my experiment and data analysis. They kindly donated a lot of time for my research, and their advice helped me to implement new ideas from a different perspective. I would also like to thank Dr Carlos Teixeira for his instructions on eye-tracking.

My deep gratitude goes to all the participants in this research who kindly accepted to participate in this experiment.

I thank the ADAPT Centre for letting me be a part of it and start a new life on this beautiful island.

I thank my friends Liting and Marissa for making me feel at home thousands of miles away. I thank my friends Daria and David for all the chats and craic throughout the four years.

My utmost gratitude belongs to my open-minded father, who has always been supporting me and giving me the freedom to do what I want.

A special thank you goes to Cian for letting me know that there are miracles in the world. Thank you for loving me, supporting me, encouraging me, and making me laugh!

Finally, I would like to thank myself for my perseverance, positive attitude and overcoming all difficulties in this PhD journey.

Doing this PhD has been an invaluable and fulfilling experience in my life. It has changed me in so many ways and I am grateful to everyone I have met during this time.

前路漫漫，昂首挺胸，英姿飒爽，克敌制胜。

Publications and Presentations from this Research

Publications:

Hu, K., O'Brien, S. and Kenny, D. (2019) 'A reception study of machine translated subtitles for MOOCs', *Perspectives*, pp. 1-18.

Toral, A., Castilho, S., Hu, K. and Way, A. (2018) Attaining the unattainable? Reassessing claims of human parity in neural machine translation, *Proceedings of the Third Conference on Machine Translation*, *Volume 1: Research Papers*. Brussels, Belgium, 31 October – 1 November 2018. Association for Computational Linguistics, pp. 113–123. Available at: https://arxiv.org/abs/1808.10432 (Accessed: 13 October 2019)

Hu, K. (2018) 'MT use in China', *MultiLingual, 29*(6), pp. 32-38.

Hu, K. (2016) 'A Comparative Study of Post-editing Guidelines', *Multilingual*, *27*(7), pp. 53-57.

Hu, K. and Cadwell, P. (2016) 'A comparative study of post-editing guidelines', *Baltic Journal of Modern Computing, 4*(2), pp. 346-353.

Presentations:

Hu, K. O'Brien, S. and Kenny, D. (2019) 'Quality Assessment of Machine Translated and Human Translated Subtitles for MOOCs', *Translation Technology in Education – Facilitator or Risk?* 5 Jul 2019, Nottingham, UK.

Hu, K. O'Brien, S. and Kenny, D. (2018) 'Machine Translation for Subtitling - a way to improve access to MOOCs', *UMAQ (Understanding Media Accessibility Quality) Conference 2018*, 4 – 5 June 2018, Barcelona, Spain.

Hu, K. O'Brien, S. and Kenny, D. (2017) 'A Reception Study of Machine Translated Subtitles for MOOCs', *The 16th Machine Translation Summit*, 18 – 22 September 2017, Nagoya, Japan.

Hu, K. and O'Brien, S. (2016) 'Applying TAM (Technology Acceptance Model) to Testing MT Acceptance', *12th International Postgraduate Conference in Translation and Interpreting (IPCITI)*, 12 – 13 December 2016, Dublin, Ireland.

Hu, K. and O'Brien, S. (2016) 'Applying TAM (Technology Acceptance Model) to Testing MT Acceptance', *TranslatingEurope Forum 2016*, 27 – 28 October 2016, Brussels, Belgium.

Hu, K. and Cadwell, P. (2016) 'A Comparative Study of Post-editing Guidelines', *The 19th Annual Conference of the European Association for Machine Translation (EAMT)*, 30 May – 1 June 2016, Riga, Latvia.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	Analysis of Variance
AOI	Area of Interest
AOI_IMA	Image AOI
AOI_SUB	Subtitle AOI
ASR	Automatic Speech Recognition
BLEU	Bilingual Evaluation Understudy
DQF	Dynamic Quality Framework
EN	English
HT	Human Translation
HTER	Human-targeted Translation Edit Rate
IAA	Inter-Annotator Agreement
ISO	International Organization for Standardization
LISA	Localization Industry Standards Association
LSD	Least Significant Difference
LSP	Language Service Provider
MOE	Ministry of Education
MOOC	Massive Open Online Course
MQM	Multidimensional Quality Metrics
MT	Machine Translation
OOPS	Opensource Opencourseware Prototype System
PE	Post-Editing
PEOU	Perceived Ease of Use
PU	Perceived Usefulness
QA	Quality Assessment
RBMT	Rule-Based Machine Translation
RTF	Real Time Factor
SMT	Statistical Machine Translation
TAM	Technology Acceptance Model
TAUS	Translation Automation User Society
TER	Translation Edit Rate
TM	Translation Memory
TQA	Translation Quality Assessment
VNC	Visual Nonverbal Cues
WER	Word Error Rate
ZH	Chinese

“It takes backbone to lead the life you want.”

Revolutionary Road

Abstract

A Reception Study of Machine Translated Subtitles for MOOCs
Ke Hu

As MOOCs (Massive Open Online Courses) grow rapidly around the world, the language barrier is becoming a serious issue. Removing this obstacle by creating translated subtitles is an indispensable part of developing MOOCs and improving accessibility. Given the large quantity of MOOCs available worldwide and the considerable demand for them, machine translation (MT) appears to offer an alternative or complementary translation solution, thus providing the motivation for this research.

The main goal of this research is to test the impact machine translated subtitles have on Chinese viewers' reception of MOOC content. More specifically, the author is interested in whether there is any difference between viewers' reception of raw machine translated subtitles as opposed to fully post-edited machine translated subtitles and human translated subtitles.

Reception is operationalized by adapting Gambier's (2007) model, which divides 'reception' into 'the three Rs': (i) response, (ii) reaction and (iii) repercussion. Response refers to the initial physical response of a viewer to an audio-visual stimulus, in this case the subtitle and the rest of the image. Reaction involves the cognitive follow-on from initial response, and is linked to how much effort is involved in processing the subtitling stimulus and what is understood by the viewer. Repercussion refers to attitudinal and sociocultural dimensions of AVT consumption. The research contains a pilot study and a main experiment. Mixed methods of eye-tracking, questionnaires, translation quality assessment and frequency analysis were adopted. Over 60 native Chinese speakers were recruited as participants for this research. They were divided into three groups, those who read subtitles created by raw MT, post-edited MT (PE) and human translation (HT). Results show that most participants had a positive attitude towards the subtitles regardless of their type. Participants who were offered PE subtitles scored the best overall on the selected reception metrics. Participants who were offered HT subtitles performed the worst in some of the selected reception metrics.

Chapter 1 Introduction

1.1 Motivation

This section outlines the motivation for the component parts of the research described in this thesis, i.e.: MOOCs, MT and the concept of reception.

1.1.1 Why MOOCs?

Massive Open Online Courses, or 'MOOCs', are increasingly being used by universities, companies and organizations globally. On the one hand, they are used to promote courses and share knowledge resources; on the other hand, they showcase brands and attract more students. Through the years, popular platforms such as Coursera,[1] edX,[2] Udacity,[3] XuetangX[4] and FutureLearn[5] have partnered with hundreds of universities to offer thousands of courses to millions of users. By the end of 2018, the number of MOOC learners had reached 101 million, and over 900 universities had announced or launched 11,400 MOOCs (Shah, 2018). With the development of MOOCs, many problems have been exposed. Taking some recent studies as examples, a study by the Massachusetts Institute of Technology (MIT) shows that MOOCs struggle to lift rock-bottom completion

[1] www.coursera.org (Accessed: 5 December 2019)
[2] www.edx.org (Accessed: 5 December 2019)
[3] www.udacity.com (Accessed: 5 December 2019)
[4] www.xuetangx.com (Accessed: 5 December 2019)
[5] www.futurelearn.com (Accessed: 5 December 2019)

rates as the researchers found that the average dropout rate for online courses over the past five years was about 96% and this figure had not improved between 2013-14 and 2017-18 (Murray, 2019). Another recent study by the University of Michigan shows that providing certificates for a fee can boost MOOC completion rates (Sriram, 2019). However, there are objections that charging fees can reduce access to education and detract from the spirit of the first MOOCs, and make MOOCs less massive and less open (Murray, 2019). Other well-known issues with MOOCs are related to intellectual property, ethics, the language barrier, etc. Among them, the language barrier is the focus of this thesis.

It is a fact that the vast majority of MOOCs are produced and made available in English nowadays. While English accounts for about 50% of Internet content and English speakers only make up 30% of the total users, over 80% of the courses on Coursera and edX are instructed in English, which is against the mission statements of Coursera and edX, whose mottos are "anyone, anywhere", [6] and "everyone, everywhere", [7] respectively (Agudo, 2019). Due to the shortage of labour and capital, only selected courses are being translated into selected languages, which can lead to imbalance in the subjects which are translated. Removing this language barrier, is of key importance to the development of MOOCs worldwide. Taking China as an example, MOOCs have become an important pillar in China's thriving e-learning industry: according to its Ministry of Education, China is now the world's largest MOOC provider in terms of volume (MOE, 2018). But despite the growth in Chinese-language MOOCs, a

[6] blog.coursera.org/about/ (Accessed: 3 March 2019)
[7] edx.org/about-us (Accessed: 3 March 2019)

considerable number of MOOCs accessed in China are in other languages, due to collaboration between Chinese and foreign MOOC platforms or universities. For instance, one of the leading MOOC platforms in China, XuetangX, is in partnership with edX and provides MOOCs from Stanford University, Queensland University, and the University of California, Berkeley, among others. The use of MOOCs in languages other than Chinese can present problems, however. In one survey on MOOC usage in China that attracted some 3,300 responses (Guokr MOOC Academy, 2014), language was cited as an important barrier to learning via MOOCs: of the more than 2,440 respondents who answered that they had tried MOOCs before, 47% claimed that the obstacle that stopped them learning via MOOCs was 'language'. For the respondents who answered they had not tried MOOCs before, 17.5% gave the 'language barrier' as the reason. There appears therefore to be scope for translation in the popularisation of MOOCs in China. Volunteer groups clearly have a role to play in such translation (Beaven et al., 2013), but there is still a large unmet demand for translated MOOCs.

1.1.2 Why MT?

The globalization of the world and the acceleration of internationalization, especially with the deepening of China's "One Belt and One Road" initiative, have promoted international political, economic, and cultural exchanges, thereby enhancing the world-wide demand for language and translation services. Meanwhile, the rapid development of the Internet, big data, cloud computing, and other modern technologies have also brought opportunities for MT (machine translation). In regard to the considerable

demand for translating MOOCs, given limited availability of human translators and budget for translating content, MT could be part of the solution.

Over the past decade or so, several projects have addressed subtitling by MT, including MovRat (Armstrong et al., 2006; Flanagan, 2009), Volk's 'Stockholm system' (Volk, 2008), SUMAT (Etchegoyhen et al., 2014) and DialogueMT (Wang et al., 2016). Research in the area has been supported by the increasing availability of large parallel corpora of subtitles (e.g.: the OpenSubtitles corpus in Lison and Tiedemann, 2016), However, these projects mostly focus on films and TV shows, and subtitling MOOCs by MT is a fledging field. That said, the recent TraMOOC project (Kordoni et al., 2016) focuses on the MT of subtitles for MOOCs. The project established an open-source online neural MT platform (http://www.translexy.com/) that can automatically translate English-language MOOCs into 11 languages including Chinese. Despite such initiatives, MT still has an image problem and is not always trusted by end users. To improve the quality of MT output, and to establish trust among human subtitlers and end-users, post-editing could be a solution. Several sources (Plitt and Masselot, 2010; Aranberri et al., 2014) have reported productivity increases when using post-editing compared to traditional human translation. Castilho et al. (2014) further show that post-editing significantly increases the usability of machine-translated (online help) text. Their work sheds light on how end-users engage with raw and post-edited machine-translated text, which is also the focus of this research.

1.1.3 Why reception?

According to Gambier (2018), reception studies of translated audio-visual texts are relatively scant. Apart from that, most of them are about films and TV shows, none of them is about the subtitles of MOOCs. As MOOCs are becoming more and more popular these days, more research is needed to explore the use of MOOCs and users' reception of the translated subtitles for MOOC videos. Another problem of the existing reception studies is that they mainly focus on studying viewers' reception of human-translated subtitles. While MT has gained considerable attention in both academia and industry, the usability of machine translated subtitles and post-edited MT subtitles is worth investigating. In addition, as MT quality has been the focus of controversy, it would be a good idea to compare the reception of raw MT output, full post-edited output and human translation. Therefore, the researcher includes raw MT, full post-edited MT and human translation conditions in the research design. As for the language pair to be studied, although Chinese viewers have been exposed to subtitles for a long time and fansubbing is very popular in China, little research has been done on the viewer's reception of English-Chinese subtitles; thus this research focuses on studying the English-Chinese language pair.

In regard to the methodology for reception studies of subtitles, translation scholars often adopt eye-tracking technology, questionnaires and sometimes a mix-methods approach. Results show that these methods and the approach are effective, and hence they have been adopted in this research.

1.2 Thesis overview

In summary, by using a mixed-methods approach, this research investigates Chinese end users' reception of MOOC subtitles that have been translated from English into Chinese in three conditions: human translation (HT); raw, i.e. un-edited, MT (RAW); and post-edited MT (PE). The research aims to answer the following question: is there a difference in reception between participants who are offered raw MT subtitles and those who are offered full PE subtitles or HT subtitles?

This thesis is organised as follows: Chapters 2 and 3 constitute the literature review, the former focuses on MOOCs, subtitles and reception studies, and the latter on translation quality assessment; Chapter 4 outlines the mixed-methods approach of this research while Chapter 5 discusses the research design, including the set-up and outcome of the pilot study, and provides a brief introduction to the main experiment; Chapters 6, 7 and 8 are the data analysis part of the research, with a focus on questionnaire data (Chapter 6), eye-tracking data (Chapter 7), and the data collected from the Translation Quality assessment (TQA) and frequency analysis (Chapter 8) respectively; Chapter 9 presents the conclusions and suggestions for future work.

Chapter 2 MOOCs, Subtitles and Reception Studies

2.1 Introduction

This chapter reviews relevant literature on MOOCs, subtitling, MT and reception studies. First, Section 2.2 outlines the focus of this research: MOOCs, and presents the options for removing the language obstacle for better online knowledge sharing: human translation and MT. Section 2.3 gives an overview of MT of subtitles. Then, a discussion about the reception of subtitles is provided in Section 2.4. Finally, Section 2.5 gives the concluding remarks.

2.2 MOOCs

Education is important to people of all ages from all places. In the Information Age, education has become highly technologized. To improve student learning, researchers have been interested in integrating technology into education for a long time. A typical practice is e-learning, which means delivering courses through the Internet. E-learning is rooted in the concept of Computer-Assisted Learning (CAL), which can be dated back to "programmed instruction delivered via mainframe computers in the 1950s and 1960s" (Mayer, 2017, p.403). Apart from being increasingly used in schools, with the advance of technology, e-learning has developed into many creative forms, such as educational

network platforms that aim to help children study after class (e.g.: ABCmouse[8]), and online self-learning services that aim to train people in specific skills (e.g.: Khan Academy[9]). Over the years, one of the most important initiatives that has been introduced in the context of e-learning is 'massive open online courses', or MOOCs.

The term 'MOOC' was coined in 2008 by Dave Cormier of the University of Prince Edward Island in Canada (Cormier, 2008). Since then, MOOCs have received widespread attention in the USA especially since Sebastian Thrun, a Stanford professor, offered an online course on Artificial Intelligence for free (Hu, 2013). Currently, universities worldwide are using platforms such as Coursera, Udacity, and edX to offer MOOCs. Taking Coursera, one of the most notable MOOC platforms (Tahmassebi et al., 2018), as an example, it is offering 3,449 courses and has 185 partners (including universities and organizations) across 28 countries up to the date of March 3, 2019.[10] In terms of the number of registered users, the top five MOOC providers in 2018 are as follows (Shah, 2018): Coursera (37 million), edX (18 million), XuetangX (14 million), Udacity (10 million), and FutureLearn (8.7 million). Andrew Ng, founding lead of Google Brain, had enrolled over 1.7 million students in his first machine learning course by the end of Feb 2019.[11]

Stephen Downes and George Siemens distinguish two types of MOOCs which they call cMOOC and xMOOC (Nobre et al., 2018). They define the differences between them as follows: the "c" in cMOOC refers to connectivist while the "x" in xMOOC means extended (Jacoby, 2014). The former is based on the principles of connectivism,

[8] https://www.abcmouse.com/ (Accessed: 3 March 2019)
[9] https://www.khanacademy.org/ (Accessed: 3 March 2019)
[10] https://www.coursera.org/about/partners (Accessed: 3 March 2019)
[11] https://www.coursera.org/learn/machine-learning (Accessed: 3 March 2019)

openness, and participatory teaching (ibid.), whereas the latter is based on the Behaviourist Learning Theory; it is teacher-centered, and the student acts as the receiver of knowledge (Gao, 2014a). Panchenko (2013) summarizes the differences between the two types of MOOC, as presented in Table 2.1. xMOOCs are usually associated with Coursera and other popular MOOC platforms. The research scope for this PhD study is limited to xMOOCs.

Table 2.1 cMOOC and xMOOC comparison (Panchenko, 2013, p.157)

cMOOC	***xMOOC***
Knowledge is created and generated	Knowledge is replicated
Creativity	More traditional approach (video presentations, quizzes, and testing)
Not sponsored	Good financial support
Individual initiative of some members of educational community	Supported by prestigious universities
Large body of non-structured information	Course content is structured
Not controlled	Controlled
A team of volunteers	A team of co-workers

Costa-Jussà et al. (2014 and 2015) provide a good example of a MOOC relevant to this thesis. They explain the process of designing, developing, and analysing a MOOC on MT that is entitled “Approaches to Machine Translation: Rule-based, statistical and hybrid” in detail, including course topic and contents, activities, recording equipment, platform, and promotion.

2.2.1 Practical perspectives

Much of the literature on MOOCs comes in the form of overviews and case studies, with government and industry white papers featuring strongly in the Chinese context (these will be discussed in the next section). Academic papers have been published on the introduction and development of MOOCs. Across this literature a number of recurring themes can be discerned, the most relevant of which are addressed below. These include advantages of MOOCs, participation in MOOCs, and problems associated with MOOCs.

Panchenko (2013) participated in and analysed several MOOCs offered by Coursera, Udacity and edX in order to study the advantages of MOOCs as an alternative form of advanced training for university teachers. He studied the dynamics of Coursera, Udacity, and edX search engines with the help of the Google Trend Service. He conducted a comparative analysis of the three providers arguing that the advantages of MOOCs are as follows (ibid., p.168):

a) participation is free of charge;
b) opportunity for on-the-job advanced training;
c) introduction to individual teaching styles of the leading professors of some well-known universities;
d) possibility to compare the methodological basis of different courses;
e) distance learning experience in the role of a student;
f) participation in forum discussions;
g) experience of peer assessment practices;
h) broadening one's horizons and knowledge of teaching methodology;
i) opportunity to expand the courses taught at the university incorporating knowledge of MOOC;
j) intercultural competency development;
k) English listening, reading, and writing skills enhancement;
l) expanding the range of software;
m) establishing new professional contacts; and

n) reflecting on one's own educational activity from a new angle.

We can see that some elements in the list, such as b), d), e), f), h), j), m) and n), are not specific to MOOCs; other e-learning forms also contain these elements. Elements like i) and l) are unclear. Element k) is questionable because, although the majority of MOOCs on those well-known platforms are in English, the lingua franca in many countries, a large number of MOOCs are in other languages. Given the worldwide accessibility of MOOCs, language can be a barrier for online learning, a point to which we return below. Regarding Element a), up to now, most MOOCs on platforms like Coursera, Udacity, and edX are free to audit, but if students want their assignments graded or to earn a certificate, they will need to pay. Element g) is worth highlighting. On Coursera, for example, students have to do assignments and also review two or three assignments of their peers to progress. In this way, everyone has to complete assignments on time and can have them graded, creating a virtuous circle for online learning.

In China, Yuan and Liu (2014) investigated how six Chinese universities dealt with MOOCs and identified their specific concerns as follows: stressing the significance of educational reform based on MOOCs; focusing on self-promotion, supplemented by the support and guidance of government; introducing foreign MOOCs and producing domestic MOOCs; and sharing resources between universities and enterprises. The development of MOOCs in China is closely related to their promotion by the government. In recent years, the scale of state financial investment in education has continued to grow. The ratio of national fiscal expenditure on education to GDP has also steadily increased and exceeded 4.0% since 2012 (iiMedia, 2018). More information about Chinese government policies will be presented in Section 2.2.2.

The rapid development of MOOCs is based on the merits of MOOCs, but also on participants' willingness to use them. As Hew and Cheung (2014) point out, students and instructors join MOOCs for different reasons. For students, they join MOOCs because of curiosity and personal challenge, and to extend their knowledge and collect completion certificates. For instructors, they join MOOCs because of a sense of intrigue, or egoism, or altruism. Meanwhile, Hew and Cheung also point out the challenges for learning and teaching using MOOCs. As they put it, the nearly 90% drop-out rate in MOOCs is due to a lack of incentive, inability to understand the content and having no one to ask for help, and having other priorities to fulfil. Instructors find it hard to teach without students' presence and immediate feedback. Assessing students' work, a lack of student participation in online forums, and the high demand for time and energy, have also increased the difficulty of their job.

For MOOCs, high enrolments come with high drop-out rates. According to Hew and Cheung (2014, p.51), the main differences between MOOCs and traditional university courses are as follows: "the large and diverse student enrolment in MOOCs, the high drop-out rate of MOOCs compared to that of traditional courses, and the relatively [sic] lack of instructor presence or support in MOOCs compared to traditional courses". Low completion rates are also a big problem for MOOCs in China (Gao, 2014a). According to the *2016 Educated Youth Report* (Guokr MOOC Academy, 2017), jointly issued by several education websites in China, 14.18% of learners completed all their online courses, 33.23% of learners completed most courses and gave up a few, 21.52% of learners completed around half of the courses, and 4.74% of learners did not complete any course. We can see from Hew and Cheung's work (ibid.) that an important reason for

people to join MOOCs is curiosity. However, curiosity can be short-lived. For example, the *2016 Educated Youth Report* shows that 15.4% of respondents enrolled in a course "because of curiosity" but didn't actually want to complete the course. The most cited reason for dropping-out was "terrible self-discipline and procrastination", which accounts for 55.88%. Another reason - "too busy and no time to complete the course" - accounts for 53.63%. MOOC users come from different backgrounds, which may lead to different user behaviour. Jia et al. (2019a) conducted a study on a Chinese MOOC platform with 57 "blended learners" (those who studied both in college and via MOOCs) and 4,049 "social learners" (those who studied only via MOOCs) to find out who can benefit more from MOOCs. Results show that the course completion rates for blended learners and social learners are 100% and 7.14% respectively, and blended learners benefit more from MOOCs than social learners. The disparity in the number of participants between the two groups may have some influence on the results, but the huge gap between 100% and 7.14% elucidates the problem to some extent.

Apart from problems on the user side, MOOCs have other issues. Gao (2014b) finds that currently MOOCs are beset by a number of problems: lack of creativity in the mode of teaching (still based on outdated behaviourist teaching theory), difficulty to achieve personalised learning, lack of real learning experience for students (such as that based on interaction between teacher and student), difficulty in assessing student work, and lack of official acknowledgement for the learning outcome. Qi (2019) also points out the current status and problems of MOOCs in China, for example, the teaching quality cannot be guaranteed, credit certification is confusing and not standardized, copyright issues exist etc.

As for other sceptics regarding MOOCs, Jacobs (2013) is concerned that MOOCs would offer a watered-down education, have a negative impact on less prestigious education institutes, and raise the risk of further cuts to public school budgets. Kolowich (2013) is not sure how viable the notion of awarding formal course credit to MOOC students is, since in the survey he conducted on 103 professors, 72% of the respondents held the opinion that students who succeeded in MOOCs did not deserve to get formal course credit from their home institutions. 66% of them believed that their home institutions would not grant formal credit to those students. Furthermore, the professors were worried that the person who submitted assignments on a MOOC was not actually the one who was enrolled in the course in question. Hew and Cheung (2014) believe MOOCs are a good platform for people who want to learn something but are not interested in receiving a credit. They conclude that MOOCs are merely another learning resource and they doubt they could replace the traditional mode of education. Haggard et al. (2013) agree that the value of MOOCs is called into doubt by many educators, and the conflicting views on MOOCs lead to a disparity between prestigious universities and small academies, which may cause a potential threat to higher education. Nevertheless, they emphasize that the feedback on learning experiences with MOOCs is positive. They hold that the borderless format of MOOCs will accelerate the globalization of higher learning content and accreditation systems.

2.2.2 MOOCs in China

China has a thriving e-learning industry which is keen to globalize. At present, the most popular Chinese MOOC platforms are as follows: 学堂在线 (http://www.xuetangx.com/),[12] 中国大学MOOC (http://www.icourse163.org/), 华文慕课 (http://www.chinesemooc.org/), and 好大学在线 (http://www.cnmooc.org/home/index.mooc).

In 2015, the Ministry of Education (MOE) launched several initiatives to stimulate the development of MOOCs. In October 2016, the MOE Research Centre for Online Education in China, in cooperation with the HCR Company (慧辰资讯), published the *White Paper on China's MOOC Industry 2016* (hereafter abbreviated as 2016 White Paper). According to the White Paper (HCR, 2016), Chinese MOOC platforms mainly focus on higher education. The development of Chinese MOOCs is led by universities, while Internet companies and online education companies establish platforms and cooperate with foreign platforms. More than 30 universities sponsored by Project 985[13] and a number of universities sponsored by Project 211[14] have participated in the establishment of MOOCs. In January 2018, the Ministry of Education claimed that China is the world's largest MOOC provider in terms of volume (MOE, 2018). Chinese colleges and universities have established over 10 MOOC platforms providing more than 3,200

[12] Or it can be called XuetangX

[13] Project 985 is a project for founding world-class universities in the 21st century conducted by the government of the People's Republic of China. On May 4, 1998, President Jiang Zemin declared that "China must have a number of first-rate universities of international advanced level", so Project 985 was launched (China Education Center, Online).

[14] Project 211 is the Chinese government's new endeavour aimed at strengthening about 100 institutions of higher education and key disciplinary areas as a national priority for the 21st century. There are 112 universities in the project 211 (ibid.).

courses from over 460 universities throughout the country, which have been accessed by 55 million people (ibid.). Meanwhile, the Ministry of Education launched 490 national premium quality online open courses, which were selected from the abovementioned 3,200 MOOCs (ibid.). This was the first time that MOOCs were certified at the national level.

As the 2016 White Paper suggests (the data in this and the following paragraphs is all extracted from this source), the users of MOOCs in China are mainly 18 – 35 year olds. Among them, 47.9% are aged between 18 - 25, 20.4% are aged between 26 - 30, and 15.2% are aged between 31 – 35. It seems that men are more interested in MOOCs than women, since 60.1% of users are male while 39.9% are female. Due to an imbalance in levels of overall development in different regions in China, it is not surprising that most users are from first and second-tier cities: 29.2% of users live in first-tier cities, 47.5% in second-tier cities, and 23.3% in third and fourth-tier cities.[15] Due to the popularity of computers and easy access to the Internet, 61.5% of users learn about MOOCs from educational websites. 37.1% of users are introduced to MOOCs through their teachers, classmates or friends, and 28.3% of users learn about MOOCs through search sites and application stores. The most popular MOOCs were those on professional skills, taken by 61% of users.

[15] The tier system in China was introduced in the 1980s as a way of facilitating urban development. In the beginning, cities were ranked by tier according to the government's development priorities (Starmass, 2019), but now ranking has become a proxy for demographic and social segmentation in China (Liu et al., 2014). In May 2019, CBN Weekly, a business news magazine in China, issued a new list of Chinese tier cities. They investigated 337 cities and graded them according to five dimensions: quantity of commercial resources, transport hub or not, activeness of urban people, diversity in lifestyles, and future plasticity. The first-tier cities in this list are, for example, Beijing, Shanghai, Guangzhou, Shenzhen along with 15 more cities (CBN Weekly, 2019).

MOOCs are popular mostly among well-educated people, as 38.5% of users are students, 26.5% are white-collar office workers, and 11.5% are middle-senior managers. In addition, 68.6% of users have a Bachelor's degree, 12.5% have a Master's degree, and 12.3% are Junior College[16] graduates. Compared to the reasons for joining MOOCs summarized by Hew and Cheung (2014) mentioned in Section 2.2.1, the White Paper reveals that Chinese MOOC users have similar motivations: some people use MOOCs out of their desire for new knowledge and self-development, some use them for specific interests and hobbies, and others use them due to the influence of their positive experience of watching TED talks and other online open courses. Knowledge improvement is an important reason for learning through MOOCs. Students hope to access better resources and teachers in famous universities, while employees hope to gain more skills. Regarding the length of video used in MOOCs, 87% of users preferred the MOOC to be 10 – 20 minutes long.

In December 2018, iiMedia Research Group, a mobile Internet organization in China focused on third-party data mining and integrated marketing, released its *2018 China Online Education Industry White Paper* (iiMedia, 2018. Hereafter abbreviated as 2018 White Paper). Here, 'online education' equals e-learning, which not only includes MOOCs, but also other types of online learning sites or apps (especially those platforms designed for preschool education, secondary education, and interests and hobbies education). According to this White Paper, 'AI + education' has become the keyword of the online education industry in 2018. AI technologies such as intelligent-checking of homework, face recognition technology, personalized recommendations, and AI

[16] Junior college program is a form of higher education in China and usually lasts 2-3 years (Embassy of the People's Republic of China in Ireland, 2019).

teachers have been applied in the multiple scenarios of online education to continuously improve user experience.

This White Paper also suggests that online learners are mainly distributed in first-tier cities in China. Regarding the gender of learners, males (59.2%) are more active in learning than females (40.8%). As for the age, Chinese online learners are mainly 16 – 35 year olds. These findings are in line with those of the 2016 White Paper. In regard to fees, the 2016 White Paper states that 22.4% of Chinese users had experience of paying for MOOCs. Their low cost could be a potential reason for this. For instance, 71.7% of the prices were below 300 yuan (approximately 40 euros), and the price paid by students was mainly below 100 yuan (approximately 13.5 euros). The 2018 White Paper suggests that nearly 80% of online education users intend to spend more than 200 yuan (approximately 26 euros) per month. Among them, 36.8% of online education users intend to spend between 500 yuan (approximately 65 euros) and 1,000 yuan (approximately 132 euros) per month. We can see that both the number of paying users and the fees themselves have increased over the two years. According to the 2016 White Paper, the anticipated number of registered users on Chinese MOOC platforms is more than 10 million for the year ending 2016. According to the updated information, by the end of 2018, the number of registered users of XuetangX alone had already reached 14 million. According to the 2018 White Paper, the number of online education users in China is expected to reach 296 million by 2020, which is nearly 30 times more than that anticipated in the 2016 White Paper (note: the figure in the 2018 White Paper includes but is not limited to MOOC users).

2.2.3 The language barrier

Despite the growth in Chinese-language MOOCs, a considerable number of MOOCs accessed in China are in other languages due to the collaboration between Chinese MOOC platforms and foreign MOOC platforms or universities. For example, XuetangX is in partnership with edX and also provides MOOCs from Stanford University, Queensland University, University of California Berkeley, etc. Wu and Bai (2018) emphasize the impact of low English level on MOOC learners. They found that most Chinese learners' English level is not high, while the required English level is high in English-speaking MOOCs. Chinese learners could not communicate easily with other foreign learners in the forums, and the learners found it difficult to learn from English MOOCs even if they had English subtitles. Wu and Bai (ibid.) argue that the language barrier reduced learners' interest in learning and is an important factor in the high dropout rate.

The results of other surveys are consistent with the argument of Wu and Bai (2018). In the *2014 MOOC Learners Survey* (Guokr MOOC Academy, 2014), approximately 3,300 responses were collected and 74% of the respondents indicated that they had taken MOOCs before. Of relevance here is that for both users and non-users of MOOCs, language was cited as an important barrier to learning via MOOCs. For the respondents who answered they had tried MOOCs before, 47% of them claimed that the obstacle that stopped them learning via MOOCs was 'language'. For the respondents who answered they had not tried MOOCs before, 17.5% of them chose the 'language barrier' as the reason for not learning via MOOCs. In the *Online Learning Survey 2015*, conducted by Guokr MOOC Academy (2015) in conjunction with several other education platforms,

for those who had not taken online courses, the reason "cannot adapt to English courses" accounts for 31%, while for those who had taken but not completed the course, the reason "language barrier" accounts for 11%. In addition, 82% of learners said they had to have the Chinese subtitles to keep up with the course.

The 2016 White Paper shows that among the factors that influence users' selection of MOOC platforms, "availability of Chinese courses" is selected by 29.6% of users, and "whether I can improve my English here" by 13.4% of users.[17] This is worth highlighting because learning English is very popular in China. English is a compulsory subject for Chinese secondary schools and even a subject in some kindergartens. Many people try to make use of all the available resources to learn English. As time goes by, rather than dubbing, adding subtitles to English-speaking videos is becoming more popular (Wang, 2013; Qian, 2014).

In summary, language is an obstacle for Chinese people to learn via MOOCs, and removing this obstacle, by, for example, creating translated subtitles, is an indispensable part of developing MOOCs and broadening access. In fact, not only in China, but also in other countries, the translation of subtitles is an important demand made by users. An interesting example of this is the MOOC entitled "Accessibility to the Scenic Arts",[18] developed under the ACT (Accessible Culture and Training) project,[19] which aims to improve the accessibility of cultural events to all people. While different languages (Spanish, Dutch/Flemish, English, German) are used for instruction in this MOOC, all of its videos have English subtitles (Orero, 2018). For those videos in English, automatic

[17] Note: all the quotes in this section are originally in Chinese and translated by the researcher.
[18] https://www.coursera.org/learn/accessibility-scenic-arts (Accessed: 9 June 2019)
[19] http://pagines.uab.cat/act/ (Accessed: 9 June 2019)

subtitles are provided; for those not in English, some of them have translated English subtitles only (see Figure 2.1), while others have embedded same language subtitles and users can choose to turn on/off the translated English subtitles (see Figure 2.2).

Figure 2.1 Translated English subtitles for Spanish-speaking video[20]

[20] https://www.coursera.org/learn/accessibility-scenic-arts/lecture/WdZmu/coordination-and-unexpected-problems (Accessed: 9 June 2019)

Figure 2.2 Embedded same language subtitles (green colour) for Flemish-speaking video along with translated English subtitles (white colour)[21]

In the discussion forum of this MOOC, a user commented:

> "I've got a little German, and English at C2 level, so if the videos in Flemish are subtitled, I can sort of understand what is being said. But when there are no subtitles, I can grasp a word here and there, but not enough to comprehend the details. Is there a way to get English or French subtitles? Even German would be better than none! Help, please? These videos are so informative, so well written and filmed!"[22]

While a discussion of the similarities between languages is beyond the scope of this thesis, it is clear that the translation of subtitles is an important requirement of MOOC users and a necessary part of creating and disseminating MOOCs.

[21] https://www.coursera.org/learn/accessibility-scenic-arts/lecture/sQk63/mini-docu-ad-in-the-theatre (Accessed: 9 June 2019)
[22] https://www.coursera.org/learn/accessibility-scenic-arts/discussions/weeks/3/threads/E9JJmetmEeixpA5S2n-CNg (Accessed: 9 June 2019)

2.2.4 Translating MOOC subtitles

From the three surveys by Guokr MOOC Academy and the two White Papers mentioned in the previous section, we can see that China has a big market for MOOCs. Since their introduction to China in 2013, MOOCs have developed very quickly. Because of the language barrier, translation plays a significant role in making MOOCS more accessible. Thus far, a number of attempts have been made to translate MOOCs, either by human translation or MT. These are discussed below. Translation quality will be frequently mentioned in this section, but will be elaborated upon in Chapter 3.

2.2.4.1 Human translation

As one of the leading MOOC platforms in the world, Coursera now has a well-established global translator community, which consists of Coursera learners who translate course subtitles from English into over 65 languages.[23] Anyone who is fluent in English and one other language can join this community. The Coursera Chinese community landing page has detailed instructions for people who want to be translators and a number of translation rules. Translators use Smartling, a cloud-based translation management platform that is integrated into Coursera, to work. Some other MOOC platforms work in

[23] https://coursera.community/gtc-news-announcements-17/join-the-coursera-global-translator-community-45 (Accessed: 5 March 2019)

partnership with crowdsourcing platforms for translating their MOOCs. For example, edX works with transifex,[24] and Udacity works with Amara.[25]

Another example is Opensource Opencourseware Prototype System (OOPS),[26] an independent grassroots project in Taiwan initially designed to translate and adopt MIT OpenCourseWare[27] (it now also includes some courses by Harvard University, Yale University and Utah State University) for the Great China Region, mainly run by volunteers from various disciplines worldwide. Similar to Coursera, OOPS also has a translation community. On the home page of OOPS, there is a clear entry point for people who want to join its translation community (see Figure 2.3). Once potential users have submitted the basic information online, including which language pair they can work with (English-Chinese or Japanese-Chinese), an editor will contact them directly. Lee et al. (2007) found that the success and sustainability of the community are related to three key issues: democratic leadership, participation incentives, and a forum for discussion and story sharing.

[24] https://www.transifex.com (Accessed: 5 March 2019)
[25] https://amara.org/en/ (Accessed: 5 March 2019)
[26] http://www.myoops.org (Accessed: 5 March 2019)
[27] A web-based publication of virtually all MIT course content (https://ocw.mit.edu/index.htm Accessed: 5 March 2019). It is open and available to the world and is a permanent MIT activity.

Figure 2.3 Home page of OOPS

Regarding the tools for translating MOOCs, the work of Beaven et al. (2013) may be insightful. They researched a MOOC on open translation tools and practices. The MOOC aims to help students learn a range of online open translation tools (Amara, Transifex, Google Translator Toolkit) for the crowdsourcing of translation, dubbing and subtitling. Participants in the MOOC undertake hands-on translating work (from and into Spanish/English, French/English, and Brazilian Portuguese/English) by using different translation tools. The findings show that while MOOCs do play a role in bringing communities together to some extent, they don't always achieve effective collaboration, so the results in translation output are variable and there are challenges in quality assurance. They concluded that crowdsourcing can be the only solution to meet the large demand for the open translation of MOOCs.

At this point in time, there is no evidence of professional translation of MOOCs in China. Rather, MOOCs are translated either by amateur subtitling groups who are in partnership with MOOC platforms or by crowdsourcing. For instance, the EduInfinity Translation Group (http://www.edu-infinity.org/) is a Chinese fansub group which focuses on translating subtitles for MOOCs. They cooperated with Coursera in 2013 and edX in 2014, providing Chinese subtitles for the two platforms. The content of 网易公开课 (http://open.163.com/), a Chinese website of open courses that is affiliated with the NetEase company, is similar to OOPS. However, unlike OOPS or Coursera, the website does not have an integrated translation platform, and the sources of its subtitles are unclear and varied. As Zhao (2013) put it, the website finds its translators either from fansub groups or by recruiting people who love subtitling and translating part-time or full-time for reimbursement. The translators usually focus on translating one type of course according to their expertise. After uploading their translated subtitles, the staff of the website will proofread them, record any errors in a form and send it to the translators for revision. Besides, the subtitles of some courses are just borrowed from the existing resources, as on the website there is a line saying: “The translation of some courses is reproduced from the free subtitles published by YYeTs, TLF and other fansub groups. NetEase disseminates and retains all subtitle copyright information. We sincerely thank them for their contributions.”

It can be seen that fansub groups play an important part in the translation of MOOCs in China. Fansubbing is a sub-set of audio-visual translation (AVT). As the name suggests, fansubbing refers to subtitling videos by fans. The emergence of fansubbing in the 1980s is a result of the work of Japanese anime fans, who produced and distributed subtitles

through anime clubs (Leonard, 2005). In 2006, Díaz-Cintas and Muñoz Sánchez defined a fansub as "a fan-produced, translated, subtitled version of a Japanese anime programme" (p.37). Looking back now, this definition is rather restrictive. As time went by, fansubbing was no longer limited to Japanese anime. It can now refer to any foreign TV show or movie, or any video of a superstar, as long as they have fans who are willing to translate the subtitles. Fansubbing is also called 'amateur subtitling' (Lepre, 2015) or 'non-professional subtitling' (Orrego-Carmona, 2015), which implies that fans do not have to be professional translators or receive any translation training. Thanks to the great accessibility of many means of interacting online and sharing different cultures in the 21st century, fans from around the world can easily and conveniently engage in collaborative work related to their idols or admired products, which leads to the proliferation of fansubbing. The fansubbing process usually involves raw video providers, translators, timers, typesetters, editors, proof-readers and encoders (Díaz-Cintas and Muñoz Sánchez, 2006). Usually, each role is played by a different person, and each member only completes the assigned task, but sometimes the same person can play different roles.

As Qi (2019) has pointed out about MOOCs, fansubbing is also entangled in copyright issues. A number of researchers have written about the legal issues that involve the copyright of releasing the subbed audiovisual programs (González, 2007; Rembert-Lang, 2010; He, 2014; Hsiao, 2014; Lee, 2014; Cai, 2015; Wang, 2015; Wongseree, 2019). The very act of translating copyright material may also be a breach of Intellectual Property Law, even before translated materials are distributed. The legal issues in fansubbing are very complex, because the distribution of anime, TV shows and movies is done via the

Internet, “a medium in which borders and nationalities” are “difficult” to delineate (Díaz-Cintas and Sánchez, 2006, p.45). The lack of enforcement of copyright laws is also a thorny problem, as different stakeholders hold different attitudes towards copyright (ibid.).

In China, people rely heavily on subtitles, both intralingual and interlingual. According to the experience of the researcher, almost all news, movies and TV shows have Chinese subtitles. As for Chinese fansubbing, Cai (2015) elaborates on its history in her study. She mentions BitTorrent, a forum providing the largest downloading and uploading of audiovisual products for Chinese netizens. This forum gained popularity between 2003 and 2004, giving rise to the emergence of fansubs in China (Hu, 2009). For instance, in 2006, the American TV series *Prison Break* was a big hit in China. Meanwhile, a report on Chinese fansubbers by Howard W. French (2006) was published in *The New York Times*. Cai mentions that both of these facts accelerated the development of fansubbing in China. The large number of fansubbing subtitles on the Internet has also become a valuable resource that can be used for constituting various corpora for training MT systems. For example, the OpenSubtitle corpus is a collection of parallel translated movie subtitles from http://www.opensubtitles.org/, which includes lots of fansubbing subtitles in various languages. The corpus (ZH-EN) has been used for the frequency analysis of the machine translated subtitles in this PhD research, see Sections 2.3.1 and 8.3 for more information.

2.2.4.2 MT and post-editing

MT has already been applied to language learning. For example, Zhang and Lu (2014) researched the application of MT in learning Chinese by Irish students. Briggs (2018) studied Korean students using MT to learn English. MT has also begun to play a role in the translation of MOOCs. This section provides an introduction to MT and post-editing. Section 2.3 will deal with MT of subtitles in depth.

Due to the fast-growing popularity of artificial intelligence, MT has gone through a series of changes and developments in the 21st century. To increase productivity, reduce cost and save time, MT has already been applied by a number of language service providers (LSPs). Milengo GmbH, a Berlin-based LSP reports in 2019 that compared to the traditional translation model, both cost and time-to-market is reduced by nearly 80% with managed MT.[28] MT is also booming in China, with various local offerings emerging in an endless stream (Hu, 2018). As Mike Dillinger (2016, p.4) stated in the keynote speech at the 12th conference of the Association for MT in the Americas, "75% of small and large LSPs have at least a year's experience with MT in 2016. In 2006, it was more like 5%." Koponen (2016, p.143) reports that "post-editing of MT is indeed already an established part of the translation workflow in many professional contexts". In the face of the large demand for MOOCs with short timelines, can MT be a solution? Before delving into this question, a brief introduction to MT and post-editing (PE) is presented.

[28] https://slator.com/sponsored-content/how-nmt-based-translation-services-can-reduce-enterprise-translation-costs-by-up-to-80/ (Accessed: 3 March 2019)

According to ISO 17100:2015,[29] machine translation is "automated translation of text or speech from one natural language to another using a computer system", while post-editing means to "edit and correct machine translation output". As defined by Allen (2003, p.297), "In basic terms, the task of the post-editor is to edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)." In 1954, a research collaboration between Georgetown University in Washington, DC, and IBM resulted in the first widely-known MT system from Russian into English. Since then, MT has evolved with rule-based MT (Attnäs et al., 2005; Dugast et al., 2007) popular in the 1970s; then, data-driven MT including statistical MT (Brown et al., 1988; Koehn, 2010) and example-based MT (Nagao, 1984) emerging with the accessibility of large parallel corpora in the 1980s; and neural MT emerging in 2016 (Wu et al., 2016). Neural MT has made significant advances in recent years (Bentivogli et al., 2016; Junczys-Dowmunt et al., 2016; Johnson et al., 2017; Hassan et al., 2018; Toral and Way, 2018), nevertheless, the ongoing question as to the 'real' increase in quality remains a bone of contention (Koehn and Knowles, 2017; Khalilov, 2018; Zhou, 2018; Zheng et al., 2019).

MT has been controversial since its inception, especially due to its quality issues (see Chapter 3 for more information) and ethical issues (de Andrade Stupiello, 2007; Ambati et al., 2010; Drugan and Babych, 2010). When users want more than a general idea of the source text, post-editing of MT becomes important. In the early days of MT, translators were mostly hostile to MT because of its low quality (Heyn, 1996), and assumptions that MT could replace human translators (Melby, 2002) caused panic

[29] http://www.iso.org/iso/catalogue_detail.htm?csnumber=59149 (Accessed: 4 December 2019)

among translators. On the latter point, Georgakopoulou's (2012) research also presents the fear of subtitlers regarding MT, and responds to it with a positive answer. She argues that MT will expand the translation industry rather than eliminate it, since human translators are needed for quality assurance. As time goes by, translators' attitudes towards MT are changing. Cadwell et al. (2016) carried out research with 70 translators at the European Commission's Directorate-General for Translation (DGT), aiming to find out why translators decide to use or not use MT for work. Their findings suggest that a large number of translators in their sample held a positive attitude to MT, but MT was not used consistently for all tasks. They argue that translators' decisions for or against using MT were influenced by ergonomic factors related to their needs, abilities, limitations and overall well-being, and these factors were also influenced by the special institutional circumstances of the DGT.

Post-editing has been studied from a variety of angles, including how it contributes to translator productivity (Guerberof, 2009; O'Brien, 2011; Zhechev, 2012; Aranberri et al., 2014; Koponen, 2016), PE interfaces (Moorkens and O'Brien, 2013; Moorkens et al., 2015; Torres-Hostench et al., 2017), PE tools (Vieira and Specia, 2011; Aziz et al., 2012; Roturier et al., 2013), PE processes (O'Brien, 2006; Koponen, 2012; Lacruz et al., 2014), etc. However, the aspects of post-editing that will be most relevant to this study are PE guidelines and the usability of PEMT (post-edited machine translation).

In the past, it was noted that there were no widely accepted general or standard post-editing guidelines (TAUS, 2010). Different organizations made their own rules according to their needs. Since needs vary, it seems that guidelines will never be general or standard. It was also suggested that post-editing might be "light" or "full", depending

on the organisations' needs. Full post-editing usually aims to reach human translation quality and light post-editing usually means the text just has to be "good enough" or "understandable" (for a detailed discussion, see Hu and Cadwell, 2016). TAUS (2010) established MT post-editing guidelines in partnership with the Centre for Next Generation Localisation [30] in 2010 with the hope that organizations could use the guidelines as a baseline and tailor them as they required for their own purposes. TAUS highlighted two main criteria which determine the effectiveness of post-editing: the quality of the MT raw output and the expected end quality of the content. Instead of using the terms "light post-editing" and "full post-editing", TAUS (ibid., Online) put forward the "Guidelines for achieving 'good enough' quality" and the "Guidelines for achieving quality similar or equal to human translation". Based on the LISA QA model, DePalma (2013) developed post-editing guidelines from a business point of view, which highlighted the difference between light PE and full PE. He suggested that clients should agree with language service providers (LSPs) on exactly what light and full post-editing included before contracting for a job. However, there are also voices that say the distinction between light and full PE is not very useful due to the lack of clarity in the different levels of quality and the lack of standards in how PE should be done (do Carmo, 2017). In particular, light PE does not seem to add enough value to make it worthwhile for commercial use, as its role can be almost achieved by free online MT services (ibid.). According to the PE guidelines Cui (2014) proposed, LSPs should lay down a post-editing style guide based on the demand for translation quality, identify the characteristics of MT raw output based on the type of MT system (rule-based, corpus-based or hybrid),

[30] Now it is called the ADAPT Centre (https://www.adaptcentre.ie/ Accessed: 5 December 2019)

establish the PE working environment (integrated translation environment or customized post-editing environment) based on the purpose of post-editing, combine the trained MT system with professional translators, and set PE workload based on the aim of quality and editors' skills.

Similar to what Cui (2014) suggests, Hu and Cadwell (2016) also call for collaboration between LSPs and their clients to make tailored guidelines as the first step in their workflow. They compare the post-editing guidelines proposed by O'Brien (2010), Mesa-Lao (2012), Densmer (2014), Flanagan and Christensen (2014) and TAUS (2016a). As they conclude (p.351):

> "The existing PE guidelines have many overlaps, especially for light post-editing. The main differences lie in the full PE guidelines and concern the requirement for style and the expected quality of the target text, which we believe depends on the use and type of the text."

Some scholars have investigated the usability of raw MT output and post-edited MT output and found that the latter had higher usability than the former. According to the ISO 9241-11:2018 definition,[31] usability refers to "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use". Adopting this definition for usability and relevant methods (eye tracking, screen recording and questionnaire), Castilho et al. (2014) made a comparison between raw machine translated text and the post-edited version. Their research indicated that after post-editing, the usability of machine

[31] http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/35/63500.html (Accessed: 4 December 2019)

translated text increased significantly, as would be expected. Still using eye tracking and questionnaires, Castilho and O'Brien (2016) observed users' behaviour when conducting tasks with the source text, raw MT output and light post-edited MT content (English into German). Results show that in regard to cognitive effort, the light PE content was significantly different from the source text, but not from the raw MT output. Meanwhile, the group of participants who used the light PE text was found to be faster, more efficient and more satisfied than those who used raw MT. Later, Castilho (2016) added simplified Chinese and Japanese to the research and found that the raw MT versions were usable, but achieved lower acceptability (the level of acceptance to the end-user, which is influenced by usability, satisfaction, text quality, and other factors) than the light PE versions.

2.3 MT of subtitles

A number of projects have been launched to research subtitling via MT over the past ten years or so. This section will focus in on those projects. While quality will be frequently mentioned, a full discussion of the evaluation of MT will be deferred until Chapter 3.

2.3.1 MT of subtitles for films and shows

Different types of MT systems have been developed and tested for translating subtitles. Flanagan (2009) carried out research (as part of the MovRat project), which aimed to

apply MT to AVT and test the feasibility of seeding an example-based machine translation (EBMT) system with human-generated subtitles to machine translate new movie subtitles from English to German and English to Japanese. This study was based on previous research conducted by Armstrong et al. (2006), who focused on improving the quality of automated DVD subtitles by using a modular EBMT system, MaTrEx. Flanagan found that increasing corpus size and the number of source language repetitions while decreasing corpus homogeneity improved the readability of the EBMT-generated subtitles. However, these interventions did not have much impact on comprehensibility, style and well-formedness of the EBMT-generated subtitles. In addition, results showed that the viewing subjects' judgements of the quality of EBMT-generated subtitles were not strongly affected by their linguistic background, while prior knowledge of the movie had an effect on how they rated the severity of observed errors in the subtitles and the style of subtitles, but this effect is training corpus-dependent.

Castilho et al. (2011) tested rule-based machine translation (RBMT) and statistical machine translation (SMT) systems for translating subtitles. They assessed post-editing for automatic and semi-automatic translation (Brazilian Portuguese and English) of DVD subtitles. She downloaded subtitles of the American TV series *X files* from three free subtitle websites, cleaned the corpus, used one RBMT system, two SMT systems and one TM (translation memory) system, and employed eleven participants to post-edit the automatic translations and to translate English subtitles from scratch. They found that post-editing was on average 70% faster than translating from scratch. More than 69% of the translation required little or no post-editing with their best system. The TM system also performed well compared to translating from scratch.

The EU-funded project SUMAT,[32] which aimed at developing an online SMT service for subtitles for nine European languages combined into 14 language pairs, also achieved promising results. The MT system was trained using corpora composed of combined professionally-created and crowd-sourced data (Fishel et al., 2012; Etchegoyhen et al., 2014). Etchegoyhen et al. (ibid.) carried out productivity measurement through SUMAT by comparing the time needed to translate a subtitled file from scratch with post-editing machine translated output. Results show that the productivity of post-editing MT output was nearly 40% higher than translating from scratch in terms of subtitle per minute. Meanwhile, their quality evaluation also demonstrated positive results about post-editing MT (see Section 3.3.2.1 for more information). Both Castilho's study and the SUMAT project have confirmed that MT can be really helpful for translating subtitles, and they are faster to post-edit than to translate from scratch.

In addition to academic research, some scholars have explored the commercial possibilities of machine translated subtitles. Volk (2008) reports on a project that aimed to build an SMT system for translating film subtitles from Swedish to Danish and Norwegian in a commercial setting. The project had access to more than 14,000 subtitle files in each language and used 4 million subtitles for training the system. The MT systems built by Volk and his colleagues have been used in large scale subtitle production by a subtitle company who is satisfied with their work. Volk et al. (2010) report their experiences in working with the consumers of the subtitles and summarized lessons for SMT in subtitle production: a) TM-MT integration is not self-evidently beneficial due to the modest contribution of TM to the translation quality; b) there are

[32] http://www.fp7-sumat-project.eu (Accessed: 5 December 2019)

conflicting interests among different stakeholders, for whom quality or time might be more important; c) integrating SMT into the subtitling workflow is the key; d) ambiguous MT output with strange constructions may influence post-editors' judgements; e) a complex SMT system requires maintenance and updates, but companies are reluctant to invest human resources in it; f) the idea of presenting alternative translations on the screen is rejected by post-editors due to the time cost; g) increasing the efficiency of post-editors and improving the MT system to filter out bad translations is important.

Georgakopoulou (2012) discusses Automatic Speech Recognition (ASR) technology in interlingual subtitling. She argues that there is an increasing demand for interlingual subtitling, and that intralingual subtitles are not only used by hard-of-hearing viewers, but also by hearing viewers as a means to retrieve information. She presents the problems that would prevent subtitling by MT from being implemented in business practice: firstly, a lack of corpora leads to unsatisfactory results, which makes post-editing a must; secondly, more technological progress in speech recognition such as speech-to-text conversion is needed; thirdly, responding to points made by Volk et al. (2010), Georgakopoulou emphasizes that it is difficult to integrate the technology into existing workflows, and for the success of MT systems, it is necessary to establish collaboration and build trust among subtitling companies, language experts and subtitlers. Georgakopoulou points out that the ideal scenario for the subtitling industry will be to create a system that can do automatic transcription of the audio streams and then machine translate the text. She also called for the training of linguists and interdisciplinarity in studies at university level to help people understand and learn subtitling by MT. The three-year project "ALST-Linguistic and Sensorial Accessibility:

Technologies for Voice-over And Audio Description" led by Matamala (2015) is also about the application of ASR and MT in video, although it is not directly related to subtitles, the results of the project also provide insights on integrating MT and ASR into subtitling.

Although there are research projects on subtitling by MT, no established subtitling group has publicly acknowledged their use of MT. On the contrary, the acceptance of MT for audio-visual products is still quite low in society. For example, OpenSubtitles.org is (see the end of Section 2.2.5.1) one of the best and largest subtitle sites. In its forum, there is a highlighted notice[33] (last edited in May 2014) saying that "Opensubtitles has decided to get rid of all 'MACHINE TRANSLATED' subtitles", since this is "best for the future of OS as we can avoid mass trash being accumulated and saved here over the coming years". The administrators of the forum believe that:

> "Viewing a movie with a machine (google) translation, you are truly missing a lot of the real dialogue with these subtitles. Even though you can make basic sense out of them, the wording does not portray the true concept, belief and emotions depicted in a particular scene of a movie/TV series".

To change this situation, apart from improving the translation quality, trustworthiness should be established among subtitlers, end users, and project managers when MT is deployed.

[33] https://forum.opensubtitles.org/viewtopic.php?f=1andt=1969 (Accessed: 15 March 2019)

2.3.2 MT of subtitles for MOOCs

There are far fewer studies on subtitling MOOCs by MT than on subtitling films and TV shows by MT, perhaps because MOOCs have just become popular in the past ten years. Although the number of relevant projects is small, the technology deployed for MT in MOOCs is advanced. For example, as with Georgakopoulou's study (2012), Orlič et al. (2014) investigated ASR technology in their project named transLectures (Transcription and Translation of Video Lectures). This project was designed to develop advanced automatic transcription and MT (between Slovene and English) for two websites containing video lectures. They particularly focused on IT tools that can be deployed to help improve the quality and the speed of automatic transcription and translation. Engaged in the same project as Orlič et al., Miró et al. (2015) evaluated the efficiency of the manual review process of automatic subtitles and compared it with the conventional generation of video subtitles from scratch through transLectures. They recruited lecturers and volunteers as reviewers and explored the evaluation from two perspectives: the quality of automatic transcription or translation; and the time the lecturers spent reviewing the automatic subtitles. There were four tasks: 1) review of Spanish transcriptions; 2) review of English transcriptions; 3) review of Catalan transcriptions; and 4) review of Spanish to English translations. Non-native speakers of English (volunteers) took task 2), and lecturers took the other three tasks. The methods adopted for user evaluations were WER (word error rate), TER (translation edit rate) and RTF (real time factor). TER is determined by the minimum number of edits required to transform a MT output into its corrected version. RTF measured the speed of monitoring

the transcriptions of a video (lecturers were asked to monitor the automatic transcriptions for their videos, which means they play the videos with automatic transcriptions, when they spot an error, they can correct it and the video is automatically paused), while WER measured the accuracy of the automatic transcriptions (see the formula below for further explanation).

> RTF = time devoted to reviewing the transcription or translation of a video / duration of the video
>
> WER = number of basic word editing operations required to convert the automatic transcription into the correct reviewed transcription / total number of words in the reviewed transcription
>
> TER = number of edit operations required to convert the automatic translation into the correct reviewed translation / total number of words in the reviewed translation

Miró et al. found that the time required for reviewing automatic subtitles (transcriptions and translations) in all three languages is significantly less than the time required for manual transcriptions and translations. Regarding the quality between the automatic subtitles and those translated from scratch, no difference was found.

Another project that used ASR and MT for lectures is a 30-month (Feb 2014 – Jul 2016) pilot action named EMMA (European Multiple MOOC Aggregator),[34] which provides a system for delivering free MOOCs from various European universities. Its website supports nine languages (Catalan, Dutch, English, Estonian, French, Italian, Portuguese,

[34] https://platform.europeanmoocs.eu (Accessed: 5 December 2019)

Spanish, and Polish). The project provides automatic transcription in seven languages (Dutch, English, Estonian, French, Italian, Portuguese, and Spanish) and MT into English, Italian and Spanish. The transcriptions and translations are reviewed by lecturers so as to reach publishable quality. Most MOOCs in the project are bilingual (original language plus English) or trilingual (Italian, French or Spanish as a third language). Miró et al. (2018) claim that both the transLectures and EMMA projects have shown that domain-adapted ASR systems have reached a mature level, allowing us to generate low-cost automatic subtitles of (nearly) publishable quality in most cases. Their quality and efficiency evaluations for the EMMA project and the UPV media repository (a large video lecture repository developed by Universitat Politècnica de València) have confirmed their statement. As in the transLectures project mentioned at the beginning of this section (Miró et al., 2015), they measured the transcription quality using WER, the translation quality using TER, and the post-editing time for automatic transcriptions and translations using RTF. Results show that compared to generating subtitles from scratch, the ASR/MT systems can result in time savings of 25% - 75%, and the accuracy of the machine translated subtitles is such that they are worth post-editing. The multilingual subtitles of MOOCs have a positive impact on attracting students and have increased the student enrolment rate on EMMA by 70%.

The latest study on machine translated subtitles of MOOCs is the TraMOOC project[35] carried out jointly by academic researchers, industrial organizations and user partners. The project has built a system named Translexy[36] to automatically translate English-language MOOCs into 11 languages (Bulgarian, Chinese, Croatian, Czech, Dutch, German,

[35] https://tramooc.eu (Accessed: 5 December 2019)
[36] http://www.translexy.com (Accessed: 5 December 2019)

Greek, Italian, Polish, Portuguese, Russian). Kordoni et al. (2016) point out that crowdsourcing was used in the project for both quality evaluation and creating parallel corpora. The evaluation of output quality in TraMOOC relies on a multimodal schema that involves error type markup, an error taxonomy for translation model comparison, explicit evaluation (automatic and human evaluation), implicit evaluation (by using topic identification which focuses on topical information elements like named entities, events and specific terms in source and target texts, and sentiment analysis that extracts users' opinions about the MT text from forums) etc. Kordoni et al. indicate that by using this model, more in-domain parallel data with better quality can be fed back to the translation engine for the purpose of higher quality output. In order to create parallel corpora for the educational domain, Behnke et al. (2018) report that they applied crowdsourcing with strict quality controls. They found that their crowdsourced data outperformed both the general-domain baseline systems and the systems fine-tuned with pre-existing in-domain corpora all the way.

A unique feature of TraMOOC is that in addition to automatically translating subtitles, TraMOOC also translates other types of MOOC text, such as forums, assignments, reading materials, etc. The TraMOOC platform supports a wide range of file formats including SRT, WebVTT, SCC, Microsoft PPT, XML, to name just a few. The platform has already been successfully integrated into and tested in the openHPI MOOC platform[37] and the VideoLectures.NET library.[38] The first MOOC that was automatically translated by the platform from English into the abovementioned 11 languages is entitled "In-

[37] https://open.hpi.de (Accessed: 5 December 2019)
[38] http://videolectures.net (Accessed: 11 March 2019)

Memory Data Management 2017" and offered by Prof. Hasso Plattner,[39] and in which over 7,000 students had enrolled by March 2019.

2.4 The reception of subtitles

According to Kovačič (1995, p.376), where subtitling is concerned, there are several interpretations for the term 'reception' as follows:

a) The socio-cultural issue of non-TV context influencing the process of receiving subtitles as part of the complex TV signal;
b) The attitudinal issue of viewers' preferences for subtitles over dubbing or vice versa;
c) The perceptual issue of subtitle decoding (reading and viewing) strategies;
d) The psychological or cognitive issue of the impact of cognitive environment on decoding and understanding subtitles.

In this PhD research, reception is operationalized by using Gambier's (2007) reception model, which is based on Kovačič's interpretations and which will be elaborated upon in Section 2.4.1. An outline of existing AVT reception studies will be presented in Section 2.4.2. The chapter ends with a discussion of the gaps in existing reception studies and the intended contribution of the current research (in Section 2.4.3). Given the importance of eye-tracking and questionnaires to these studies, my review of the literature will start with studies based on Gambier's (2007) reception model, but will then bifurcate along these methodological lines.

[39] https://open.hpi.de/courses/imdb2017 (Accessed: 11 March 2019)

2.4.1 Gambier's reception model

As early as 2003, Yves Gambier called for reception studies in translation studies. Fifteen years later "reception of translated text" had still received "relatively scant, uneven attention in Translation Studies" (Gambier, 2018, p.43). Gambier (ibid.) also called for interdisciplinary collaboration and argued this would contribute to the improvement of AVT reception studies. When referring to AVT, Gambier (2003, p.178) also uses the term 'transadaptation', which he claims might allow us "to take target audiences into consideration more directly".

This PhD research has adopted Gambier's (2007) classification of the elements contained in his proposed reception process, which are 'response', 'reaction' and 'repercussion'. According to Gambier (2009, pp.53-54), 'response' refers to "the perceptual decoding", 'reaction' refers to "the psycho-cognitive issue", and 'repercussion' is linked with the "attitudinal issue" and "the social-cultural dimension of the non-TV context which influences the receiving process". The three terms were originally suggested by Chesterman (2007), who used the plural forms (responses, reactions, and repercussions). He pointed out that the research on translation reception had not adopted standard terminology at that time. In his research, the importance of quality assessment was emphasized. The term *reactions* was adopted in light of relevance theory, which "talks of the 'cognitive effects' of any act of communication" (Chesterman, 2007, p.179). According to Chesterman, *responses* manifest feedback and pertain to the general notion of discourse when they are communicative acts. In addition, Chesterman held that when describing the effects of translations at the

cultural level, people turn to translation *repercussions*. It should be noted that regarding *responses*, the term has been differently defined in Gambier's classification. It seems that Gambier made some changes to the term according to his own context. His classification is not only based on Chesterman's work, but also Kovačič's (1995). As mentioned at the beginning of this section, we can see Kovačič's interpretations have covered several aspects. Related to Gambier's classification, issues a) and b) in Kovačič's interpretations are linked to repercussion, issue c) is linked to response, and issue d) is linked to reaction.

2.4.2 Existing AVT reception studies

Gambier's reception model has been adopted by Orrego-Carmona (2015), who explores the audience reception of professional subtitles and non-professional subtitles (English to Spanish). He emphasizes that because audiences are impatient to wait for legitimate resources and subtitling is easy to do nowadays, non-professional subtitling has become increasingly used. He recruited 52 participants and showed them three excerpts from *The Big Bang Theory* with three different types of subtitles: one with professional subtitles, and two with non-professional subtitles produced by two different subtitling communities (one in "neutral Spanish" produced by an Argentinian community, and one in Iberian Spanish produced by an Iberian group). By using methods including eye-tracking, questionnaires, and interviews, he found that non-professional subtitles can work as well as professional subtitles, and non-professional subtitles do not necessarily have a negative impact on participants' reception. In addition, the level of English

proficiency has an effect on subtitle-reading behaviour, which means participants with low English proficiency relied more on subtitles, while some of those who speak better English almost paid no attention to subtitles.

While Orrego-Carmona's study shows that audiences receive non-professional subtitles well, Secară (2015), who also based her work on Gambier's reception model, went further. Secară challenged uniformity in the subtitling production and presentation process and carried out an empirical study on the effect of non-traditional subtitling practices on viewers. She focused especially on the use of creative spellings in subtitling, specifically the shortened forms used in texting, for example, 'C U', 'gr8', 'Thx', etc. Secară adopted eye-tracking to examine the behaviour of two groups of English native speakers in reading two subtitled versions of a French Internet clip. One version used creative *txt lingo* in interlingual English subtitles, while the other one used traditional English subtitles. She also used questionnaires to measure participants' *txt lingo* production and consumption behaviour, comprehension, access to subtitled content and attitudes regarding txt form usage. Her results show that people have a very low tolerance for non-standard practices in subtitling, although they all reported they could understand them. What's more, those with high levels of messaging media exposure had neither more relaxed acceptability constraints nor higher acceptance of non-standard forms. Since people have been exposed to standard subtitles for many years, this result may not be surprising. As the Chinese saying goes, "It's easier to move a mountain than change a man's habit."

Also by using eye-tracking and questionnaires, Iriarte (2017) adopted Gambier's reception model for studying the deaf and hard-of-hearing viewers' reception of

subtitles for two English-speaking film videos (one more focused on verbal information, the other more focused on visual information). In Spain, she recruited 72 viewers with three types of hearing and communication profile: hearing, deaf-sign language users, and deaf-spoken language users. Along with the three types of viewers' profiles, she used three versions of subtitles that differed in the speed of exposure. Results show that viewers' profile had an impact to their reception of subtitles in all stages, however, rather than the individual differences in hearing, language and communication, it was more the differences in reading skills that matters. In regard to the three R's, it was found that the subtitling speed of exposure had an impact to viewers' "response", but not much to their "reaction" and "repercussion". In addition, viewers' reception to the two videos not only varied in terms of their different profiles, but also throughout the three reception levels, which means there was a strong correlation between the content of the video and the reception of the viewers. An earlier PhD study on AVT reception of Spanish deaf and hard-of-hearing viewers was conducted by Arnáiz-Uzquiza (2012), who did not adopt Gambier's reception model, but also used eye-tracking and questionnaires (including comprehension testing) as methods for the research. Interestingly, she found that while hearing-impaired users demand verbatim subtitles, results suggest that this speed is not conducive to their comprehension at all. Besides, results also show that adapted subtitles do not have the most successful results in comprehension, even among deaf people. More research on the reception of subtitles for the deaf and hard of hearing viewers in Europe can be found in the DTV4ALL (Digital Television for All) project (Romero-Fresco, 2015).

2.4.2.1 Use of Eye-tracking

As O'Brien (2006, p.185) explains, "an eye-tracker is a device that monitors and records the movement of an individual's eyes… [it] contains special diodes that reflect light off the pupil and monitor fixations, gaze paths and pupil size as the subject interacts with an object on-screen." According to Duchowski (2017), eye-tracking has been applied in various areas, and the roles it plays can broadly be divided into two categories: diagnostic and interactive. As he put it (p. 247):

> "In its diagnostic role, the eye tracker provides objective and quantitative evidence of the user's visual and (overt) attentional processes… Equipped with an eye tracker as an input device, an interactive system is expected to respond to, or interact with the user."

Duchowski (ibid.) points out that diagnostic eye-tracking techniques can be applied to the fields of psychology, marketing/advertising, human factors and ergonomics. As he stated, in recent years, eye-tracking has been frequently adopted as a method for translation process studies and mainly plays a diagnostic role. It has been used as a method for investigating mental activities, cognitive workload, emotional stimulation, etc. (O'Brien, 2006). Among them, the use of eye-tracking in investigating cognitive workload is relevant to this PhD study. It has to be noted that using eye-tracking data to indicate cognitive processing is based on the assumption that "eye-tracking data can be interpreted as correlates of on-going cognitive processing of source and/or target texts" (Alves et al., 2012, p.6). The advantage of using eye-tracking is that "integrated eye trackers are unobtrusive, resulting in lower interference with viewer behaviour, and

they also generate a large volume of objective data" (Caffrey, 2009, p.92). Since O'Brien (2009, 2010, 2014) first introduced eye-tracking to the study of MT, more and more scholars have used eye tracking to study MT and post-editing (Carl et al., 2011; Lacruz et al., 2012; Green et al., 2013; Sanchis-Trilles et al., 2014). Regarding how to use eye-tracking technology in translation process research, Teixeira and O'Brien (2018) provide a fine-grained introduction, including how to select an eye-tracker, find suitable participants, configure the software, synchronize logs across tools, segment the recordings and whether to use data from one eye or both eyes.

In the 1980s, d'Ydewalle and colleagues (e.g.: d'Ydewalle et al., 1985; d'Ydewalle et al., 1987) started using eye-tracking to explore the various aspects of the reception of subtitles. Since then, some scholars have used it to study reception of additional information on the screen. Caffrey (2008a), for example, uses eye-tracking data to study the distribution of viewers' attention over VNC (visual nonverbal cues) and screen text. Viewers were asked to watch some anime clips on a computer with an eye-tracker, and their pupillary contractions and dilations were measured to study their emotional reactions to the clip. Caffrey defined two areas of interest on the screen: the subtitle area and the VNC area. He analysed only the scenes where VNC was present, and reported the raw data for each participant (four in total) rather than using summary statistics. The fixation data shows that in most cases, when there were subtitles on screen together with the VNC in question, the fixations on the subtitle area were more than those in the VNC area, which implies that subtitles tend to distract visual attention from the VNC area. By analysing the eye-tracking data, Caffrey (2008b) found that the use of pop-up gloss in addition to the subtitles on the screen could increase viewers'

processing effort; viewers tended to skip a number of subtitles and also spend less time gazing at the subtitles. Adopting eye-tracking technology and questionnaires, Künzli and Ehrensberger-Dow (2011) conducted an empirical study investigating reception capacity and audience response to subtitled movies (FR-DE) with 27 undergraduate students as participants. They used four excerpts: each excerpt had two subtitled versions, one with standard subtitles, and the other one with subtitles and surtitles (additional information in the upper part of the screen). Different from the work of Caffrey (2008b), they reported that in the version with surtitles, viewers spent less time gazing at the rest of the image, thus the time spent gazing at surtitles was not at the expense of the time spent on subtitles. Their research indicated that young people are able to process additional information on the screen of a subtitled video, and the additional information can help viewers understand the video better. Perego et al. (2016) acknowledge that previous research has shown no significant trade-off between subtitle processing and image processing, meanwhile, they also point out that many of the AVT reception studies are conducted mono-nationally, thus they carried out a cross-national (Italy, Spain, Poland and Belgium) study on the reception of inter-lingual film subtitles with viewers from countries with different AVT traditions. Rather than eye-tracking, they used questionnaires and cognitive tests for their research. Results show that subtitles were effective regardless of how familiar the users were with them, although their level of enjoyment varied.

Some scholars have used eye-tracking to study subtitle segmentation. Perego et al. (2010) carried out an eye-tracking experiment to study the cognitive effectiveness of subtitle processing. Participants were divided into two groups and exposed to a

Hungarian video with Italian subtitles. One group was presented with the video with well-segmented subtitles, while the other one was presented with the video with ill-segmented subtitles (subtitles that were not segmented according to the norm). The participants were tested individually in the lab. They sat in a comfortable chair (with headphones) with their eyes about 60cm from the screen and watched the 15-minute clip with subtitles. There were two areas of interest: the area below a threshold line (subtitle area) and the area above it (main film zone). Results indicated that participants had a good understanding of the film content regardless of the quality of subtitle segmentation and without significant trade-off between image and text processing. More eye-tracking research on viewers' preferences regarding subtitle segmentation can be found in Gerber-Morón et al. (2018) and Gerber-Morón (2019). The former used screenshots as materials and found that viewers prefer syntactically segmented line breaks; the latter used videos as materials and found that subtitle segmentation does not seem to be a decisive factor influencing media accessibility. Szarkowska et al. (2016) conducted an eye-tracking study on the influence of text editing and subtitle presentation rates with deaf and hard of hearing Polish participants. Their results show that text editing brings no benefit to the audience, especially when the subtitles are intra-lingual and viewers are deaf. They found that viewers understood the verbatim subtitles with higher subtitling rates better, skipped subtitles less frequently, and their reading pattern was more effective.

Other reception studies using eye-tracking as a method centre on topics such as spotting translation problems, game subtitling, cultural awareness, etc. For example, Alves et al. (2012) note that eye-tracking has been used for a variety of experimental studies in

translation process research. However, they point out that the significant variance of eye fixations among participants can make studies rather complex. To solve this issue, they recruited eight participants and used overlapped heat maps of different participants to analyse translation problems. Heat maps generated by eye-trackers can show both fixation count and fixation duration from a graphic perspective. By overlapping correlated sets of scenes of the eight participants, Alves et al. generated heat maps which represented each task according to eye fixations. In the heat maps, instances where fixations were longer "were considered to be potential candidates for translation problems that were cognitively relevant for all subjects in terms of processing effort" (2012, p.11). Calling for the standardisation of subtitling practices in game localization and the improvement of accessibility for all players, Mangiron (2016) conducted a small-scale empirical study on the reception of subtitles for games by using user tests with eye-tracking technology and a questionnaire. Regarding the eye-tracking method, she focused only on the variable of fixation duration, which is the most-used measure in eye-tracking research (Holmqvist et al., 2011). Results show that hearing users and deaf users have different preferences for subtitle presentation, which allows for the personalization of the gameplay interfaces, and hearing users read faster than deaf users. As for what they had in common, all users preferred the subtitles to be presented in the centre of the screen and read two-line subtitles slightly faster than one-line subtitles. Łabendowicz (2018) studied how various modes of AVT (dubbing, voice over, fansubbing, etc.) affect Polish viewers' reception of humorous audio-visual materials from American culture by analysing eye-tracking data and viewers' feedback. She found that subtitles did not necessarily make it easier for viewers to identify a

cultural reference, rather, they could divert viewers' attention from such references and make them more difficult to notice.

Although eye-tracking has become more and more widely used in the reception studies of subtitles, it is not yet firmly established in this field and has its methodological limitations, which have been summarized by Doherty et al. (2018, p.57) as follows: 1) inconsistency of terminology and operational definitions; 2) limited sampling and participant profiling; 3) limited data analysis and statistical testing; 4) inconsistent reporting of the technical specifications of eye trackers and their related software. They point out that these limitations can lead to a lack of standardization in analysing results and to incorrect inferences. Their points have been echoed by Orero et al. (2018), who call for a common framework and best practice for experimental research in AVT. They indicate that methods such as eye-tracking, EEG, heart rate, etc. have been heavily borrowed by AVT scholars from fields including psychology, psycholinguistics, and cognitive science, where the experimental tradition has been established for a long time. However, the application of these methods in AVT research is usually not as rigorous as the fields from which they originate, with adverse consequences for the validity and reliability of the research. Hence, they emphasize the significance of triangulation, and recommend the combination of online (eye-tracking, EEG, etc.) and offline measures (self-report scales, interviews, etc.), as well as quantitative and qualitative analysis for experimental AVT research.

2.4.2.2 Use of questionnaires

Relying solely on eye movements, it is impossible to deduce whether the difficulty of comprehending subtitles will affect the success of viewers' comprehension or not, or to deduce their thoughts and feelings about what they see (Doherty et al., 2018). Doherty and Kruger (2018) maintain that recent AVT research has tended to use mixed research methods like eye-tracking, EEG (Electroencephalography), questionnaires, and surveys, to study viewers' cognitive process when viewing AVT products. It can be seen in the studies mentioned in Sections 2.4.1 and 2.4.2, that eye-tracking is usually used along with questionnaires in translation studies. Indeed, Toury (1995, p.228) points out that the most common way to empirically study addressees' response to translated texts is through questionnaires.

Italian researchers have been particularly active in using questionnaires to study viewers' perceptions and reception of subtitles, mostly to study the transmission of humour. In 2004, Chiaro carried out a study on Italian viewers' perceptions of translated verbally expressed humour in a North American TV series using an online self-reporting questionnaire. Viewers were asked to rate their understanding and also explain what they had understood of each clip on the questionnaire (Chiaro, 2004). Similarly, Antonini (2005) used questionnaires in her research project on Italian viewers' perceptions of a subtitled episode of the Irish TV series *Father Ted*. She focused on studying the visual humour elements as well as the translated verbally expressed humour. Similar to Chiaro's research design, after viewing each clip, participants of Antonini's study were asked to rate their appreciation of the clip on two 6-item scales, and then they were

asked whether they understood the humour in the clip and also explain it briefly. In 2007, she adopted a large-scale online questionnaire to measure Italian viewers' perceptions of culturally marked items in dubbing (Antonini, 2007). In the same year, Chiaro carried out another study on visual humour and verbally expressed humour (Chiaro, 2007). She recruited both British and Italian participants and used questionnaires to compare their humour responses. Her questionnaire contained two sections: Section A gathered demographic information and participants' attitudes towards comic films, in addition to a self-evaluation of their sense of humour and their mood when doing the questionnaire; Section B included two questions on each clip and a self-evaluation about participants' responses to the verbally expressed humour in the clip on a 7-point scale. Bucaria (2005) focused on comparing Italian viewers' perception of humour in dubbing (reduced humorous elements) and subtitling (complete humorous elements) of an American TV show. Her questionnaire was composed of three sections: the first section aimed to assess viewers' enjoyment of each clip, the second section aimed to ask whether viewers found the clip funny or annoying on a 6-point scale, and they were asked to explain their feelings in the third section. Bucaria and Chiaro (2007) distributed 150 questionnaires to Italians who were shown a tape containing twenty excerpts from several TV programmes dubbed into Italian. In this way, they studied the participants' understanding of the culture-specific aspects of the AV content. To discover how existing subtitle norms meet the needs of Italian viewers, Perego and Del Missier (2008) turned to questionnaires to test readability, and to measure comprehension and the recollection of images from a Hungarian film excerpt subtitled in Italian. Their questionnaire had two parts, one part on viewers' opinions about the usefulness of the subtitles and whether they paid attention to the subtitles (this part contained five

questions on a 7-point scale and two multiple-choice factual questions); the other part on viewers' general comprehension about the content of the video, which contained 12 multiple-choice questions.

As with some of the Italian scholars mentioned above, Schauffler (2012) also studied humour by using questionnaires. She investigated audience reception of different strategies for subtitling English into German of a short animated film. The only difference between the two subtitled versions lay in the fact that two different strategies were used for wordplay translation. One was based on formal equivalence but lost parts of the extensive humorous content of the source text, while the other one was less formally similar to the source text but maintained more linguistic humour and prioritised equivalence of effect. The participants were German native speakers with various levels of English proficiency. Questionnaires were given to participants to express their reactions after watching the film. Results show that the participants had significantly better reception of the version that maintained more linguistic humour than the one that lost parts of the humorous content. The findings indicate that the AVT industry should adopt more flexible and creative translation strategies.

2.4.3 Discussion

From the literature discussed in Section 2.4, we can see that although Gambier (2018) reports that reception studies of translated texts are still relatively scant, this area has drawn more and more attention from scholars in recent years. Gambier's three-R

reception model has been adopted by a small number of scholars for studying viewers' reception of translated subtitles with good results. However, it can be seen that none of the reception studies is about the subtitles of MOOCs, they are all about films and TV shows. Nowadays, MOOCs are pushing education to become more borderless. Considering the fast-growing market for MOOCs and the increasing number of worldwide viewers of MOOC videos, more research is needed to explore the use of MOOCs and users' reception of the translated subtitles for MOOC videos.

In addition, the majority of reception studies only focus on human-translated subtitles, especially on comparing professional subtitles and non-professional subtitles. Here, it has to be noted that there is no shared terminology for different types of subtitles yet. For example, Orrego-Carmona (2015) refers to "professional subtitles" and "non-professional subtitles", Lee (2014) refers to "professional" and "amateur subtitling", and Secară (2015) refers to "traditional" and "non-traditional". In this PhD research, the terms 'professional subtitles' and 'non-professional subtitles' are used. It is important to distinguish between the two, however, Orrego-Carmona (2016) points out that the concept of professional translation is complex and there is no evidence in any literature that a straight-forward boundary has been established between professionals and non-professionals. He distinguishes the two groups by monetary reward and defines them as follows (ibid., p.165):

> "[P]rofessional translators are hired and paid as translators as opposed to outsiders who do not receive payment for their translation. It is true that other rewards might be obtained in non-professional translation settings (Fernández Costales, 2013; O'Hagan, 2011), however, when a translation is not paid for, the

> conditions under which it is produced, according to my definition, make it non-professional."

The author agrees with Orrego-Carmona's definition and distinguishes the two groups by monetary reward. Professional subtitles are produced by people who are hired and paid for the work, while non-professional subtitles are produced by people who work out of interest and do not receive any payment. Apart from comparing viewers' reception of professional and non-professional subtitles, other types of subtitles are also worth researching, for example, machine translated subtitles and post-edited MT subtitles, which is the focus of this research. The primary reason for comparing the reception of raw MT output, full post-edited output and human translation is due to the recurrent theme: MT quality. The literature discussed in Section 2.2.4.2 has shown the prevalence of MT nowadays, and Section 2.3 has revealed the fact that MT quality has improved significantly over the last ten years. Claims are being made (e.g., Koehn, 2016; Sennrich et al., 2016; Wu et al., 2016) that neural machine translation (NMT) will result in higher quality raw MT output. However, post-editing is still necessary and has been widely implemented by LSPs with the aim of improving the quality of the target text (see Section 2.2.4.2). Meanwhile, previous work (e.g., Castilho et al., 2014; Doherty and O'Brien, 2014) has shown that although raw MT output is less usable and acceptable when compared with post-edited content, people can still successfully perform tasks when reading raw MT content. In addition, whether MT can achieve quality equal to that and take the place of a professional translator has also been the focus of controversy (see Section 2.2.4.2). Therefore, the researcher includes raw MT, full post-edited MT and human translation conditions in the research design. In addition, little attention has been given to Chinese viewers' reception of translated English subtitles,

even though Chinese have been exposed to subtitles for a long time and fansubbing in China is so popular. The existing studies on English to Chinese subtitles are mostly about translation strategies; thus in the area of reception of translated subtitles, more research is needed on the English-Chinese language pair. For all the reasons above, the researcher chooses to study Chinese viewers' reception of raw and post-edited machine translated subtitles and human translated subtitles for English-speaking MOOCs.

As discussed, eye-tracking technology has been adopted in translation studies for over ten years now. It has become a well-established method for measuring human cognitive activities and has been widely adopted in reception studies of subtitles. Translation scholars usually use eye-tracking, questionnaires and some other methods together as a mixed-methods approach for reception studies of subtitles, and their results have shown that this approach is effective and robust. The current research, therefore, uses this mixed-methods approach to study machine translated subtitles, in an attempt to gain new insight into both reception studies and AVT.

2.5 Conclusions

This chapter presents an overview of recent literature on MOOCs, subtitling, MT, and reception studies. We can see that a number of research projects have been carried out on subtitling audio-visual products by MT, but mostly focus on films and TV shows. Very few are on MOOCs. As a special genre of audio-visual media, MOOCs are based online, targeted to various people in various countries. MOOCs are developing quickly all over

the world, including China. Some reports have shown that language is a barrier to many users when learning via MOOCs. As the number of MOOC providers is increasing, there is a large number of MOOCs to be translated, bringing about a considerable market for subtitling MOOCs. However, the translation of MOOCs is mostly done by humans, either by crowdsourcing online or by fansub groups. Given the large quantity of MOOCs and the considerable demand for them, MT appears to be a logical solution, and this provides the motivation for this research. Besides, with the exception of TraMOOC, little research has been done on subtitling English-speaking MOOCs into Chinese, while this language is the most spoken language in the world. Section 2.2.3 shows that the development of MOOCs in China is in full swing, and the demand for translated subtitles is considerable. With the expanding market for MOOCs in China, the English-Chinese language pair deserves more attention from academia. Therefore, I would like to extend this PhD research to Chinese as a target language and MOOC content.

Subtitling MOOCs by MT is a new trend and represents a potential means of distributing more MOOC content internationally. The literature reviewed in this chapter shows that in most cases, post-editing MT output is faster than translating from scratch, and post-editing requires less effort than human translation. To investigate whether subtitling of MOOCs by MT (and then post-edited) can be adopted on a large scale, more work needs to be done on researching the establishment of trustworthiness, discovering the factors affecting viewers' reception, and most importantly, testing the quality of machine translated subtitles. The next chapter will focus on the translation quality issue, especially on the human and automatic evaluation of MT.

Chapter 3 Translation Quality Assessment (TQA)

3.1 Introduction to TQA

Translation quality is always a topic that can induce a heated discussion. In the research community, there is no commonly agreed definition for translation quality (Castilho, 2016). After reviewing various terms similar to translation quality assessment (e.g.: translation control, translation evaluation, translation quality assurance, etc.), Thelen (2008) concludes that the term itself is not important in the process of assessing translation quality. According to House (2014, p.2), "translation quality assessment means both retrospectively assessing the worth of a translation and prospectively ensuring the quality in the production of a translation."

According to House's (1997) categorization, there are a number of approaches to TQA: 1) a subjective approach, proponents of which believe that translation quality depends to a large extent on the translator's subjective interpretation and transfer decisions; 2) a behavioural approach, which is associated with Nida's (1964) dynamic equivalence theory and tends to be communication-oriented and takes readers' reactions to translation as the main criterion; 3) a linguistically-oriented approach, which considers the source text as the key element in the translation process and focuses on the systematic relationship between the source language and the target language; 4) a functionalist approach, as opposed to the linguistically-oriented approach, which is

based on Vermeer's (1978) skopos theory and considers the target text as the key element in the quality assessment, in other words; it indicates that translation quality depends on whether the intended function of the translation in the target language is fully realized; 5) a descriptive approach, proponents of which believe that the target text is considered an integral part of the target culture, not just the reproduction of another text, and the importance of the source text is thus greatly reduced; 6) a deconstructionist approach, proponents of which aim "to make politically pertinent (and 'correct') statements about the relationship between features of the original text and the translation text." (House, 2001, p.246); 7) an interpretative approach, proponents of which emphasize the communicative needs of the target audience and believe that translation quality depends on the clarity and intelligibility of the translation, as well as the acceptability of the translation in the target culture and the communicative function of verbal or written discourse.

House herself employed the term TQA in two monographs: *A Model for Translation Quality Assessment* (1977) and *Translation Quality Assessment: A Model Revisited* (1997). Based on Halliday's systemic functional grammar, House develops a general TQA model that focuses on comparing the source and target texts. In 2014, she proposes a new integrative model, in which she considers the relevance of cognition in TQA, and also stresses the link between comparative pragmatics, globalization and her notion of cultural filtering. Another translation scholar that needs to be highlighted here is Larose (1998), who argues that the most important aspect of measuring translation quality is the purpose of the translation. He emphasizes the context of a translation in his approach and innovatively attempts to include human factors, for example, the working

conditions of translators, into the assessment. However, House (2014) argues that Larose's model is not a worked-out model, because it is extremely difficult to include the complexity of the real environment in any assessment, especially the actual working conditions of translators. As she puts it, "the aim of uniting product and process evaluation is just too difficult (or impossible?) to reach" (p.18).

With the increasing use of technologies such as eye-tracking, keystroke logging, and EEG in translation studies however, it is not as difficult as before to study the human factors of translation, or to look at both product and process in TQA. The researcher agrees with Larose on the significance of context in TQA. As a matter of fact, TQA is carried out by humans, so human factors necessarily play a role. As Zehnalová (2013, p.43) puts it, "The aim of TQA is not to produce absolute judgements or perfectly objective assessments. Such goals are not attainable." Rather, it needs to be targeted "for a specific situation, a specific purpose and a specific audience" (ibid.). MOOCs are educational, so they need to be accurate. An error could mislead thousands of students. Hence, MOOCs require a high level of translation quality. However, similar to TED Talks, Wikipedia and Facebook, the translation of MOOCs is usually carried out by amateur subtitling groups or through crowdsourcing, which means the translator could be anyone, for example, a government official, a college student or a professional translator. To the best of my knowledge, there is no evidence of any organized professional translation of MOOCs so far. From the literature described in Chapter 2, we can see that post-edited MT can be faster than human translation, though this is not always the case. If the quality produced is similar or does not differ much from human translation, this could be a good option for subtitling MOOCs. Therefore, translation quality assessment forms an important

topic for this PhD research. The rest of this chapter is organized as follows: Section 3.2 briefly discusses the quality assessment of human-translated subtitles; Section 3.3 revolves around an important part of this PhD research – quality assessment of MT, and is divided into three sub-sections: human evaluation, automatic evaluation, and generic frameworks used in industry; Section 3.4 provides the conclusion of this chapter.

3.2 Quality assessment of human-translated subtitles

As mentioned in Section 2.2.4, subtitling MOOCs is mostly carried out by humans, including fansubbers, and crowdsourced translators using online platforms. Due to the variety of subtitlers, how to assure the quality of subtitles is an important issue.

In regard to the TQA of professional subtitles, Jan Ivarsson and Mary Carroll (1998), two well-known figures in the subtitling business, drew up the *Code of Good Subtitling Practice* that has become a widely-used industry standard. Based on this code and other related literature, Kuo (2014) discusses the theoretical factors that determine subtitling quality from three dimensions: the temporal dimension, the spatial dimension and the stylistic dimension. For the temporal dimension, she elaborates on the minimum and maximum duration time of subtitles, and points out most subtitling service providers follow the "six-second rule", which turns out to be enough time for an average viewer to read two full lines of subtitles (about 35 to 37 characters per line). She discusses some literature concerning Chinese subtitling and argues that it seems reasonable to apply the duration rules of English subtitles to Chinese subtitles. She also mentions that the golden rule for spotting is that subtitles should keep pace with utterances as much as

possible, and then discusses the pause between consecutive subtitles, overlapping dialogue, the use of blank spaces between dash and text, etc. For the spatial dimension, she stresses the importance of the safe area and position on the screen,[40] the number of lines, line length, font type, font colour and background. She argues that currently, the practice for the positioning of subtitles is almost the same in most countries, including Chinese-speaking countries. For the stylistic dimension, she took the actors' tones, language register and offensive language in subtitling as examples, and concludes that good stylistic quality means that the subtitler can combine the content and the form, and convey the spirit of the original as much as possible but not beyond the client's stylistic constraints.

To investigate the practical elements that can affect subtitling quality, Kuo conducted two rounds of online surveys with professional subtitlers, first with professionals in the subtitling industry mainly from European countries, second with professionals working with the Chinese language. She summarizes that remuneration, assignment deadlines, the quality of support materials (scripts, genesis files, glossaries etc.), and subtitling software programs, can all have an impact on the quality of the subtitles. She points out that both preventive measures (adequate guidelines and support materials etc.) and remedial measures (lexical and syntactic checks, verification of maximum reading speed and number of characters per line, proofreading, feedback from clients etc.) can help control the quality of subtitles.

[40] The 'safe area' refers to the fact that "the lowest line of the subtitles should appear at least 1/12 of the total screen height from the bottom, with a margin of at least 1/12 of the total screen width on both left and right sides for the sake of eye movement and to make it easier for the audience to read the subtitles" (Kuo, 2014, p.78).

Fansubbers enjoy more freedom than professional subtitlers, which means they are not constrained by the conventions or norms set in mainstream subtitling (Cai, 2015). However, Cai points out that too much freedom can cause high turnover in the fansubbing group, since fansubbers can leave the group at any time without notifying anybody. She also states that due to the intimate relationship between fansubbers and fans, and the voluntariness of fansubbers, fans are usually tolerant of low levels of quality. This point is questionable, since as time goes by, viewers have higher and higher expectations of quality. Especially after the emergence of 'Bullet Screen', or 'DanMu', a feature on online video sites in China and Japan that "allows the viewer to send real-time comments, called bullets, that publicly fly across the screen when watching a video" (Niu et al., 2019, p.1), it has become much easier and convenient for subtitlers to know whether their work meets the demands of viewers, who usually send either complimentary or snarky messages about the quality of subtitles on the screen. For example, Figure 3.1 is a screenshot of a Japanese anime where there are some complimentary messages on the screen, such as "字幕组干得好" (the line in the yellow box), "字幕君又调皮了" (the line in the purple box), which means "Well done fansubbing" and "naughty fansubbers" ("naughty" is used in the literal sense for "调皮", which has become Internet slang in China nowadays to mean someone is funny, or humorous in a good way. Here it can also mean the subtitles are too free, although the viewer likes them).

Figure 3.1 A screenshot of bullet screen[41]

Lv and Li (2015) point out that many fansubbers tend to value the entertainment effect more than the accuracy of the translation. They call for new standards for fansubbing. The entertainment effect of subtitles is also mentioned by Wen (2017), who points out that fansub groups have the problem of re-creation. For example, some fansubbers overused the strategy of domestication and translated “eBay” into “Taobao” (the largest online shopping website in China), and “Skype” into “QQ” (an instant messaging software in China). Wen argues that some fansubbers also like to comment on the content of the video in their subtitles, for example, in the American TV show *Desperate Housewives*, a main character passed away in the show and there was a line added by the fansubber in the subtitles saying “May you continue to be calm and sexy in heaven.

[41] Source of the screenshot:
https://www.bilibili.com/video/av26560497?from=searchandseid=4004799615723733034
(Accessed: 2 May 2019)

Thank you for the smiles and tears you've brought us over the past five years".[42] As for this kind of behaviour of the fansubbers, some viewers found it to be "very humane and heart-warming".[43] It can be seen that it is difficult to cater for all tastes. How to balance the translation quality issue and the entertainment effect of fansubbing remains a question. Meanwhile, every fansub group seems to have their own translation guidelines, but questions such as "Who made those guidelines?" "Are they qualified?" "Can the guidelines really assure the translation quality?" are worth pondering.

As for crowdsourcing, take OOPS (see Section 2.2.4.1) as an example. Due to the open-door policy in accepting volunteer translators, translation quality is always a big concern to OOPS. Whether OOPS should rely on an outside authority or get better reviewers for quality assessment remains a hot topic in the OOPS forum, and the conversation on improving quality continues at OOPS. Unfortunately, no one can come up with a final answer to the quality issue (Lee et al., 2007). In 2004, OOPS set some rules regarding its review process a few months after establishment. The review process requires that a translator and an editor go through each assignment before publishing it online. In OOPS, there are several measures being taken to establish trustworthiness: data triangulation, peer debriefing, and member checking. When analysing how members negotiated meanings and formed a collective identity in this online translation community, Lee et al. emphasize the importance of 'joint enterprise' in Etienne Wenger's concept of Communities of Practice (Wenger, 1998). According to them, "As participants in the various stages of the translation process define the joint enterprise, their responses to and interpretations of these processes were always being negotiated" (p.11), from

[42] Translated from Chinese to English by the researcher
[43] Same as above

determining word definitions to more macro issues like quality control. The fundamental question of the quality issue lies in whose knowledge is to be prioritized and why, and another question is how to decide this and who should be the decision maker. The leader of OOPS answered that all readers are proof-readers and everything is forever up for discussion and modification. Some people call for standardizing the terms and concepts, and the need for standards set by the authority. Lee et al. also argue that the high level of autonomy of the community could be an impediment to effectively managing quality. They remark that the OOPS community should consider the critiques of 'outsiders' as a peripheral form of participation which could influence the identity and sustainability of the community. The criticism and suggestions should be recognized and used to support the community, but not to create a sense of defensiveness.

From the discussion of the three types of human-translated subtitles above, we can see that TQA relies on both the internal negotiation and cooperation of the translation community and the supervision and suggestions of the external audience and clients. Quality control should be implemented throughout the whole subtitling procedure. Standards and guidelines for translation are essential to the quality assurance of subtitles and the sustainable development of any translation community. Meanwhile, it has to be noted that the TQA process and standards will differ depending on the type of product being subtitled. TV shows/Anime, for example, might have a freer approach to subtitling, but MOOCs may require a more standard approach, given their aim and audience. Therefore, while based on the general or widely-accepted standards for subtitling, specific problems require specific analysis. As previously mentioned, there is

no absolute standard for translation quality (Sager, 1989), because translation quality is context dependent (Secară, 2005).

3.3 Quality assessment of MT

Regarding MT quality evaluation, Mitchell (2015) reports that there are three major strands: 1) human evaluation, which revolves around the two concepts 'adequacy' and 'fluency'; 2) automatic evaluation, using measures such as BLEU, METEOR, and TER; and 3) usability evaluation, which provides insights from the perspective of end users and plays a key part in determining adequacy. Here, Mitchell distinguishes between human evaluation and usability evaluation, but in this PhD thesis, usability evaluation is seen as a type of human evaluation. Meanwhile, there are frameworks (e.g. MQM) that can potentially combine these two main approaches. Hence, this section contains three parts: human evaluation, automatic evaluation, and generic frameworks used in industry.

3.3.1 Human evaluation

When Microsoft claimed their neural MT system had achieved "human parity" for news translation from Chinese into English (Hassan et al., 2018), it immediately raised eyebrows in the academic community, notably in Toral et al. (2018) and Läubli et al. (2018), who both question Microsoft's claim and report that their research suggests the claim surpasses the fact. Interestingly, "human parity" is a concept that was invented by

the NLP (natural language processing) community rather than the translation community. According to Microsoft, the definition of "human parity" is as follows:

> "If a bilingual human judges the quality of a candidate translation produced by a human to be equivalent to one produced by a machine, then the machine has achieved human parity" (Hassan et al., ibid., p.2).

What Microsoft used for translation was the WMT 2017[44] news translation task from Chinese to English. News translation has long been an interest of the MT community because of its practical and commercial importance. Also, there is a wealth of parallel data in this domain on the Internet, and related government-funded projects and evaluation campaigns have a long history. Hassan et al. (2018), Toral et al. (2018) and Läubli et al. (2018) all used human evaluation for the quality assessment of MT output. Hassan et al. recruited bilingual annotators to evaluate the output of Microsoft's MT system by adopting a source-based direct assessment methodology. That is, for each sentence, annotators were shown both the source text and the candidate translation, and answered the question about the accuracy of the translation on a Likert scale. Looking back on House's (1997) categorization of approaches to TQA in Section 3.1, we can see that Hassan et al. adopted a subjective approach and a linguistically-oriented approach in TQA. Judging by their definition of "human parity", the key to their evaluation was the bilingual annotators, or "bilingual crowd workers", as they called them (ibid., p.15), without mentioning who the bilinguals were or how many they were. Hence, it remains a question how reliable these bilinguals were. In a word, it can be seen

[44] The Second Conference on Machine Translation (at EMNLP-2017): http://www.statmt.org/wmt17/ (Accessed: 5 December 2019)

that Microsoft's TQA is context dependent, their results are limited to the domain of news translation and depend on the bilinguals they recruited (the human factors). More in-depth discussion on the different aspects of human evaluation will be presented below.

3.3.1.1 Human evaluation based on judgements at segment level

Castilho et al. (2018) discuss human evaluation from six perspectives: 1) adequacy and fluency; 2) readability and comprehensibility; 3) acceptability; 4) ranking; 5) usability and performance, and 6) evaluators. Adequacy is sometimes also named accuracy, fidelity, correctness or precision in other literature (Mitchell, 2015). It refers to "the extent to which the translation transfers the meaning of the source-language unit into the target" (Castilho et al., 2018, p.18), while fluency refers to "the extent to which the translation follows the rules and norms of the target-language" (ibid.). Doherty and Kruger (2018, p.183) clarify that when applied to the field of AVT, adequacy "measures the correspondence between the words uttered in the auditory mode and the words appearing on screen (between auditory and written modes), rather than between text written in a source and target-language (within the written mode)". Adequacy and fluency are the two most well-established TQA measures. As Denkowski and Lavie (2010) put it, adequacy and fluency are two concepts that emerged very early in MT evaluation and have been used consistently over the years. The ALPAC report in 1966 was an early example of human evaluation of MT. The two characteristics used for evaluation were intelligibility and fidelity (Pierce et al., 1966). Adequacy and fluency in Pierce et al. are usually used together and assessed at the sentence level in the form of Likert scales

(ibid.). The assessment of adequacy requires the evaluator to have some bilingual proficiency, while that of fluency only requires the evaluator to be proficient in the target language. The abovementioned research by Hassan et al. (2018) is an example of recruiting bilingual evaluators for assessing the adequacy of MT output.

Readability is often used for evaluating the complexity of source/target text, and it "relates to the ease with which a given text can be read by one or more person(s)" (Castilho et al., 2018). Popular measures of readability are Flesch-Kincaid (Flesch, 1948; Kincaid et al., 1975) and Dale-Chall (Dale and Chall, 1948). An example can be seen in Perego and Del Missier (2008), who tested the readability of the Italian subtitles of a Hungarian film. However, the researcher is not sure whether readability belongs with human evaluation at all, because most readability scores are simply properties of texts, and there is no human judgement involved, apart from the judgement involved in creating the formulae. Comprehensibility "represents an attribute of the (source or target) text which indicates how understandable it is for a reader" (Castilho et al., 2018, p.19), thus comprehensibility changes based on the reader's background or familiarity with the text. Ranking is often used for comparing different MT outputs for the same source text, and it is usually adopted for large-scale MT evaluation campaigns because it is fast and efficient. Toral et al. (2018) adopted pairwise ranking to evaluate the output of Microsoft's MT system. They evaluated sentences in order, not randomly, and they showed all three translations of each source sentence to the evaluators at the same time as well as the previous and next source sentences.

Another example of human evaluation is described in Etchegoyhen et al. (2014), who recruited professional subtitlers for a large-scale MT evaluation through SUMAT (see Section 2.3.1). The process involved two rounds, a first round on quality evaluation, and a second on productivity measurement. The round of quality evaluation included quality rating, translation pair comparison, correlation measures, and error collection. The feedback provided by subtitlers who participated in the evaluation includes both positive and negative comments.

Ive et al. (2018) attempt to control MT quality at an early stage. They point out that in order to achieve better MT output, human-computer interaction is usually carried out after the fact. They propose a "Pre-Edition" protocol to conduct early resolution of translation difficulties, so as to reduce downstream errors. They declare that their study can help to address translators' concerns about being replaced by machines, that translators can contribute to the pre-evaluation of the text before MT and solve the translation difficulties in the early stage to prevent MT errors in the later stage.

3.3.1.2 Human evaluation based on error classification

Before delving into the major strands, this section provides an outline of error classification, which is of critical importance to both human and automatic evaluations.

Error classification is a good way of evaluating the quality of MT (Aziz et al., 2014). Usually in human evaluation of MT, bilingual evaluators will assign a quality score for a given task based on some criteria (Popović, 2018). Here, the criteria refer to error

schemes. Examples of simple error schemes can be seen in Farrús et al. (2010) and Federico et al. (2014), with five and seven categories respectively. The two schemes are quite broad (e.g.: category “Syntactic errors” and “Semantic errors” in Farrús et al., 2010) and ambiguous (e.g.: category “Too many errors” in Federico et al., 2014), which is not very helpful to identify the specific type of error, thus creating difficulties for improving the MT system. Other schemes are more hierarchical, for example, two-level hierarchical schemes can be seen in Kirchhoff et al. (2012), and Stymne and Ahrenberg (2012); three-level schemes can be seen in Vilar et al. (2006) and Costa (2015). These schemes all overlap, whether hierarchical or not. For the hierarchical ones, some researchers put the same categories in different levels, for instance, Kirchhoff et al. (ibid.) put “Missing words” in Level 1, while Stymne and Ahrenberg (ibid.) put it in Level 2. Some hierarchical schemes have many categories on the first level, for example, there are 12 categories in Level 1 in the scheme presented by Kirchhoff et al. (ibid.); some schemes only have a few, for example, there are four categories in Level 1 in the scheme presented by Blain et al. (2011). A more recent and widely-used hierarchical scheme is called MQM (Multidimensional Quality Metrics), which will be elaborated upon in Section 3.4. Lacruz et al. (2014) investigated intrinsic measures of MT quality and conducted error classification from a novel angle. They followed the framework of the American Translators Association (ATA) grading rubric, and classified MT errors into mechanical ones (those that could be routinely fixed without reference to the source text) and transfer ones (those that could be fixed only with reference to the source text).

A number of researchers use error classification to research the nature of or effort involved in post-editing (De Almeida and O’Brien, 2010; Depraetere, 2010; Li and Zhu,

2013), and the classification of post-editing operations, in turn, can benefit both manual and automatic error classifications (Popović, 2018). Popović (ibid.) lists six challenges for manual error classification: the annotator's profile, consistency, number of error classes, definition of error classes, unification and generalization of error typologies, and annotating post-editing operations. To solve these challenges and ensure the TQA is valid, apart from providing detailed guidelines to the evaluators, the inter-annotator agreement (IAA), which is also known as inter-rater reliability, needs to be determined. However in TQA, a low level of IAA is not uncommon (see Sections 3.3.1.3 and 4.2.3 for more information). It could be the result of a low number of evaluators or the various levels of their professional knowledge, which is often beyond remedy due to the difficulty in recruiting adequately qualified evaluators. Nevertheless, a low IAA might still occur even with a high number of evaluators, even if evaluators agree on most error types, they may not agree on the severity level. While certain inconsistencies cannot be completely avoided, precise guidelines and intensive training are necessary for human evaluators regardless of their background (ibid.).

3.3.1.3 Human evaluation based on usability judgements

Castilho et al. (2018) call for usability tests for translated content. They point out that both academia and industry rarely take end users into account when evaluating translation quality, thus the acceptance of translated content has rarely been measured. Van Slype (1979) states that MT acceptability can only be effectively measured by its end users. Roturier (2006) echoes his point and argues that the evaluator of quality has to be the real user of the translated text, in order to maximize the ecological validity of

the evaluation study. Suojanen et al. (2015) point out that traditional TQA has the problem of only focusing on the end product, which means any changes can lead to a substantial loss of money and time. They thus suggest the user-centred translation (UCT) approach, which emphasizes the central role of users in both the production and evaluation of translation. They assert that the use of the innovative UCT model can help the translation industry better face the competitive market and meet the needs of clients.

The usability of MT output has been investigated in a number of studies (see Section 2.2.4.2). Similarly, performance-based measures (objective or subjective) are used to assess how users use the final product or service with translated content. They provide real usage data and are often adopted in the localization industry. Usability, performance, and acceptability are user-centred concepts. Castilho et al. (2018, p.20) treat acceptability as a part of the concept of usability and define it as "the degree to which the target or output text meets the needs and expectations of its reader(s) or user(s)". A related term, more commonly used in AVT research, is 'reception'. Section 2.4 has elaborated on viewers' reception of subtitles, which is similar to usability evaluation to some extent.

An example of usability evaluation for subtitles is the transLectures project (see Section 2.3.2). According to Orlič et al. (2014), the quality of the subtitles was evaluated by five undergraduate and graduate students of translation studies in a Slovenian university through the transLectures player. The students participated in a two-phase evaluation process with the help of a Spanish university. Whenever they spotted an error of

transcription or translation, they could pause to do post-editing. In addition, the power users and authors of transLectures were free to edit the translations and transcriptions. After the two testing phases, the results were evaluated using RTF and WER (see Section 2.3.2 for their definitions). Another example is the ALST project, where Ortiz-Boix and Matamala (2015) propose a quality assessment of post-edited and human translated wildlife documentary films from English to Spanish at three levels: language experts' assessment, dubbing studio's assessment, and end users' assessment. The end users' assessment is the highlight of their research. Based on a mixture of holistic and analytic approaches (Lommel et al., 2014), the language experts carried out the experiment with three evaluation rounds. The dubbing studio received all the scripts and videos, and made a professional recording according to standard procedures. A researcher made observations and collected data on changes made by the dubbing director during the recording process. Following Gambier's reception model (2009), the last level of the research was carried out with end users from three perspectives: understanding, enjoyment, and preferences. Data was collected through both pre-task and post-task questionnaires. The pre-task questionnaire was designed to collect end users' demographic information, while the post-task questionnaire was to collect their comprehension, enjoyment, and preferences. Results show that human-translated texts performed better than post-edited texts, but the difference was not significant and could be considered not meaningful.

While it is important to do usability evaluation, Castilho et al. (2018) acknowledge that evaluators for research-based MT quality assessment are mostly students and amateurs who have never received any professional training due to limited resources. In fact, "it

appears from the data available that professional or trained evaluators are the exception in MTE tasks, rather than the rule" (p.23). They report that TQA may be conducted individually, in groups, or in crowds (specifically referring to large-scale crowdsourced evaluation campaigns). The more people involved, the more time and money will be required, but the evaluation will have higher validity. Meanwhile, they also admit that it is not really practical to conduct large-scale acceptability tests with the real end users of MT. In fact, crowdsourcing evaluation is often the chosen option. Graham et al. (2017) agree that crowdsourcing evaluation is more common with MT now, especially for large-scale projects. Mitchell (2015) also agrees that crowd evaluation is gaining popularity, usually with the aid of Amazon Mechanical Turk.[45] However, it can be considerably subjective and inconsistent when evaluators have no professional training. In this case, the results of their TQA usually have low intra- and inter-annotator agreement (see the discussion at the end of Section 3.3.1.2). Meanwhile, even evaluators who have received professional training can have a low inter-annotator agreement. For example, Jia et al. (2019b) recruited 30 first-year postgraduate students specialized in translation for comparing post-editing NMT output and human translation (English to Chinese). The kappa scores for fluency and accuracy were 0.0334 and 0.0744 respectively, which indicates only slight agreement among the 30 evaluators based on Landis and Koch's (1977) standard. Why there was such low agreement is perhaps because postgraduate translation students are still novices rather than experienced professionals.

[45] https://www.mturk.com (Accessed: 5 December 2019)

It is a fact that human evaluation can result in a low inter-annotator agreement, although it does not mean their TQA is useless, rather that there will be some impact on the study's validity. To better assess the quality of MT output, triangulation is necessary. Saldanha and O'Brien (2013, p.23) endorse the practice of triangulation in translation studies and acknowledge that it can help cross-check the results from different sets of data. Therefore, a combination of human evaluation and automatic evaluation can be a good choice. A discussion of automatic evaluation is thus provided in the next section.

3.3.2 Automatic evaluation

In 2002, IBM introduced the BLEU (Bilingual Evaluation Understudy) metric in the 40th annual meeting of ACL and led MT evaluation to a new era. From then on, statistical MT systems moved towards a rapid evaluation of their performance using automatic metrics. BLEU is a precision-based algorithm for evaluating the quality of machine translated text. Precision "denotes the proportion of Predicted Positive cases that are correctly Real Positives" (Agrawal and Gupta, 2018, p.139). In MT, as Papineni et al. (2002, p.312) put it, "To compute precision, one simply counts up the number of candidate translation words (unigrams) which occur in any reference translation and then divides by the total number of words in the candidate translation." NIST (Doddington, 2012) was developed at around the same time as BLEU, but is not as commonly used. At a later point, Banerjee and Lavie (2005) introduced METEOR, a related metric which measures both precision and recall. Recall "is the proportion of Real Positive cases that are correctly predicted positive" (Agrawal and Gupta, 2018,

p.139). While BLEU, NIST, and METEOR are string-based comparison metrics (comparing MT output with one or more reference translations), other metrics exist that are based on edit distance, such as TER and HTER. Before elaborating on TER and HTER, there is one construct that needs to be explained: edit distance. As stated by Przybocki et al. (2006, p.2038):

> '[E]dit distance' is defined to be the number of modifications a human editor is required to make to a system translation such that the resulting edited translation contains the complete meaning in easily understandable English, as a single high-quality human reference translation… Edit distance, as a Machine Translation (MT) metric, is an intuitive measure of the rate of errors in MT output, (number of errors, divided by number of reference words), with each edit viewed as fixing an error.

As a commonly used and well-established metric to evaluate MT quality, HTER (Human-targeted Translation Edit Rate) measures the edit distance between the machine translated text and its minimally post-edited version produced by human (Specia and Farzindar, 2010). Compared to fluency and adequacy, HTER is more complex and semi-automatic (because it is based on post-editing, see Denkowski and Lavie, 2010). Snover et al. (2009) argue that by using HTER, humans do not score translations directly, but rather generate, via post-editing, a new reference translation that is closer to the MT output but meanwhile retains the fluency and meaning of the original reference. This new targeted reference is then used as the reference translation when scoring the MT output using Translation Edit/Error Rate (TER) or when used with other automatic metrics such as BLEU or METEOR (ibid.). TER is similar to HTER and it is easy to confuse them. As Chatterjee et al. (2015, pp.158-159) put it, they both "measure the minimum edit distance between the MT output and its correct version." Their difference lies in the

"correct version", which "can be either a reference translation created independently from the MT output (TER) or a human post-edited version obtained by manually correcting the MT output (HTER)" (ibid.). TER is derived from WER (Word Error Rate), a metric initially used for automatic speech recognition systems.

Castilho et al. (2018) provide an overview of automatic MT evaluation metrics. In addition to almost all the above-mentioned metrics in this section, there are other automatic evaluation metrics such as GTM, CHRF (Popović, 2015), ULC (Giménez and Màrquez, 2008), MaxSim (Chan and Ng, 2008), RTE (Padó et al., 2009), TERp (Snover et al., 2006) and wpBleu (Popović and Ney, 2009). Popović (2018) explained the similarities and differences between Word Error Rate (Levenshtein, 1966) and Position-independent Word Error Rate (PER: Tillmann et al., 1997), and assessed the automatic evaluation tools Hjerson (Popović, 2011) and Addicter (Zeman et al., 2011). She discussed approaches like identification and analysis of unmatched patterns, and checking and evaluating specific linguistic features. She acknowledges the effectiveness of automatic evaluation tools because they are faster, cheaper and can achieve more consistent results than human evaluation. She indicates that those tools can be used for "estimating the distribution of error classes in a given translation output, as well as for comparing different translation outputs. They can also facilitate manual error classification by pre-annotation" (p.129), thus saving a lot of time and effort.

Although there are benefits to automatic evaluation metrics, such as helping to provide objective and reliable results (Castilho et al., 2018) and those mentioned in Popović

(2018) above, Way (2018) points out some limitations with them. For instance, BLEU is often used at the sentence level and not aimed at measuring overall translation quality, METEOR prefers longer MT output than the reference translation while TER prefers shorter MT output. These limitations can affect the accuracy of the results, especially when TQA is context-based. These automatic metrics lack the ability to evaluate articles as a whole in context. Way also argues that usually only one reference translation is provided for automatic evaluation, and implies that bad quality of this 'perfect' human translation can lead to questionable evaluation results of an MT system. Hassan et al. (2018) and Popović (2018) echo Way's (2018) point by admitting automatic evaluation relies too much on the reference translation, and the quality of the reference translation can influence the result. Even if the quality of the reference might be fine, it may be just one rendition of several possible, accurate renditions. Some of the automatic metrics take this into account by allowing for more than one reference sentence (though they do not usually check the quality of those sentences) (examples are Lin and Och, 2004; Owczarzak et al., 2007; Qin and Specia, 2015). Castilho et al. (2018) state that the automatic metrics work in a way that "is based on a number of assumptions that can raise some concerns as to their actual value" (p.26), and "their ability to assess syntactic and semantic equivalence in MT output is severely limited" (ibid.). In addition, they recommend automatic metrics to be used to evaluate MT on a document level instead of sentence/segment level in order to obtain more precise results. Koponen et al. (2012) point out the shortcomings of HTER while conducting two experiments investigating the connection between post-editing and cognitive effort. In the experiments, they highlighted the post-editing time (TIME), seconds per word (SPW), keystrokes (KEYS) and HTER as the main post-editing effort indicators. The results verify their point in that

HTER reflects what the final translation looks like, but it fails to reveal much about the cognitive effort required to produce that final result.

As discussed here, there are both advantages and disadvantages of automatic evaluation metrics, as with human evaluation discussed in the previous section. As Way (2018) emphasizes, even though BLEU has many well-known limitations, it is still the most popular metric used by MT system developers. Again, to assure the validity of TQA, triangulation is needed and is adopted by the researcher for this PhD study.

3.3.3 Generic frameworks used in industry

Doherty and Kruger (2018) provide a detailed discussion of human and machine-generated subtitles and captions. They remark that there is no international standard for AVT quality assessment, and the assessment is largely based on specified industry guidelines (see Section 3.3.3.1 for examples), varying by institution, media, language, and country. Most LSPs have their own way of evaluating the quality of subtitles. In this section, a discussion about the most popular industry approaches to TQA will be provided, with a focus on the recent MQM model.

3.3.3.1 Industry approaches

Translation service providers usually stick to standard approaches to assure translation quality, such as the ISO 17100[46] translation process standard, ISO 18587[47] process for post-editing of MT output, and ISO/TS 11669[48] guidance for translation projects. For ISO/TS 11669, Castilho (2016) states that it has provided guidance for the best practice in all phases of a translation project. The standard contains 21 translation parameters divided into five categories (source content, requirements for the target, production tasks, environment, and relationships). In addition to the ISO standards, the European Commission's Directorate-General for Translation also sets good standards for translation quality (Drugan et al., 2018), at least for institutional circumstances. However, in the translation industry, the evaluation models adopted usually rely on error-based approaches (Mitchell, 2015; Castilho et al., 2018; Lommel, 2018), which allow translators, reviewers, or linguists to count and classify errors. In the late 1990s, the industry started to use the LISA (Localization Industry Standards Association) QA Model (Jiménez-Crespo, 2009), which consists of a series of error categories and three severity levels (minor, major, or critical). The translation is scored by evaluators in terms of how many errors it contains, and the translation is defined as 'pass' or 'fail' based on a threshold set by the evaluator. Many translation companies have tailored their own models based on this model. However, the major limitation of this type of model is the 'one-size-fits-all' approach (Castilho, 2016). TQA models are rather static and few variables are considered, such as content type, communications function, end user

[46] https://www.iso.org/standard/59149.html (Accessed: 5 December 2019)
[47] https://www.iso.org/standard/62970.html (Accessed: 5 December 2019)
[48] https://www.iso.org/standard/50687.html (Accessed: 5 December 2019)

requirements, context, perishability, or mode of translation generation (TAUS, 2017). On the one hand, the 'one-size-fits-all' approach is helpful for standardization, but on the other hand, it is difficult to meet the various requirements of localization projects. After all, as mentioned in Section 3.1, translation quality is context dependent.

The SAE J2450[49] translation quality metric was created in 2001 and stabilized in 2016. The metric consists of seven primary categories (wrong term, syntactic error, omission and addition, word structure and agreement error, misspelling, punctuation error, miscellaneous error), two subcategories (serious and minor), numeric weights for each category, and two meta-rules to help users decide in case of ambiguity. The metric can be used regardless of language or translation method. However, a feature of the metric is that it does not measure errors in style.

In order to create a common standard framework that can meet the needs of users, TAUS has partnered with a number of companies to develop the Dynamic Quality Framework (DQF). [50] Quality is considered dynamic because translation quality requirements vary depending on the content type (e.g.: sales and marketing material, patents, financial documentation, support content etc.), the purpose of the content, and its end users. Using this framework, the quality issue can be considered in advance of the translation process. Similarly, as part of the QT21 project,[51] drawing on the LISA QA Model, an open source, custom quality assessment tool, MQM (Lommel et al., 2014), was developed by DFKI (German Research Center for Artificial Intelligence) to address

[49] https://www.sae.org/standards/content/j2450_201608/ (Accessed: 5 December 2019)
[50] https://www.taus.net/evaluate/dqf-background (Accessed: 5 December 2019)
[51] http://www.qt21.eu (Accessed: 5 December 2019)

the difficulty of one-size-fits-all quality assessment and make it accessible to all. Later, TAUS and DFKI agreed to harmonize DQF and MQM,[52] and the newly harmonized model can be used 'stand-alone' in all contexts (TAUS, 2015). According to TAUS (2016b, Online), "Both the analytic method of DQF and the MQM hierarchy of translation quality issues have been modified to share the same basic structure."

3.3.3.2 MQM

As Lommel (2018) and Eyckmans et al. (2009) put it, there are two approaches within translation evaluation: holistic and analytic. The former involves evaluating translation as a whole and evaluating its quality based on comprehensive criteria (see Garant, 2009; Jiménez-Crespo,2009; Williams, 2013). With this approach, evaluators can quickly determine whether the translation meets specifications (discussed later in this section) and get an overall impression of the translation, but no detailed feedback or remedy can be provided (Lommel, 2018). The latter is to analyse individual errors (see Phelan, 2017). This approach is useful for identifying specific errors, but it is time-consuming and does not provide an overall impression (only if you add up all the errors, which is also time-consuming). In addition, training of the evaluator is required for this approach (Lommel, 2018). MQM supports both holistic and analytic approaches (ibid.), which is an advantage.

[52] https://www.taus.net/evaluate/qt21-project (Accessed: 5 December 2019)

MQM is one of the latest initiatives that attempts to standardize TQA. It provides a systematic framework for describing and defining metrics for evaluating the quality of translated text and identifying specific issues within those texts (MQM, 2015). 'Customisability' is one of the key features of MQM: users can just use parts of the model based on their needs (Castilho, 2016). Note that this feature was not initiated by MQM, but already appeared in the study of Flanagan (1994), who presents a flexible and simple classification system for errors in MT. MQM is flexible, but not simple. On the contrary, it is very detailed and hierarchical. There are eight primary dimensions in the top level MQM hierarchy. Each of them contains a number of issue types, listed as follows: accuracy (18 issues), design (33 issues), fluency (39 issues), internationalisation (49 issues), locale convention (14 issues), style (7 issues), terminology (7 issues), and verity (7 issues) (MQM, 2015). The simplified version of MQM, MQM Core (see Figure 3.2) consists of 20 of the most common issue types arising in TQA (ibid.). It has to be noted that when identifying these issue types, Lommel et al. (2014) took inspiration from some existing metrics and tools, especially House (1997), thus many of them are similar with those identified in House (ibid.).

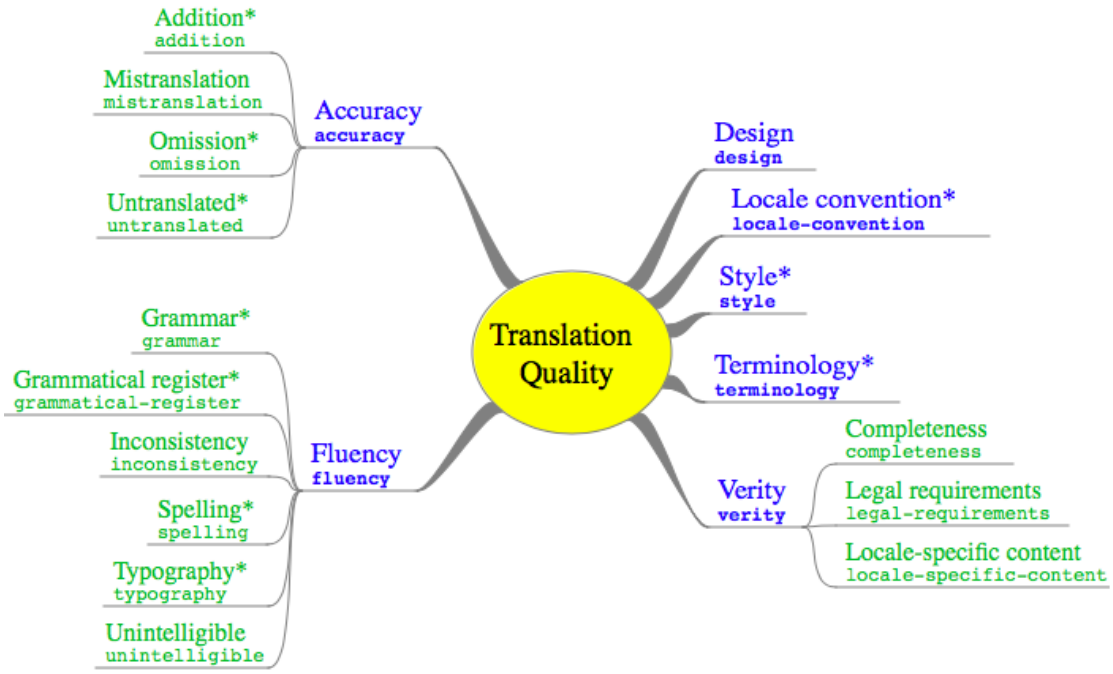

Figure 3.2 MQM Core (MQM, 2015)

The definitions of the categories in Figure 3.2 are presented in MQM. However, some of the definitions are not very precise. Take the definition of "overly-literal" as an example: it is defined tautologically as follows: "the translation is overly literal".[53] The repetition of "overly-literal" makes the definition non-sensical. Therefore, it is suggested to be selective when referring to the definitions of the categories in the model. The editors of the model were aware how problematic it is to standardize TQA; they try "to set forth a set of criteria that *all* translations *should* follow, regardless of purpose or other requirements" (Lommel, 2018, p.118). In other words, when designing TQA models, most translation experts ignore a frequently mentioned issue in this chapter: context. Or, in the words of TAUS, they ignore that quality is 'dynamic'. To tackle this issue, based on a functionalist perspective, MQM metrics are associated with a set of specifications

[53] Available at: http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html#overly-literal (Accessed: 6 December 2019)

that follow ASTM F2575-14, “which defines a standard set of 21 ‘parameters’ that describe the information needed to complete a translation project” (ibid., p.119). Therefore, before creating a customized MQM model, the creator needs to check the specifications first, then add or delete the issue types in compliance with the specifications. In this way, the evaluation process can be controlled from the very beginning.

When conducting TQA with MQM, errors are classified as belonging to one of four severity levels (the number in the parentheses refers to the weighted penalty points of the severity): critical (100 points), major (10 points), minor (1 point) and none (0 point). MQM also has a default scoring algorithm, which will be elaborated in Section 5.3.3. Since its launch, MQM has been widely used in the translation industry and integrated into the main CAT tools and TMSs (Translation Management Systems). Its users include, for example, Trados, XTM Cloud, Mozilla, eBay, to name just a few (Lommel, 2018).

3.4 Conclusions

Following an introduction of MOOCs, subtitles, MT and reception studies in Chapter 2, this chapter gives an overview of translation quality assessment, by both human and automatic methods, and also an outline of existing TQA models, especially the MQM model, which is adopted in this PhD research (see Section 5.3.3). It has been emphasized a few times in this chapter that triangulation is an important tool for research, a mixed-methods methodology is thus preferred to assure the validity of the research. This approach will be elaborated upon in Chapter 4.

Chapter 4 Methodology

4.1 Introduction

This chapter presents the methodology adopted in the research described in this thesis, shared by both the pilot and the main experiment. The chapter is structured as follows: Section 4.2 clarifies the mixed-methods approach adopted in this research, including eye-tracking, questionnaires, and TQA.

4.2 A mixed-methods approach

A number of scholars (Onwuegbuzie and Teddlie, 2003; Teddlie and Tashakkori, 2003; Creswell and Plano Clark, 2007) have discussed the benefits of adopting a mixed-methods approach to research. Over twenty years ago, Frey et al. (1991, p.124) proposed that combining quantitative and qualitative methods could help increase the measurement of validity and reliability of studies, and emphasized the importance of triangulation, which has been frequently mentioned in the previous chapter. Shreve and Angelone (2010) also stressed the significance of triangulation for achieving reliable data and results. Referring specifically to media audience research, Schrøder (1999, p.50) points out that triangulation "usually implies that the researcher seeks primary data about a research question in two (or more) different ways." In this PhD research, triangulation is used in a similar way to Schrøder (ibid.), to refer to the use of more than

one method, and specifically to a combination of qualitative and quantitative methods, to elicit different types of data about the reception of machine translated subtitles.

In recent years, a mixed-methods approach has been used in a number of translation studies, especially when cognitive effort forms part of the research question. Usually, mixed methods involving quantitative eye-tracking data and quantitative or qualitative questionnaire data are adopted for this kind of research. Examples can be found in Caffrey's (2009) investigation on the effects of an abusive subtitling procedure on the perception of TV anime, Perego et al.'s (2010) analysis of the cognitive effectiveness of subtitle processing, Doherty's (2012) research on the effects of controlled language on the reading and comprehension of machine translated texts, Secară's (2015) findings on creative spellings in subtitling, Castilho's (2016) study on measuring acceptability of machine translated enterprise content, and Filizzola's (2017) work on Italians' perception and reception of British stand-up comedy humour.

As for the current research, a mixed-methods approach is chosen because it is expected that the combination of quantitative and qualitative methods will provide different types of data that will help to explore in more depth the phenomenon being researched. The mixed-methods approach of this research follows Morse's (2016) mixed method sequential design. As she put it (ibid., p.17):

> "Sequential mixed method designs are those projects in which the supplemental project is conducted after the core project has been completed."

Morse (ibid.) points out that sequential design is usually planned at the proposal stage, when it is not, then it should be called emergent design, a subset of sequential design.

> "Emergent design is a study in which the supplementary component is added *after* the core component of the study is underway or completed and the researcher realizes that the present study is inadequate and that additional information – or answering a supplemental question – would greatly improve the research." (ibid.)

In regard to this research, eye-tracking and questionnaires are core, while TQA and frequency analysis are supplementary. The researcher uses eye-tracking to elicit quantitative and objective data on the cognitive processing of translated subtitles, and questionnaires and TQA tools to elicit qualitative, perception data regarding human and machine translated subtitles. Frequency analysis is used to generate quantitative data about these subtitles, from which qualitative inferences can be made. Eye-tracking data was collected to assist in interpreting the behaviour of participants. It was analysed using statistical tests, which supports the objectivity and reliability of the research. Questionnaire data, on the other hand, reveals the traits and subjective experiences of participants, which helps us to discover any links between their perception and reception of subtitles (see Section 2.4). As supplementary methods, TQA and frequency analysis were used to explore the difference between human translated subtitles and post-edited machine translated subtitles.

Data from the supplementary component and that from the core component are analysed separately, and the results of the former are incorporated into the results of

the latter, so as to provide information and insight for the research findings of the latter and improve the whole research (Morse, 2016).

Consistent with the prior relevant studies, the researcher believes that the combination of eye-tracking and questionnaire can provide a robust set of methods to investigate the perception and reception of subtitled AV content. The addition of TQA and frequency analysis, which is new compared to previous studies, can help provide more insight from a translation-product perspective. It has to be noted that the researcher used a combination of just eye-tracking and questionnaires for the pilot study (see Section 5.2), later, TQA and frequency analysis methods were added to the main experiment (see Section 5.3). The reason for doing this is because when conducting the main experiment with a larger sample, the researcher realized the study was inadequate and needed additional information. Therefore, as the research matured, the ideas of doing TQA and frequency analysis (supplementary component) *emerged*. They were thus added in an attempt to explain findings of eye-tracking and questionnaires (core component). As Morse (2016, p.15) states, "The supplementary component is regarded as complementary to the core component of the project, providing answers of adequate quality, so that researchers may progress with their research with certainty."

In short, a mixed-methods study involving eye-tracking and questionnaire was used for the pilot; and a mixed-methods study (of emergent design) involving eye-tracking, questionnaire, TQA and frequency analysis, was adopted for the main experiment. To the best of my knowledge, these particular combinations of methods have not been adopted in any other research on machine translated subtitles of MOOC content to date.

In particular, the design used for the main experiment has not been seen in any other source in translation studies.

4.2.1 Eye-tracking

Section 2.4.2.1 has reviewed the literature in which eye-tracking is used to investigate AVT. Here, the description of an eye-tracker is reiterated as follows: "an eye-tracker is a device that monitors and records the movement of an individual's eyes… [it] contains special diodes that reflect light off the pupil and monitor fixations, gaze paths and pupil size as the subject interacts with an object on-screen" (O'Brien, 2006, p.185). It can be seen in Section 2.4.2.1 that eye-tracking has been commonly used for exploring the mental activities of viewers and their allocation of attention. For this research, on the one hand, the eye tracker records gaze behaviour data from which participants' shifts of attention between subtitle and image can be inferred, the order in which parts of the image are perceived, and how often subtitle regression (where the eyes travel back across words already read) occurs. In this way, the perception of subtitled AV content can be objectively detected. On the other hand, the eye tracker records data including fixation duration, fixation count, visit count and visit duration that can reflect participants' processing effort when watching the subtitled AV content (see Section 2.4.2.1 for more information).

As mentioned in Section 2.4.2.1, the use of eye-tracking in reception studies is based on the eye-mind hypothesis that what a person is looking at is strongly correlated to what he/she is thinking about (Just and Carpenter, 1980), which suggests that eye movement recordings can provide a dynamic track that shows the direction of a person's attention

relative to the visual display. Here, eye movements "exist in humans to either maintain or shift an observer's gaze between visual areas, while keeping the projected visual image on a specialized, high acuity area of the retina" (Goldberg and Wichansky, 2003, p.9). Measuring eye movements, such as fixations, can show how much is being processed at the "point-of-regard"[54] (Poole and Ball, 2006, p.206). The different eye-tracking metrics generated may be related to the "subjective scales of perceived workload" (Goldberg and Wichansky, ibid., p.8). As the cognitive workload increases, the saccade[55] or coverage on the screen decreases, which is consistent with a narrower focus of attention (ibid.). If cognitive processing demands are difficult, fixation duration may increase (Goldberg and Kotval, 1999). In some cases, visual attention may precede or lag behind the current fixation point (Rayner and McConkie, 1976), thus reducing the sensitivity of eye movement as an indicator of cognitive processing, for example, one's semantic processing is usually not synchronized with the perceptual stimulus input when reading a text (Goldberg and Wichansky, 2003). Apart from this disadvantage, the effect of eye-tracking varies significantly from person to person, an eye tracker can track over 90% of eye movements of some people, but fails to track the eyes of others (this problem was encountered by this research, see Section 6.2 for more information). Researchers should bear in mind these issues before implementing an eye-tracking experiment. Another thing that needs to be highlighted is calibration, which is of critical importance for eye-tracking. It is conducted as the first step to see how much the participant's point-of-regard matches the corresponding location on the screen. During

[54] Point-of-regard: point in space where a person is looking. Usually used in eye tracking research to reveal where visual attention is directed. (Poole and Ball, 2006, p.216)

[55] Saccade: an eye movement occurring between fixations, typically lasting for 20 to 35 milliseconds. The purpose of most saccades is to move the eyes to the next viewing position. Visual processing is automatically suppressed during saccades to avoid blurring of the visual image. (ibid.)

the calibration process, there is usually a moving dot on the screen. The dot continuously moves to a series of locations, usually near the corners of the screen, and the participant needs to follow the dot and fixate the locations on the screen. Good calibration is a prerequisite for valid eye-tracking. To ensure accuracy, it is better to conduct calibration multiple times for each participant.

In regard to what data is gathered by an eye-tracker, as Goldberg and Wichansky (2003) note:

> "The eye tracking system gathers x/y location and pupil size information at typical rates from 30 Hz – 250 Hz, with 60/50 Hz typical for usability evaluation needs. Slower sampling rates do not provide sufficient resolution of visual attention, especially for scrolling or other tasks involving moving visual targets. Faster sampling rates can create a massive data reduction problem for typical usability evaluations. The x/y sample locations are then further processed into fixations, which may be assigned to experimenter-defined areas-of-interest on the viewed display or scene. Fixations and saccades can be combined into scanpaths[56], providing further information about one's cognitive processing during a task..."

Here, "areas-of-interest" (AOI) is an analysis method used in eye tracking. Poole and Ball (2006) state that in the practice of HCI (Human-Computer Interaction) studies, researchers define "areas of interest" on certain areas of the screen or interface being evaluated and analyse eye movements within those areas. In this way, useful information, such as visibility, placement and meaningfulness of specific interface elements, can be inferred from eye movement recordings, and then be used for the better design of the interface. As with HCI studies, by using eye-tracking as a research

[56] Scanpath: an eye-tracking metric, usually a complete sequence of fixations and interconnecting saccades. (ibid.)

method, researchers studying the reception of subtitles also need to define "areas of interest" (see Figure 5.1) before any further analysis of eye movement recordings.

The eye tracker used for this PhD research is the portable SMI REDn Scientific eye-tracker (which collected data at a rate of 60 Hz). Reviewing the relevant literature on eye-tracking, it is found that the definitions of eye-tracking metrics are inconsistent and often have overlaps, both across researchers and manufacturers. Meanwhile, there is also some terminological variation in the literature. Because an SMI eye-tracker was used for this research, it has been decided to adopt the terminology defined by the eye tracker manufacturer as follows (examples of the other terms for the same concept are noted in parentheses):

- **Fixation Count**: number of fixations inside the AOI (area of interest) (SMI, 2015, p.295).
- **Glance Duration**: Saccade duration for entering the object + sum of all fixation durations and saccade durations before the eyes begin to leave the AOI = dwell time + duration of saccade entering AOI (SMI, 2015, p.280). (This term corresponds to the 'visit duration' in Castilho and O'Brien, 2016, for example.)
- **Glance Count**: number of glances to a target (saccades coming from outside) within a certain period (SMI, 2015, p.295). (This term corresponds to the 'visit count' in Castilho et al., 2014, for example.)
- **Average Fixation Duration**: sum of durations of all fixations divided by number of fixations in the trial (SMI, 2015, p.287). (This term corresponds

to the ‘average fixation length’ in Alves et al., 2009, and the ‘mean fixation duration’ in Castilho and O’Brien, 2016).

4.2.2 Questionnaires

This section focuses on discussing the advantages and disadvantages of questionnaires, a widely-used measure in questionnaires – Likert scales, and the design of questionnaires which includes an introduction of the Technology Acceptance Model (TAM).

4.2.2.1 Advantages and disadvantages

Questionnaires are perhaps the most common way of collecting data from people, whether they are a small or a large group. According to Gratton and Jones (2010), questionnaires are ideal to collect a large volume of simple data, but unlikely to yield more complex information. They have summarized the advantages and disadvantages of questionnaire survey as follows (pp.128-129):

1) Advantages

- Accessibility: allows the collection of data with a low cost and make the investigation of a large sample possible;
- Potential reduction in bias: a well-designed questionnaire can avoid bias in the results, which may be the case in an interview;

- Anonymity: respondents may feel more comfortable when answering questions (especially sensitive ones) without the presence of the researcher, thus may help improve the validity of the responses in certain cases;

- Structured data: questionnaires provide highly structured quantitative data that can be easily compared and converted into tables or charts for statistical analysis;

- Increased time for respondents: participants can complete questionnaires at any time and they can always go back to the questions at a later time if they recall anything or want to correct answers.

2) Disadvantages

- Potential problems over complex questions: simple and clear questions are preferred due to little opportunity for the respondents to seek clarification;

- No control over who completes the questionnaire: the person who completes the questionnaire cannot be controlled whether the researcher specifies the respondent or not;

- No opportunity to probe: no opportunity to get back to respondents for clarification;

- Potentially low response rates: response rates from questionnaires are notoriously low, and this may have a bad impact on the reliability of the study.

The role of questionnaires is to elicit the information needed to help researchers find the answers to the objectives of the survey (Brace, 2013). The first task of any survey is to determine the objectives to be answered (ibid.), which are correlated to the researchers' research questions. Before asking questions, the participants, in other words, the sample, must be defined, and the sampling method and the data collection medium (interviewer-administered interviews or self-completion survey) must also be

determined (ibid.). Self-completion surveys can be based on paper questionnaires or web-based self-completion, both of them having their benefits and drawbacks. For web-based self-completion, as used in this PhD research, the advantage is that it can be designed to have visual appeal and to allow participants to decide the time to complete the questionnaire flexibly, and to answer questions more honestly with less social desirability bias due to the absence of an interviewer; the disadvantage lies in that because there is no interviewer present, no one can clarify a question or correct a misunderstanding for the participant, which strengthens the requirements for the questionnaire designer to make the questionnaire "clear, unambiguous and engaging" (Brace, 2013, p.28). Another disadvantage of questionnaires worth mentioning is that it is a self-report method, so it can be very subjective. Again, this underlines the importance of using a mixed-method approach, which can provide multiple types of data for the research and lead to a more comprehensive result.

4.2.2.2 Likert scales

Questionnaires usually consist of a set of questions with a range of answers for gathering structured[57] data. Open questions can also be included to allow participants to answer questions in their own way (Matthews and Ross, 2010). All the participants get the same questionnaires, and they are asked the same questions in exactly the same way. After all, "the role of the questionnaire is to provide a standardized interview across all subjects" (Brace., 2013, p.3). Different types of data collected through questionnaires

[57] Structured: describes data, or a data collection method (such as an interview or questionnaire), in which the questions are the same for each participant, and typically there is a common set of answers for each question. (Matthews and Ross, 2010, p.201)

are as follows: facts, descriptions, knowledge, opinions, attitudes/values, and background information (Matthews and Ross, 2010). While some questionnaires are factual and objective, researchers have developed ways to measure participants' opinions, attitudes and values, for example, through the Likert scale.

> The Likert scale is "a rating measure used widely in survey questionnaires to assess respondents' attitudes, preferences, and subjective responses to statements. Developed by Rensis Likert, a set of items presents users with statements and standardized response categories on a continuum, such as *strongly agree, agree, neither agree nor disagree, disagree,* and *strongly disagree*". (Sullivan, 2009, p.293)

The goal of using a Likert scale is to measure the intensity of participants' feelings about a certain issue or theme (Bryman, 2016), or "single 'latent' variable (phenomenon of interest)" (Joshi et al., 2015, p.398). In the questionnaire, the 'latent' variable is expressed by the 'manifested' statements (Joshi et al., ibid.). Due to its measuring purpose, scoring is connected to the Likert scale. For instance, "strongly agree" is assigned a score of 5, indicating that the participant has a strong positive feeling towards the issue, and "strongly disagree" is assigned a score of 1, indicating that the participant has a strong negative feeling towards it. The Likert scale is usually a 5- or 7- or 9-point ordinal scale. However, there is no consensus on the optimal number of scale points (Brace, 2013). The only consensus is that it is between 5 and 10. Many researchers think 7 is the ideal number (Krosnick and Fabrigar, 1997), but there is also debate about that, and whether extending the number to 10 or above increases the validity of the data. When deciding the scale points, the questionnaire designer must take into account the possible differences between these points and the ability of the participants to distinguish these points (Brace, ibid.).

When designing the attitude survey, some issues must be taken into account (Brace, 2013, p.99): a) whether or not the statement is balanced; b) whether it leads the respondent to a specific answer; and c) how the addition or removal of wording may affect how respondents answer. In regard to the number of statements in the attitude survey, on the one hand, there should be sufficient statements to adequately address all of the attitudes to be studied; on the other hand, the number shall not exceed the participants' boredom threshold, which varies according to their interest in the subject, so as to avoid pattern responding (ibid.).

According to Bryman (2016), when considering whether a measure is reliable, one of the prominent factors is the "internal reliability" (p.157), which means the statements that make up the scale should be interrelated and consistent. In other words, they should indicate whether participants' scores on any one statement are related to their scores on the other statements. Because the overall score of each participant is formed through aggregating his/her answers to each statement, it is important to ensure coherence across all statements. Many researchers use Cronbach's alpha to test the internal reliability.

> "Essentially it calculates the average of all possible split-half reliability coefficients. A computed alpha coefficient will vary between 1 (denoting perfect internal reliability) and 0 (denoting no internal reliability). The figure 0.80 is typically employed as a rule of thumb to denote an acceptable level of internal reliability…" (Bryman, 2016, p.158)

According to Iarossi (2006, p.11), after the questionnaire is finalized, it should be coded immediately and a data entry form shall be developed. A good data entry form will have built-in consistency checks to prohibit invalid entries. However, it is not easy to develop a data entry form. It usually requires a professional programmer for the job because the form has to include some intricate cross-references and checks (ibid.). Thanks to the rapid development of technology, some online questionnaire websites, for example, www.wjx.cn, which was used for this PhD research, has already provided this consistency check service. By one click, researchers can simply get the Cronbach's alpha for all their questionnaires automatically.

4.2.2.3 Questionnaire design

There are some key issues to be considered when designing a questionnaire: the range and scope of questions to be included, question types (open-ended, closed and semi-open-ended), content of individual questions, question structure, question order, question wording, layout and presentation of the questionnaire (Catley, 1999; Lancaster, 2007; Matthews and Ross, 2010;). Matthews and Ross (2010, pp.207-211) have summarized the most common questions in questionnaires:

1) Yes/No question
   This type of question is usually used for distinguishing between different groups of respondents.

2) Which category? How many? How much?
   This type of question is usually used when you want to ask the respondents about themselves or a situation, and it needs to be presented in a way for the respondents to answer it easily and accurately.

3) Choose from a list
   This type of question is usually used when you want the respondents to choose one or more from a set of answers. A “tick all that apply” and an “other” categories (often needs specification) are usually included.

4) Agree/disagree with a statement
   This type of question is used for gathering respondents’ opinions, attitudes and values, usually by Likert scale or rating scale.

5) Open question
   This type of question is usually used for explaining an answer to a previous question, or inspiring respondents to give their own ideas, or asking respondents to add anything they want to say about the topic (in this case the question will be put in the end).

6) Filter question
   This type of question is usually used to help respondents find their own way through the questionnaire, and researchers can thus ‘filter’ certain types of participants based on whether a question is related to them or not.

From the literature reviewed in Section 2.4.2.2, it can be seen that questionnaires are widely used for measuring viewers’ perception and comprehension of AV content. In such questionnaires, most of the abovementioned question types are included. For example, Likert scales are commonly used for viewers to rate their understanding of and opinions about a video (Antonini, 2005; Chiaro, 2007; Fernández-Torné and Matamala, 2016), multiple-choice questions are used for testing viewers’ comprehension of the video content (Perego and Del Missier, 2008), and viewers are allowed to add comments to explain their answers (Chiaro, 2004; Antonini, 2005; Bucaria, 2005; Schauffler, 2012). In addition, questionnaires are also used for collecting participants’ demographic information and their attitudes towards some specific type of videos (Chiaro, 2007). Besides, in the experiment of Caffrey (2009), subjects were often asked to answer questions to rate the processing effort they required, either by answering directly about how much effort they required to view a video, or by answering indirectly about their

presumption of the relation between other values and processing effort. Therefore, we can see that questionnaires can also be used as a complementary way of measuring processing effort.

4.2.2.3.1 Technology Acceptance Model

To design questionnaires, especially those for user research on a particular technology, a model can be drawn upon, which is the Technology Acceptance Model (TAM). This model was firstly proposed by Fred Davis in his doctoral thesis in 1985. Davis (1985) suggests that users' motivation can be explained by three factors: Perceived Usefulness (PU), Perceived Ease of Use (PEOU), and Attitude Toward Using the system. The attitude of the user, in turn, was considered to be influenced by the other two factors, with PEOU also having a direct influence on PU. During later experimentation stages, he refined his model to include other variables and modified the relationships that he initially formulated (see the first modified version of TAM in Davis et al., 1989; the final version of TAM in Venkatesh and Davis, 1996; TAM 2 in Venkatesh and Davis, 2000). Meanwhile, other researchers have also applied and improved TAM (see the extended TAM in Venkatesh, 2000; UTAUT model in Venkatesh et al., 2003; TAM 3 in Venkatesh and Bala, 2008). TAM has evolved into a popular model in explaining and predicting system use over time and has been cited in many studies that deal with user acceptance of technology (Lee et al., 2003). For example, Schepers et al. (2005) conducted a research project in a Dutch high-technology company. They modelled and tested two leadership styles (transactional and transformational) as antecedents to PU and PEOU of new technologies. In the surveyed company, transformational leadership positively

influenced PU of the technology whereas transactional leadership did not display any significant effects. Park et al. (2009) collected data from 16 institutions in Africa, Asia, and Central/Latin America, and did a survey on using a digital library system. A path analysis revealed that PEOU of the library system had a significant impact on PU, which ultimately led to behavioural intention to use.

Van Raaij and Schepers (2008) report that the core TAM relationships hold just as well in a Chinese setting as they do in Western countries. The majority of hypothesized relationships are supported by the data. This suggests that, contrary to the findings of Straub et al. (1997) and McCoy et al. (2005), but consistent with the findings of Ong et al. (2004) and Pituch and Lee (2006), TAM does hold across cultures. Besides, Mao et al. (2005) and Schepers and Wetzels (2007) discovers that PU seems important in Western cultures, while PEOU has more relevance in non-Western studies. Obviously, different research leads to different findings. In China, Qi et al. (2009) investigated the reasons why people use mobile data services in China. They presented an extended TAM model and tested it with the data collected from 802 mobile subscribers. Their findings indicated that mobile subscribers' PEOU had obviously positive influence on usage attitude directly. It also positively affected PU. Additionally, brand experience also had a large influence on subscribers' attitudes towards mobile data services. This study reports that sufficient training is still an important way to secure the adoption of technology. Yoon (2009) explored the effect of Hofstede's six dimensions of national culture on consumer acceptance of e-commerce in China. He found that uncertainty avoidance and long-term orientation had moderate effects on the relationship between trust and intention to use, and masculinity also had a moderate effect on the

relationship between PU and intention to use and the relationship between PEOU and intention to use, while power distance and individualism had no significant effects. Guo et al. (2010) researched the social network services (SNS) in China by the use of an integrated model of centrality, trust and technology acceptance. Their findings suggest that centrality, technology acceptance, familiarity and user trust are important variables in users' intention of using e-socializing services, and SNS providers should take factors like the number of social ties, channel of service promotion and web interface design into consideration when developing strategies. A study by Wu and Chen (2017) is close to this PhD research. By integrating the technology acceptance model (TAM), task-technology fit (TTF) model, features of MOOCs and social motivation, they proposed a unified model to research Chinese users' continued intention to use MOOCs. They recruited 252 Chinese MOOC users as participants and results show that PU and attitude are key to the continued intention to use MOOCs; PEOU, task-technology fit, reputation, social recognition and social influence have a big impact on predicting continued intention, while PU is an important mediator of the effects on continued intention; PEOU can be affected by individual-technology fit, task-technology fit, and openness; PEOU and social influence have no major influence on attitude; and PU is not affected by individual-technology and openness.

For this PhD research, the Technology Acceptance Model (TAM) was utilised for designing the attitude survey. Some concepts in TAM, such as perceived usefulness, perceived ease of use, perceived enjoyment, perceived quality, intention to accept the technology, and compensation, have been borrowed for designing the statements in the attitude survey (see Section 5.2.3). The definitions are as follows (Hu and O'Brien, 2016):

Perceived usefulness: The degree to which a person believes that machine translated subtitles would enhance his or her job performance.

Perceived ease of use: The degree to which a person believes that using machine translated subtitles will be free of effort.

Perceived enjoyment: The extent to which the activity of using machine translated subtitles is perceived to be enjoyable in its own right.

Perceived quality: The perceived level of the quality of machine translated subtitles.

Intention to accept machine translated subtitles: A person's behavioural intention to accept machine translated subtitles.

Compensation: The degree to which a person believes that he or she has the ability to comprehend machine translated subtitles using additional inputs.

From the literature reviewed on TAM (e.g.: Venkatesh and Davis, 2000; Legris et al., 2003; Wu and Chen, 2016; Wu et al., 2007), it can be seen that there are rules to follow for the questions used for investigating each concept. Examples are as follows:

1) Perceived usefulness:
   Using the system/application improves my performance.
   Using the system/application increases my productivity.
   I find the system/application to be useful in my job.
   I believe the system/application improves my performance.

2) Perceived ease of use:
   My interaction with the system/application is clear and understandable.
   Interacting with the system/application does not require a lot of my mental effort.
   I find the system/application easy to use.

3) Perceived enjoyment:
   I think using the system/application shall be interesting.
   I enjoy using the system/application.

4) Perceived quality:
   The overall quality of the system/application is high.

The system/application generally function well.
I'm satisfied with the system/application.

5) Intention to use:
   Assuming I have access to the system/application, I intend to use it.
   Given that I have access to the system/application, I predict that I would use it.
   I plan to use the system/application in the future.
   I would recommend the system/application to my friends if they need.

6) Compensation:
   I could complete the job using the system/application if there was no one around to tell me what to do as I go.
   I could complete the job if I could call someone for help if I got stuck.
   I could complete the job if I have a lot of time.
   I could complete the job if I had just the built-in help facility for assistance.

While the literature on questionnaires reviewed in this chapter helps the researcher to structure the questionnaires and individual questions, TAM approaches help to create the 'content' or 'substance' of those questions.

4.2.3 TQA

Chapter 3 surveyed the approach to TQA taken in contemporary translation research. This section will add to the discussion on what TQA method has been adopted for this research. It can be seen that TQA plays a significant part in both MT and human translation studies, and both in academia and industry. In this PhD research, TQA is used as a supplementary method for the main experiment to gain more insight into the unexpected results of the main experiment with participants receiving HT subtitles and PE subtitles, and to explore the quality of HT subtitles and PE subtitles and their differences, which is also why raw MT subtitles were excluded here. Besides, it was believed that the quality of raw MT subtitles is the lowest of the three, while the

significance of this TQA lies in 'comparison', it is not of great value to add raw MT subtitles into it.

There are plenty of different TQA approaches and models for use. For example, the MQM model that has been discussed in detail in Section 3.4.2. This model has many advantages: it is recent, flexible, detailed and hierarchical. It can be customized, which means users can just use parts of the model, add or delete issue types based on their needs. Therefore, it has been decided to use this model for this research. More specifically, the TQA metrics are customized and adapted from the MQM Core (see Section 5.3.3).

According to Lommel et al. (2014, p.33), "Demonstrating a high degree of IAA is a necessary step to showing that an assessment metric is reliable." Cohen's kappa coefficient (κ) is usually adopted for measuring pairwise agreement among annotators who assess machine translated output (Bojar et al., 2014 and 2016; Lommel et al., 2014; Klubička et al., 2017). However, because of its 'pairwise' feature, it can only work when not more than two raters (evaluators) are involved. When there are more than two raters, some researchers still use Cohen's kappa: they calculate the κ between each pair of raters and take the average score (Lommel et al., 2014). An alternative way is to use Fleiss' kappa, which assesses "the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items" (Grůber and Matoušek, 2010, p.288). Fleiss' kappa is defined as follows:

$$\kappa = \frac{\bar{P}_o - \bar{P}_d}{1 - \bar{P}_d}$$

where:

$\overline{\mathrm{P}}_{\mathrm{d}}$: the agreement by chance

$\overline{\mathrm{P}}_{\mathrm{o}}$: the average agreement between the raters

As Schmitt and Minker (2012, p.32) state, "the factor $1 - \bar{P}_d$ gives the degree of agreement that is attainable above chance, and $\bar{P}_o - \bar{P}_d$ gives the degree of agreement actually achieved above chance." According to Landis and Koch (1977), the interpretation of κ is as follows: 0.00-0.20 is slight, 0.21-0.40 is fair, 0.41-0.60 is moderate, 0.61-0.80 is substantial, and 0.81-1.00 is almost perfect. It has to be noted that the inter-annotator agreement for TQA can be considerably low (e.g.: 0.25 – 0.34 in Lommel et al., 2014; 0.35 – 0.55 in Klubička et al., 2017).

In this research, ten evaluators completed the quality assessment. If Cohen's kappa is adopted, then the κ for 45 pairs needs to be calculated in order to get the average score. However, Fleiss' kappa can be easily used for calculating the results of ten evaluators. Therefore, it was decided to adopt Fleiss' kappa for calculating IAA in this research (see Section 7.3 for more information).

4.2.4 Frequency analysis

Frequency is at the heart of corpus linguistics as a method (Ellis, 2002; Wray, 2002; Gries, 2010; McEnery and Hardie, 2011). Researchers are interested in absolute or relative frequencies (e.g.: number of instances per million words) of a given word or multi-word expression. Also, they are interested in comparing the distribution of words across different corpora, or between two different expressions in one corpus (see Kilgarriff, 1996; Rayson and Garside, 2000; Oakes and Farrow, 2006; Gries, 2010). Methods based on frequency are not as trivial as they might seem because researchers need to consider whether they are dealing with lemmatised frequencies or frequencies across languages. In this PhD research, the researcher was comparing monolingual frequencies, and the language in study was Chinese, hence there is no need to discuss lemmatisation here. The interest in conducting a frequency analysis emerged only after conducting the TQA. The reason for doing so is to explain why no difference was found between PE and HT subtitles in TQA. In order to get more insights, it was decided to conduct the frequency analysis by using both an out-of-domain corpus and an in-domain corpus, which means a corpus of general content and a corpus of subtitles. The motivation for comparing strings (a given word or multi-word expression) from HT and PE subtitles to a subtitle corpus on the one hand and a general corpus on the other hand was to find out how consistent HT subtitles were with these two language varieties as compared to how consistent PE subtitles were with the same two varieties.

As for what frequencies can indicate, Gries (2010, p.5) raised an example by looking for two word forms “give” and “bring” in the British Component of the International Corpus

of English (contains 1,061,263 words), returning 441 and 197 matches respectively. Therefore, “give” has more occurrences than “bring” in this corpus. In the same way, in a Chinese corpus, different strings will return a different number of matches, in other words, different frequencies of occurrence. The interest of collocation of formulaic language for psycholinguists is in its implications for an understanding of processing and acquisition. Psycholinguists are interested in the collocation of formulaic language because of its importance to understanding processing and acquisition (McEnery and Hardie, 2011). If a form is seen as very frequent in a corpus, this is taken as being ‘familiar’ to speakers of that language and if a multi-word string is frequent, the same applies – this means they are easier to process than strings that are not so frequent. This is especially the case with multi-word units. Higher frequency multi-word units mean a text is easier to process than those that are less frequent. In the scenario of this PhD research, the researcher was interested in the frequency occurrences of different Chinese strings in one Chinese corpus. The strings that were chosen for analysis were those for which an English word or phrase was translated differently in the human translation and post-edited MT and played an important role in differentiating the two translations. It has to be highlighted that the segmentation for this analysis is based on the English word or phrase, rather than on Chinese n-grams, the length of which can differ. For example, the PE output for “move about” is “动一动” (1-gram), while the HT output is “四处走走” (2-gram) (see Section 8.3 for more information).

It has to be emphasized that frequency analysis is a supplementary method for the main experiment only. Here, raw MT subtitles were excluded again for the same reason as in the TQA. Note that only the ten sentences with the lowest BLEU scores were analysed.

Because the frequency analysis is just a supplementary method emerged after the unexpected TQA results, and the lower the BLEU score, the greater the difference between raw MT and the reference translation. Also, due to little post-editing was made (though it was full post-editing) to raw MT subtitles, it was believed that a great difference between raw MT and HT means a great difference between its post-edited version and HT (see Appendix G). Therefore, only ten sentences with the greatest difference between PE and HT were compared to obtain the most obvious results.

Chapter 5 Design and Implementation

5.1 Research design

This research sets out to investigate machine translated subtitles' impact on viewers' reception of MOOC content. The main research question and main hypothesis can be described as follows:

> **Main research question:** is there a difference in reception between participants who are offered raw MT subtitles and those who are offered full post-edited subtitles or human translated subtitles?
>
> **Main hypothesis:** participants who are offered full post-edited subtitles or human translated subtitles will score higher on the reception metrics compared with those who are offered raw MT subtitles.

As suggested above, the researcher is interested in the difference (if any) between viewers' reception of raw MT output, post-edited output and human translation. It is assumed that the human translated content and full post-edited content is more acceptable and comprehensible than raw machine translated subtitles. Hence, it is assumed that people who are offered full post-edited and human translated subtitles score higher on the reception metrics than those who are offered raw machine translated subtitles. Moreover, it is also assumed that people who are offered human translated subtitles score higher on the reception metrics than those who are offered

full post-edited subtitles. The participants were divided into three groups for this purpose. One group (Group RAW) was presented with raw machine translated subtitles, one (Group PE) was presented with full post-edited subtitles, and another one (Group HT) was presented with human translated subtitles.

5.1.1 Reception model

The concept "reception" is operationalized in this study using Gambier's (2009) model, which has been discussed in Section 2.4.1. Table 5.1 shows how Gambier's classification is implemented throughout the study.

Table 5.1 Reception Model and Associated Measurement Tools

Element	Related to	Reflected in	Measured by
Response	Perceptual decoding	Attentional processes	Eye-tracking
Reaction	Psycho-cognitive level	Processing effort and comprehension	Eye-tracking and comprehension testing
Repercussion	Attitudinal issues and sociocultural dimensions	Attitudes and beliefs	Attitude questions

Response refers to the initial physical response of a viewer to an audio-visual stimulus, in this case, the subtitle and the rest of the MOOC image. In the experiment in this study, it was measured using glance count, glance duration, and fixation count, which indicate where the viewer's attention is directed. These data and other eye-tracking data were

captured using an SMI REDn Scientific eye-tracker, which collected data at a rate of 60 Hz. (See Section 4.2.1 for a more detailed methodological discussion on eye-tracking)

Reaction involves the cognitive follow-on from the initial response, and is concerned with how much effort is involved in processing the initial stimulus and what is understood by the viewer. It was measured partly through average fixation duration, a typical effort indicator in eye tracking, and partly through testing viewers' comprehension of the MOOC content using specific questionnaire items. (See Section 5.2.3 for a more detailed methodological discussion on comprehension testing)

Repercussion refers to attitudinal and sociocultural dimensions of AVT consumption. It was captured using targeted questionnaire items. (See Section 5.2.3 for a more detailed methodological discussion on attitude survey)

5.1.2 Participants

According to the 2016 White Paper discussed in Section 2.2.2, in China, MOOCs are mostly popular among university students who are between 18 and 25 years old. Since English is a compulsory subject in Chinese schools, most Chinese university students have been learning English for more than 5 years. Most would therefore have some knowledge of English, but their level of English proficiency varies considerably. According to the English Proficiency Index 2018 issued by EF (Education First),[58] a well-

[58] http://liuxue.ef.com.cn/epi/ (Accessed: 3 March 2019)

known international English education company, the index for China is 51.94 out of 100, which indicates Chinese people have low proficiency in English, broadly speaking. In terms of this index, China ranks 9 in the 21 countries in Asia, and 47 in the surveyed 88 non-native English speaking countries in the world. Taking this fact into account, in this experiment, I decided to recruit Chinese undergraduates who have a low level of English proficiency as participants. Note the participants here only refer to those who participated in the main experiment. Participants in the TQA were a group of different people and will be elaborated in Section 5.3.3.

As a PhD student myself, I recruited participants for the pilot and main experiment in universities through my own networks, both in Ireland and China. More information about the participants in the pilot is given in Section 5.2.1, while more information about participants in the main experiment is given in Section 5.3.2.

5.1.3 MOOC video

The White Paper (HCR, 2016) discussed in Section 2.2.2 suggests that most Chinese users of MOOCs prefer the length of a MOOC video to be 10-20 minutes. In order to be of interest to a considerable number of potential participants, it was decided that the MOOC used in the current research should be general and not specialise in the sciences or the humanities. In addition, the MOOC must have no existing Chinese subtitles, because if there were existing subtitles and they were included in the MT training data, this would have an impact on the validity of the research. Imagine if there were existing

human translated subtitles for this MOOC, and coincidently they were included in the training data of the MT engine (Google Translate) used for this research, then if we use this MT engine to translate the original English subtitles of this MOOC, the result is likely to be influenced by the human translation that was previously used as training data. The best practice is to ensure that test data has not been included in MT training data.

After consideration of these requirements, three MOOC videos on Coursera were shortlisted as follows: "What is Physical Activity?"[59] (6:59 in length) under the MOOC "Sit Less, Get Active" (produced by The University of Edinburgh), "Decision Making and Problem Solving in Organizations"[60] (10:02 in length) under the MOOC "Introduction to Problem Solving" (produced by the University of California, Irvine Extension), and "Welcome"[61] (9:14 in length) under the MOOC "Influencing People" (produced by the University of Michigan). All three MOOC videos have no Chinese subtitles (as of April 2017) and allow people to audit for free. Eventually, the first MOOC on the topic of physical activity was selected for the following reasons: 1) it did not simply present an instructor talking with slides, but consisted of various interviews and scenes with sporting activity; 2) it was composed of vivid talks and colourful images and there were many interesting shot changes; 3) speakers ranged from children to older people with unfamiliar accents (for Chinese students). All of these features made the video interesting and challenging for viewers with limited English, which made the use of

[59] https://www.coursera.org/lecture/get-active/what-is-physical-activity-CyvKV (Accessed: 2 April 2017)
[60] https://www.coursera.org/lecture/problem-solving/1-1-decision-making-and-problem-solving-in-organizations-Sn8cY (Accessed: 2 April 2017)
[61] https://www.coursera.org/lecture/influencing-people/01-01-welcome-8TWsi (Accessed: 2 April 2017)

subtitles more important. The use of the video and transcripts in this research was approved by Coursera.

5.1.4 Google Translate

The MT tool used for both the pilot and main experiment was Google Translate.[62] The motivation for doing so was that MOOC users who need free translation are likely to turn to free, online systems such as Google Translate, which boasts some 500 million monthly users and translates over 140 billion words a day across 100 language combinations (Turovsky, 2016). As Hu (2018) explains, the web page of Google Translate is always accessible in mainland China, although other Google sites are blocked. However, the mobile app of Google Translate was not launched in mainland China until March 2017, when it was updated to version 5.8. In short, Google Translate is available to Chinese users both on laptop and mobile phone now.

While customized engines have since been created for the MT of MOOC content (for example, in the TraMOOC project as outlined in Section 2.3.2), these engines were not sufficiently developed at the time the pilot and main experiment were conducted, and Google Neural MT outperformed competing systems in the initial tests that I ran. Therefore, it was decided to use the best available system at the time.

[62] https://translate.google.com/ (Accessed: 3 April 2017)

5.2 Pilot Study

Based on what has been discussed in Chapter 2 and the goal of this research (see Section 5.1), an initial methodology was developed to investigate viewers' reception of MT output (both raw and post-edited). For the consideration of the validity of this methodology, the researcher made the decision to run a pilot prior to the main experiment. It is good practice, especially in resource-intensive empirical research, to conduct a pilot in which the researcher tests all the research instruments (technology, questionnaire items, etc.) before conducting the larger-scale experiment. The pilot study for this research was conducted in April 2017.

Ethics approval was granted by the Research Ethics Committee of Dublin City University prior to the recruitment of participants for the pilot study. The researcher recruited four participants and divided them into two groups to compare the cases of users receiving raw MT subtitles with users receiving full PE subtitles. Participants were first asked to complete an online English test in order to identify their English proficiency level and a pre-task questionnaire relating to their use of MT. They were then asked to watch the selected MOOC video on a computer fitted with an SMI REDn Scientific eye-tracker and to complete a post-task questionnaire relating to their comprehension of the video content and attitudes towards the subtitles.

Section 5.1 has put forward the main research question and main hypothesis for this PhD research. However, because the human translated subtitles were only included in

the main experiment, the main research question and main hypothesis for the pilot study are thus slightly different and presented as follows:

> **Main research question:** is there a difference in reception between participants who are offered raw MT subtitles and those who are offered full PE subtitles?
>
> **Main hypothesis:** participants who are offered full PE subtitles will score higher on the reception metrics compared with those who are offered raw MT subtitles.

Based on the reception model and the main hypothesis, several sub-hypotheses were derived as follows:

Regarding "response":

Hypothesis 1a: Less demand on attention is required if the subtitles are easy to process.

This hypothesis was measured by the glance count in the subtitle AOI of each video. The higher the number of visits in AOI_SUB, the more attention participants were deemed to give to the subtitles.

Hypothesis 1b: Relatively more attention is allocated to the image area (AOI_IMA) when full PE subtitles are displayed than when raw MT subtitles are displayed.

As for this hypothesis, a comparison of glance count in the image AOI of each video was conducted.

Regarding "reaction":

Hypothesis 2: The level of comprehension is higher with full PE subtitles.

This hypothesis was measured by comprehension testing, which was part one (full score: 12) in the post-task questionnaire (see Appendix D1). Here, 1 was assigned to a correct answer, while 0 was assigned to a wrong answer.

Hypothesis 3: Average fixation duration is shorter when full PE subtitles are displayed.

As indicated in Section 4.2.1, average fixation duration (or mean fixation duration) is obtained by adding up the fixation duration in an AOI and dividing the result by the fixation count in the same AOI (Orrego-Carmona, 2015). As Orrego-Carmona (ibid., p.191) puts it, "Mean fixation duration is commonly regarded as an indicator of cognitive effort… not only in studies within Translation Studies but also in reading studies."

Regarding "repercussion":

Hypothesis 4: Attitudes are more positive among participants shown full PE subtitles than those shown raw MT subtitles.

This hypothesis was measured by 14 attitude questions (see Appendix D1), which constituted the second part of the post-task questionnaire. The questions were based

on a five-point Likert scale. The data collected was coded into numerical values ranging from 1 for “strongly disagree” to 5 for “strongly agree”. The higher the score, the better the attitude to MT.

5.2.1 Participant overview (Pilot)

According to the research design, the ideal participants would be Chinese undergraduates with a low English proficiency level since it is expected that these might be typical users of MOOCs in English and in need of translated content, but unable to pay for professional translation services. However, since the pilot was conducted in a university located in an English-speaking country, it was difficult to find participants that could completely meet the needs. Besides, the purpose of the pilot was mainly to test the methodology to be implemented in a larger-scale experiment later. Four Chinese participants were selected for the pilot study: a year-one PhD student (female, 27), a final-year PhD student (male, 25), a post-doc (male, 33), and a final year undergraduate (male, 22). All of them came from a science background. Three participants had experience of learning via MOOCs, but none of them was familiar with the MOOC selected for the pilot. The online English proficiency test[63] for identifying their English proficiency levels follows the Common European Framework of Reference for Languages. [64] Three participants answered 20 out of 25 questions correctly, corresponding to level B2 (vantage or upper intermediate), and one participant answered 15 questions correctly, corresponding to level B1 (threshold or intermediate).

[63] http://www.cambridgeenglish.org/test-your-english/general-english/ (Accessed: 3 April 2017)
[64] https://www.coe.int/en/web/common-european-framework-reference-languages/ (Accessed: 3 April 2017)

5.2.2 Subtitles

The researcher of this thesis is an English specialist who has completed a three-year professional translator training programme and has two years' practical experience translating English subtitles to Chinese for a fansub group in China. For the convenience of the study, the full post-editing of the raw machine translated subtitles was done by the researcher according to the post-editing guidelines issued by TAUS (2016a).

The number of subtitles for both videos (i.e. with/without post-editing) is 114. In order to provide a reference translation for calculating the BLEU score, the automatic translation evaluation of the raw MT output, a human translated version of the subtitles for the video was produced by a Chinese person who is a professional high school English teacher. By using the online Tilde BLEU score calculator [65] and taking the human translated subtitles as a reference, the BLEU score for the raw MT subtitles was calculated and found to be at 58.63%[66]. According to Google's interpretation,[67] if the BLEU score is between 50 and 60, it indicates MT provides "very high quality, adequate, and fluent translations". Note that for calculating BLEU, both raw MT and HT subtitles were adjusted to reach the same number (110). For alignment, the subtitles shown on the video were still the original ones (114 for both raw MT and PE subtitles). Calculating by the use of the tercom software (Snover et al., 2006), the HTER for post-editing was

[65] https://www.letsmt.eu/Bleu.aspx (Accessed: 3 April 2017)
[66] For Bleu scores, the high, the better (Vinyals et al., 2015).
[67] https://cloud.google.com/translate/automl/docs/evaluate (Accessed: 3 April 2017)

19.69%, indicating that the quality of the raw MT subtitles was relatively good since not much post-editing was required.

As for the subtitling procedure, Aegisub[68] (version: 3.2.2) was used to convert the original English subtitles into Chinese subtitles. FormatFactory (version: 4.1.0),[69] a multifunctional media converter, was used to add Chinese subtitles to the video. When subtitling the video, the norm that was adhered to was the Simplified Chinese (PRC) Timed Text Style Guide by Netflix.[70] According to this guide, the character limit is 16 characters per line, but this can be increased to 18 characters per line as needed.

5.2.3 Questionnaires

The researcher distributed two types of questionnaires for the pilot. One is the pre-task questionnaire, which was used to collect information about the participants' backgrounds and prior behaviours, such as their consumption of MOOCs and MT, exposure to subtitled content, and opinions regarding the quality of subtitles and MT. Note that all pre- and post-task questionnaires were administered in Chinese and translated into English in this thesis by the researcher. The reason why the questionnaires were administered in Chinese is that the participants were supposed to have a low level of English proficiency, and correct answers could only be obtained if the questionnaire was designed in Chinese. The pre-task questionnaire is presented below.

[68] http://www.aegisub.org (Accessed: 20 March 2017)
[69] http://www.pcgeshi.com (Accessed: 20 March 2017)
[70] https://backlothelp.netflix.com/hc/en-us/articles/215986007-Simplified-Chinese-PRC-Timed-Text-Style-Guide (Accessed: 20 March 2017)

1) Are you a native speaker of Chinese?
    - Yes
    - No

2) Age:

3) Sex:
    - Male
    - Female

4) What year are you in at university?

    _____

5) What is your major?

    _____

6) Do you have experience in using machine translation applications (Google Translate/ Baidu Translate etc.)?
    - Yes
    - No

(If you replied 'yes', please go to question 7. If you replied 'no', please go to question 10.)

7) When did you start to use machine translation?

    _____

8) How frequently do you use it?

    _____

9) What do you think of the quality of machine translation?

    _____

10) Please indicate the reason why you don't use machine translation.

    _____

11) Do you have experience in watching movies or other videos with machine translated subtitles?
    - Yes
    - No
    - Uncertain

12) Do you have experience in learning via MOOCs?
- Yes
- No

(If you replied 'yes', please go to question 13. If you replied 'no', you can click on the "Submit" and complete the questionnaire.)

13) Do you have experience in learning via MOOCs with machine translated subtitles?
- Yes
- No
- Uncertain

14) Have you studied the MOOC entitled "Sit Less, Get Active" before?
- Yes
- No

15) Do you believe the machine translated subtitles will adequately transfer the meaning of the source text?
- Yes
- No

A question in a questionnaire has to have a reason and that should be linked to collecting data that will serve to answer a specific research question. Using the data types mentioned in Section 4.2.2.2, the type of data collected through this pre-task questionnaire should be facts and background information. This questionnaire aims to collect all participants' factual and background information, and filter out those that do not meet our needs. Based on the question types summarized by Matthews and Ross (2010) in Section 4.2.2.3, we can see that Questions 1), 6), 11), 12), 13), 14), and 15) are Yes/No questions that are used for distinguishing between different groups of participants. Among them, Questions 6), 12) and 14) are also filter questions, the researcher can thus filter the participants based on the two questions on whether they have experience in using MT/MOOCs or not. Although Question 14) belongs to the Yes/No question type, the effect is to filter out those that have accessed the MOOC

selected for this research, because if they had accessed it before, it would affect their answers to the comprehension testing part of the post-task questionnaire and the overall results. Questions 2) to 5) aim to collect demographic information of participants. Questions 7) and 8) belong to the type “Which category? How many? How much?” and aim to ask participants about themselves and their experience with MT. Questions 9) and 10) are open questions for collecting participants’ opinions towards MT, Question 15) also has the same effects.

The second was the post-task questionnaire with two parts: comprehension testing and attitude survey.

a) Comprehension testing

This section was designed to test whether there was any difference in the comprehension of the MOOC between the two groups, which is linked to “reaction” in Gambier’s model. Multiple-choice questions were presented here, and each question offered four possible answers, with only one being correct. Each correct answer had a one-point score attached to it while zero points were scored for an incorrect option. All the questions were related to the key information provided in the subtitles throughout the clip. Because the clip is less than seven minutes long, it was not easy to design adequate questions. When designing the questions, it was kept in mind that the questions should not be too simple or too difficult, they must relate to the clip content

and cannot be inferred from common sense. In the end, 12 questions were designed for this part as below:

1) What counts as physical activity?
   A. Dancing
   B. Walking the dog
   C. Climbing the stairs
   D. Housework or gardening
   E. All of the above

2) Which example below counts as a muscle strengthening activity?
   A. Running
   B. Yoga
   C. Lifting weights at the gym
   D. Gardening

3) Which of the following would NOT be considered as a sedentary behaviour?
   A. Driving in your car
   B. Sitting at a desk working
   C. Standing on the train to work
   D. Reading a book on the sofa

4) Which of the following is not active transport?
   A. Walking
   B. Cycling
   C. Driving a car
   D. None of the above

5) Which of the following is not caused by sedentary behaviour?
   A. Obesity
   B. Type 2 diabetes
   C. Eye disease
   D. Some cancers
   E. Premature mortality

6) Which of the following is not exercise?
   A. Training to take part in a marathon
   B. Walking along the pavement
   C. Going to gym for strength training
   D. Going to a yoga class

7) According to physical activity guidelines, children should do 60 minutes of activity every day, that's ____
   A. Running
   B. Jumping
   C. Both of the above

8) According to physical activity guidelines, adults should ____
   A. have moderate activity at least 150 minutes every week
   B. have strength building activities
   C. minimize the time they spend sitting
   D. All of the above

9) True or false: Doing moderate activity is enough to meet the physical activity guidelines.
   A. True
   B. False

10) True or false: If you want to do physical activity only 10 minutes a day, 5 minutes a day, that's not going to be helpful.
   A. True
   B. False

11) True or false: Children get more active as they age.
   A. True
   B. False

12) True or False: Physical activity is a very specific subset of exercise.
   A. True
   B. False

This part contains multiple-choice questions (Questions 1) to 8)) and true-or-false questions (Questions 9) to 12)), which belong to the "fixed format questions" according to Jensen and Mostrom (2012, p.49). As they put it, true-or-false questions have the advantages of being "easy to write" and they "can be answered quickly" (ibid.), while the disadvantages are participants have a high chance (50%) of getting it right, and it is not easy to avoid ambiguity. The latter can be remedied by "focusing on the accuracy of key names, actions, or concepts rather than on obscure points" (ibid.), which is exactly what the researcher did in Questions 9) to 12). Regarding multiple-choice questions,

they are used here for measuring comprehension and knowledge. The advantages are they are easy to grade and process by computer. Jensen and Mostrom (ibid.) provide their suggestions for designing multiple-choice questions as follows: avoid triviality and irrelevance when writing distractors; do not overuse “all/none of the above”; refrain from using words such as “always”, “never”, “all”, or “none”; avoid keeping the correct options in certain places (e.g.: in the middle, such as position B or C) all the time; avoid trick and complicated questions such as negatively worded options that test semantics rather than knowledge, etc. The researcher accepted these suggestions when designing the multiple-choice questions in the comprehension testing part. It can be assured that all the questions are relevant to the video clip, none of them is too complicated or tricky. There is no use of words like “always”, “never”, “all”, or “none” in the questions. “All/Both/None of the above” has been used four times (Questions 1), 4), 7) and 8)). Their use in these questions is not due to a lack of inspiration, but rather because they are logically possible answers which force participants to make finer distinctions based on the content of the video.

b) Attitude survey

This section of the questionnaire was linked to “repercussion” in Gambier’s model. It was designed to record participants’ self-evaluation of their comprehension, processing effort, reading ease and perceived enjoyment, and to find out if the reception level of raw machine translated subtitles differed from that of full post-edited subtitles. Questions were answered by both groups, for example, do they understand all subtitles?

Are the subtitles easy to understand? Do they enjoy reading subtitles? 14 statements were designed for the attitude survey as follows:

1) The subtitles allow me to fully understand the contents of the MOOC.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

2) The subtitles are useful to me.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

3) The subtitles are easy to understand.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

4) Interacting with the subtitles does not require a lot of my mental effort.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

5) I would find it easy to get the information I need from subtitles.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

6) The subtitles are clear and understandable.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

7) I enjoy reading the subtitles.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

8) I am satisfied with the subtitles.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

9) If I have a chance, I would use machine translation to translate English subtitles in the future, because I know it will do a good job.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

10) I would recommend machine translation to my friends if they need to translate subtitles.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

I could comprehend the subtitles…

11) If there was no one around to tell me what to do as I go.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

12) If I could call someone for help if I got stuck.

☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

13) If I had a lot of time.

☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

14) If I had just the built-in help facility (e.g.: online dictionary) for assistance.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

As mentioned in Section 4.2.2.2, internal reliability is of critical importance for Likert scales, and one method of testing it is by calculating the Cronbach's alpha. Due to the small sample of the pilot, the consistency check for the attitude survey was only conducted for the main experiment, see Section 5.3.2 for more information. As mentioned in Section 4.2.2.3.1, the statements in the attitude survey were designed based on TAM. Concepts in TAM such as perceived usefulness, perceived ease of use, perceived enjoyment, perceived quality, intention to accept the technology, and compensation, have been borrowed for designing the statements. In this attitude survey, all the statements were designed following the rules mentioned in Section 4.2.2.3.1. Statements 1) and 2) were designed for studying participants' perceived usefulness of the subtitles, Statements 3) to 6) were designed for studying participants' perceived ease of use of the subtitles, Statement 7) was designed for studying participants' perceived enjoyment of reading the subtitles, Statement 8) was designed for studying participants' perceived quality of the subtitles, Statement 9) and 10) were designed for studying participants' intention to accept machine translated subtitles, Statement 11) to 14) were designed for studying the degree to which participants believe that they have the ability to comprehend machine translated subtitles using additional inputs.

5.2.4 Pilot study findings

This section includes a discussion on the findings of both the questionnaires and eye-tracking experiment.

5.2.4.1 MT usage

Given the widespread use of MT nowadays, it is not surprising that in the pre-task questionnaire (see Appendix C1), all participants claimed they had experience of using MT. More specifically, three of them used MT every day, and one of them used it three times a week. One participant commented that "there is no reason for not using MT, it brings a lot of conveniences to our life, though the quality is just so-so". Three participants rated the quality of MT as "average", one participant answered that "MT for long sentences is not accurate". Regarding the experience of watching videos with machine translated subtitles, two participants answered they had prior experience, the other two were not sure, possibly because many Chinese people watch foreign videos online, but sometimes there is no indication whether the subtitles are translated by a human or by an MT system. In response to the question "Do you believe the machine translated subtitles will adequately transfer the meaning of the source text?" two of the four participants answered positively, while the other two answered negatively.[71]

[71] The original answers are in Chinese and translated into English by the researcher

5.2.4.2 Data analysis

Two areas of interest (AOIs) were defined for analysis on the video: AOI_SUB (a small rectangle surrounding the subtitles) and AOI_IMA (a large rectangle surrounding the image area). Figure 5.1 shows an example of the two AOIs.

Figure 5.1 Screenshot of Two AOIs (the orange rectangle is AOI_IMA and the blue one is AOI_SUB)

Regarding "response":

Hypothesis 1a: Less demand on attention is required if the subtitles are easy to process.

The glance count in AOI_SUB for the group who were offered raw machine translated subtitles (hereafter abbreviated as Group RAW), namely P01 and P02, was hypothesized

to be higher than the group who were offered full post-edited subtitles (hereafter abbreviated as Group PE), namely P03 and P04 (see Table 5.2).[72]

Table 5.2 Glance Count in Subtitle AOI of Each Group

		AOI_SUB
Group RAW	**P01**	130
	P02	95
	Mean	112.5
Group PE	**P03**	94
	P04	96
	Mean	95

Table 5.2 shows the related data exported from the SMI eye-tracker for the pilot. According to the means, the glance count in AOI_SUB of Group RAW was higher (112.5) than for Group PE (95), which corresponded to the hypothesis. However, given the size of the group, no tests for statistical significance in differences were conducted.

Hypothesis 1b: Relatively more attention is allocated to the image area (AOI_IMA) when full PE subtitles are displayed than when raw MT subtitles are displayed.

[72] Note that the eye-tracking data of one participant was compromised because the eye-tracker lost connection with her eye movements for the last minute of the video. As a result, data from only the first 5'30'' (330,000ms) for all participants was used for data analysis for this pilot study.

Table 5.3 Glance Count in Image AOI of Each Group

		AOI_IMA
Group RAW	**P01**	129
	P02	106
	Mean	117.5
Group PE	**P03**	93
	P04	94
	Mean	93.5

According to the means in Table 5.3, Group RAW was higher (117.5) than Group PE (93.5). It is not possible to support Hypothesis 1b with this result.

Table 5.4 Fixation Count in Both AOIs of Each Group

		AOI_IMA	AOI_SUB
Group RAW	**P01**	335	634
	P02	434	363
	Mean	384.5	498.5
Group PE	**P03**	479	413
	P04	264	387
	Mean	371.5	400

Table 5.5 Glance Duration [s] in Both AOIs of Each Group

		AOI_IMA	AOI_SUB
Group RAW	P01	142.84	176.30
	P02	187.49	96.26
	Mean	165.17	136.28
Group PE	P03	190.14	93.26
	P04	173.18	135.00
	Mean	181.66	114.13

Nevertheless, Table 5.4 shows that for fixation count, Group RAW had a higher number in both AOIs than Group PE. For glance duration, Table 5.5 shows that Group PE's average was higher at a mean of 181.66 s vs. 165.17 s for Group RAW in AOI_IMA, and was lower than Group RAW in AOI_SUB. This means that while Group PE had a lower fixation count, they spent longer on average in the AOI_IMA than the other group. Considering Group RAW also had a higher glance count and glance duration in AOI_SUB, we can infer that the raw MT subtitles might be of lower quality than the full PE subtitles, so that Group RAW had to keep visiting the image area for help, while the image could not provide enough information for them to comprehend the MOOC content, they also had to keep visiting subtitles and spent more time processing them because of their low quality. However, according to the BLEU score and HTER score mentioned in Section 5.2.2, the quality of raw MT subtitles is not low. Therefore, what the eye-tracking data shows is contradictory to what the BLEU and HTER indicate. This result may be caused by the small size of the two groups.

Regarding "reaction":

Hypothesis 2: The level of comprehension is higher with full PE subtitles.

Table 5.6 Comprehension Testing Results of Each Group

		Score
Group RAW	P01	12
	P02	11
Group PE	P03	10
	P04	9

The results are presented in Table 5.6. and show that Group RAW scored higher than Group PE, which was contrary to the hypothesis. In addition, the English proficiency level of P01 was the lowest (15 correct out of 25); however, he achieved the highest score for comprehension testing. Again, this unexpected result may be due to the small number of participants, however, and would be addressed in the larger-scale experiment.

Hypothesis 3: Average fixation duration is shorter when full PE subtitles are displayed.

Table 5.7 Average Fixation Duration [ms] in Both AOIs of Each Group

		AOI_IMA	AOI_SUB
Group RAW	P01	388	251
	P02	395	237
	Mean	391.5	244
Group PE	P03	351	178
	P04	625	328
	Mean	488	253

According to Table 5.7, this hypothesis was not supported by the results, because compared to Group PE, Group RAW had lower average fixation duration in both AOIs.

Regarding "repercussion":

Hypothesis 4: Attitudes are more positive among participants shown full PE subtitles than those shown raw MT subtitles.

Table 5.8 Attitude Survey Results of Each Group

		Score
Group RAW	P01	65
	P02	61
Group PE	P03	60
	P04	57

The results are presented in Table 5.8. In contrast to the hypothesis, it shows that Group RAW scored higher than Group PE, which means the attitude of Group RAW towards MT was better than Group PE. Again, this result may be simply due to the very small sample size.

5.2.5 Reflections

The hypotheses were not all supported by the results, but one of the reasons could be the small number in the sample. Early indications were that the use of raw MT subtitles does not negatively affect comprehension, attitude or cognitive processing when compared with the presentation of full PE subtitles. However, we also recall here that the main objective of the pilot experiment was to test the methodology. As for the main experiment, more than 60 participants would be recruited, and ANOVA (analysis of variance) and t-tests would be used for statistical analysis in order to show the likelihood that the results occurred by chance.

Regarding the post-task questionnaire, Participant 01 commented that he could answer some questions in part one without watching the video, which indicated that the questions needed to be modified, to make it more likely that participants' answers to comprehension questions would be based on information they had managed to glean (or not) from the video subtitles, rather than on their prior knowledge. Participant 01's comment is also a reminder that a high comprehension score does not necessarily mean that subtitles are of good quality. The converse is also true: a low comprehension score does not necessarily mean that the subtitles are of poor quality; after watching a MOOC

video, viewers might not remember all the content well, and they might not be focusing on the video all the time when they watched it. Put briefly, memory, concentration, and even the IQ of the participant, could all be possible independent variables that had an effect on the results of part one. While the latter problem is difficult to solve, two possible adjustments to the research design were considered to deal with the first problem: either change the MOOC video to a more technical one; or refine the questionnaire with more appropriate questions. The second option was taken, that is, the post-task questionnaire would be altered to make it less possible to answer the questions using common sense.

As for part two, the attitude survey, again, a possible reason for the unexpected results could be that the quality of raw MT subtitles and full PE subtitles was not significantly different, since the HTER was only 19.69%. According to the pre-task questionnaire, generally speaking, all participants had a positive attitude towards MT. The BLEU score for raw MT subtitles was good. Hence, there is a reason to believe that participants' attitudes might not have changed much after watching the video. It is worth emphasizing that the pilot tested the technology and the questionnaire, and that it was run in only two conditions, as this was all the researcher needed to do such testing. The results of the pilot were never going to be of interest given the sample size.

It was decided to include human translated subtitles for the video as an independent variable to determine whether there are differences between machine translated and human translated content. The human translated subtitles were the ones that were used as reference translation for calculating the BLEU score in Section 5.2.2. In the pilot

study, it was found that the segmentation of subtitles might be a confounding variable. Hence, the subtitles of the three types were readjusted in order to better follow the Netflix guidelines. After that, the number of lines of subtitles was 135, 138 and 141, for raw MT, PE and HT respectively. The difference in the number of subtitles is because for the same source sentence, MT, PE and HT can have different outputs, and one can be shorter/longer than another one. In order to follow the character limit (18 maximum), for instance, one line of the raw machine-translated version was cut into two lines, but the corresponding line of the post-edited version was not. See the example below:

> Source:
> And this seems to be even in individuals who are achieving physical activity recommendation.
>
> Raw MT (18 characters in total, one line):
> 这似乎甚至在实现体育活动推荐的个人中
>
> PE (21 characters in total, two lines):
> 甚至那些实现身体活动的人
> 也好像有这样的情况
>
> HT (22 characters in total, two lines):
> 并且甚至那些实现身体活动的人
> 似乎也有这个情况

In addition, based on the results of the pilot, the researcher made some revisions (see below) to the first part of the post-task questionnaire, which is the comprehension testing part, while the attitude survey remained the same without edits; thus the post-task questionnaire used for the main experiment was the revised version (see Appendix E1).

Question 3 in the pilot was removed due to the fact that it was too easy to answer:

> Which of the following would NOT be considered as a sedentary behaviour?
> A. Driving in your car
> B. Sitting at a desk working
> C. Standing on the train to work
> D. Reading a book on the sofa

Two new questions were added to increase difficulty:

Which of the following benefit of physical activity is not mentioned in the video?

A. It can help reduce our risk of multiple diseases.
B. It can help us to maintain healthy weight.
C. It can help us to improve the quality of our life.
D. None of the above.

Which of the following statement about physical activity is not right?

A. It is any movement that uses energy.
B. It is not structured.
C. It is different from exercise.
D. It is pursued for fitness benefits.

5.3 Main experiment

The main experiment was conducted in two mid-sized universities (Anhui Normal University and Tongling University) located in Anhui province, China, during October 2017. Compared to the pilot, the main experiment was based on a larger sample size. In total 66 participants who were not from the School of Foreign Languages were recruited as participants for the main experiment. Prior to recruitment, ethics approval was granted from the Research Ethics Committee of Dublin City University, in line with the requirements of all three universities.

5.3.1 Similarities and differences with the pilot

The pre-task questionnaire, MOOC video, eye tracker, the MT system, and the English proficiency test adopted in the main experiment were the same as those used for the pilot study. The major difference between the pilot and main experiment arises in the types of subtitles being studied; that is, alongside raw machine translated subtitles and post-edited subtitles, human translated subtitles were added to the main experiment. Other differences lie in the segmentation of subtitles and the post-task questionnaire, as mentioned in Section 5.2.5.

5.3.2 Conduct of the experiment

Participants were divided into three groups to compare the cases of users receiving raw MT subtitles with users receiving full PE or HT subtitles. The procedure of the main experiment bears a close resemblance to the pilot. There are three steps as follows in the main experiment:

Step 1: Pre-task questionnaire and English proficiency test

The pre-task questionnaire was conducted through 问卷星 (www.wjx.cn), a Chinese website providing an online survey service. Participants who either had already watched the selected MOOC video or did not meet the requirement for English proficiency would be removed from the experiment.

Step 2: Eye-tracking

Participants were divided into three groups: Group HT, Group PE and Group MT. They were asked to watch a MOOC video on a laptop with an eye-tracker.

Step 3: Post-task questionnaire

The post-task questionnaire was also conducted through 问卷星. The questionnaire included two sections: comprehension testing and attitude survey. The list of questions can be found in Appendix E1.

Only Step 1 was conducted without the supervision of the researcher. Most participants completed Step 1 using their mobile phones in their dorm or other places. Steps 2 and 3 were carried out in an independent room on campus (an office free from disturbances) under supervision of the researcher, one participant at a time. In Anhui Normal University, 48 undergraduates were recruited as participants, of whom 45 completed the full experiment. Due to unknown reasons, the eye-tracking failed to track three of the participants to a level that was of acceptable quality. In Tongling University, 18 undergraduates were recruited as participants, of whom 16 completed the full experiment, and two participants only completed the first step. In summary, 66 participants from two Chinese universities were recruited for the main experiment, 61 of whom completed the full experiment. To compare the cases of users receiving subtitles in the three different conditions, the 61 participants were randomly assigned to three groups: 24 participants were in Group PE – this group would watch the video

with full PE subtitles; 22 participants were in Group RAW – this group would watch the video with raw MT subtitles; and 15 participants were in Group HT – which would be provided with the video with HT subtitles. The consistency check of the 61 post-task questionnaires (the attitude survey part) was conducted by using the automatic built-in service on the questionnaire website. The result of the Cronbach's alpha was 0.854, indicating a high reliability of the questionnaire data (see Section 4.2.2).

5.3.3 TQA

As mentioned earlier, TQA and frequency analysis were used as two supplementary methods in the main experiment only, which were adopted post-hoc to see if further explanations for the reception metrics could be established.

For TQA, the researcher adopted the MQM Core (Section 3.4.2) and customized it in order to suit this research. The fact that professional translators are not usually involved is a negative feature of evaluations, hence it was planned to recruit professional translators (certified or registered in a translation agency) and conduct the TQA for this research on a group-based level (in other words, at an initially individual level where the scores are then grouped and averaged). In this way, several evaluators can be involved and the averaging of their scores will balance the personal biases. In order to balance limited resources with the need for a high number of evaluators, it was decided to recruit 10-20 evaluators for this study, and all of them would evaluate both the HT subtitles and PE subtitles. Because the research was carried out in Dublin where not many Chinese professional translators could be reached, the original plan of recruiting

professional translators was difficult to carry out and the researcher had to compromise. As a result, some native Chinese students who major in translation studies (mostly) in Xi'an Jiaotong-Liverpool University, an international joint university (between the University of Liverpool and Xi'an Jiaotong University) based in Suzhou, China, were contacted by the researcher. Rather than professional translators, it would be more appropriate to call the students translation novices or trainees. The researcher liaised closely with 12 students and all of them agreed to participate in the TQA eventually. The application for research ethics approval was approved by the DCU Research Ethics Committee (REC).[73] Before processing the data, each participant was assigned a number and those numbers would be used throughout the research project to store, process and present the data. To measure the English proficiency level of the participants, the researcher adopted the Cambridge online English proficiency test (same as the pilot), and only those who achieved a B2 level or above (corresponding to achieving 18 points or more in the test) were considered eligible evaluators. In the end, ten of the students passed the English test. All of them evaluated both the HT subtitles and PE subtitles.

Figure 5.2 shows a screenshot of the work interface of the evaluators, which was designed by the researcher. In Figure 5.2, Chinese words in the header row are glossed in red in English, and the original source-language sentence and its two translations are also labelled in red (in rows 2, 3 and 4). These English glosses and labels are provided for the convenience of the reader of this thesis. The Chinese evaluators saw only Chinese on their interface. On the left side, it can be seen that the English source text comes first,

[73] REC Reference Number: DCUREC/2017/053

and two Chinese translations named “Translation 1” and “Translation 2” follow. In this way, the evaluator would not know which one was PE or HT. The reason for not identifying the source of the two translations is because evaluators may presume the HT output is better than PE output, thus leading to biased and unreliable quality assessment. Also, the two types of subtitles are in randomized order to prevent participants seeing a pattern during the evaluation.

	B	C	Error 1	Error type	Severity	Error 2	Error type	Severity
			错误1	错误类型	严重度	错误2	错误类型	严重度
		Physical activity is any movement that uses energy. Source						
	翻译1	身体活动是任何使用能量的活动。 Translation 1	使用能量	术语	轻微			
√	翻译2	身体活动是消耗能量的任何运动。 Translation 2						
		And being physically active is important for people of all ages.						
	翻译1	保持身体的积极状态对所有年龄的人都重要。	积极状态	语法	轻微			
√	翻译2	保持身体活跃对所有年龄段的人都很重要。						
		Physical activity can help reduce our risk for multiple diseases such as coronary heart disease, type 2 diabetes, some types of cancer.						
√	翻译1	身体活动可以帮助我们降低多种疾病的风险，如冠心病，2型糖尿病，某些类型的癌症。						
	翻译2	身体活动能帮助减少患多种疾病的风险，比如冠心病，2型糖尿病，某几种癌症。	某几种	过度直译	轻微			
		It can also improve our bone health, help us to maintain healthy weight, and also improve our overall quality of life.						
	翻译1	它还能改善我们的骨质健康，帮助我们维持健康的体重，并且还能提升我们整个生活骨质健康	术语	轻微		维持健康体	语法	轻微
√	翻译2	它也可以改善我们的骨骼健康，帮助我们保持健康的体重，并提高我们的整体生活质量。						

Figure 5.2 Screenshot of the QA work interface

The translation quality categories used for this TQA are adapted from the MQM Core (see Figure 3.2), because MQM is one of the latest initiatives that attempts to standardize TQA that provides a systematic framework for describing and defining metrics for evaluating the quality of translated text and identifying specific issues within those texts (see Section 3.3.3.2).

Table 5.9 Translation quality categories in this study

Accuracy	Addition
	Mistranslation
	Overly-literal
	Omission
Fluency	Grammar
	Spelling
Style	Style
Locale convention	Locale-convention
Terminology	Terminology
Others	Others

Table 5.9 shows the customized translation quality categories and sub-categories used in the evaluation component of this study. The definitions of the categories are cited from MQM as below. However, the definitions themselves often seem tautological or circular, or they are often just paraphrases of the term used to label the category (e.g.: overly-literal, spelling, and style). Besides, the definitions in MQM may not be entirely appropriate in the context of MT. Despite these issues, on balance, MQM was judged to be a suitable metric and its categories and definitions were adopted, with some customisation for the context.

- Accuracy: Accuracy issues address the relationship of the target text to the source text and can be assessed only by considering this relationship. Changes in intended meaning, addition and omission of content, and similar issues are considered under this category.
    - Addition: The target text includes text not present in the source.
    - Mistranslation: The target content does not accurately represent the source text content.
    - Overly-literal: The translation is overly literal (e.g.: A Hungarian text contains the phrase *Tele van a hocipőd?*, which has been translated

as “Are your snow boots full?” rather than with the idiomatic meaning of “Feeling overwhelmed?”).
- Omission: Content is missing from the translation that is present in the source.

- Fluency: Fluency includes those issues about the linguistic “well-formedness” of the text that can be assessed without regard to whether the text is a translation or not.

    - Grammar: Issues related to the grammar or syntax of the text, other than spelling and orthography.
    - Spelling: Issues related to spelling of words (e.g.: The German word *Zustellung* is spelled *Zustetlugn*).

- Style: The text has stylistic problems (e.g.: The translation of a light-hearted and humorous advertising campaign is in a serious and “heavy” style even though specifications said it should match the style of the source text).

- Locale convention: Issues in locale convention relate to the formal compliance of content with locale-specific conventions, such as use of number formats. If content is otherwise correctly translated and fluent but violates specific locale expectations (as defined in the translation specifications), it is addressed in this dimension. This dimension does not cover issues related to whether the content itself is appropriate for the locale (these issues are covered under verity).

- Terminology: Terminology issues relate to the use of domain- or organization-specific terminology (i.e., the use of words to relate to specific concepts not considered part of general language).

- Others: Issues that do not belong to any of the issue types listed above (defined by the researcher).

Compared to MQM Core (Figure 3.2), we can see that for this research, in regard to the translation quality categories, it has been decided to remove the issue types “untranslated” under “accuracy”, and “grammatical register”, “typography” and “unintelligible” under “fluency” from the MQM Core, because the researcher had scrutinized all the subtitles before conducting TQA with the evaluators, and it was found that there were no errors of “untranslated”, “typography” and “unintelligible”, for the subtitles were all translated into Chinese language in plain format without typos. Also,

the MOOC video consists of interviews and voice-over about sports, and the language used by speakers is informal and mostly non-specialized. The translated subtitles are consistent with the source text and do not have any incorrect use of grammatical markers of formality, that is why "grammatical register" was removed.

"Inconsistency" was removed because in our quality assessment, the two types of subtitles are in randomized order. "Style" was retained because our subtitles are idiomatic and informal, and the translation should adhere to the style of the original text. "Locale convention" was included because due to the difference in punctuation between Chinese and English (the MOOC video in this study was made in UK), the locale convention is necessary for evaluation. For example, when separating a series of the same kind of things, a comma is used in English while a slight-pause mark "、" is used in Chinese. The use of this special symbol is similar to "/" in English, see Example 5.1 (this example is taken from the subtitles) as follows:

Example 5.1:

English: So, and we know that higher levels of sitting are linked with obesity and an increased likelihood of developing type 2 diabetes and cardiovascular disease and some cancers, and even premature mortality.

Chinese: 所以，我们知道坐久了会产生肥胖，增加发生 2 型糖尿病、心血管疾病和某些癌症，甚至过早死亡的可能性。

"Verity" issues relate to "the suitability of content for the target locale and audience". The content of the MOOC video is suitable for the Chinese locale and audience, thus "verity" was removed from the categories because there was no evidence of such an issue emerging. "Terminology" was included because it relates to "the use of domain-

or organization-specific terminology". The MOOC video used for this study is about sports and health, and domain terminology is included (though not much), thus this issue is worthy of inclusion. For example, "physical activity" is a specialized term in the area of health, however, it is not that 'specialized' and is easy for people to understand, the same is true of other terms in the video. Hence, even though the MOOC video contains some domain terminology, this does not contradict my earlier suggestion that the MOOC selected was general and not specialized (see Section 5.1.3), because the subject is so close to people's daily life. "Design" is not included because it is "related to the physical presentation of text, typically in a 'rich text' or 'markup' environment".[74] The evaluators were seeing the subtitles just on the TQA work interface on their computers, and the subtitles used for TQA were all in plain text, thus there is no need to take this issue into account. It has to be noted that the issue type "overly-literal" is a sub-issue of "mistranslation" in the MQM model, however the two issues can usually be properly differentiated and the definition of "mistranslation" in MQM would not apply to a "literal translation". Also, considering MT has long been criticized as being literal (Coldewey, 2018; Hofstadter, 2018), it has been decided to highlight the issue "overly-literal" and treat it as parallel with the issue "mistranslation". The category "others" was added in order to provide the evaluators with an option were they to find an issue that does not belong to any of the listed issue types.

[74] Available at: http://www.qt21.eu/mqm-definition/definition-2015-12-30.html (Accessed: 6 December 2019)

When conducting the quality assessment for the two types of subtitles, the severity levels were also taken from the MQM metric (MQM, 2015). The scoring algorithm below is cited from MQM:

$$TQ = 100 – TP + SP$$

where:

TQ = quality score

The overall rating of quality

TP = penalties for the target content

Sum of all weighted penalty points assigned to the target text

SP = penalties for the source content

Sum of all weighted penalty points assigned to the source text

All penalties are relative to the sample size (in words) and are calculated as follows (assuming default weights and severity levels):

$$TP = \frac{\text{Issues}_{minor} + \text{Issues}_{major} \times \text{SeverityMultiplier}_{major} + \text{Issues}_{critical} \times \text{SeverityMultiplier}_{critical}}{\text{Word count (target)}}$$

where:

Issuesminor = Number of issues with a “minor” severity

Issuesmajor = Number of issues with a “major” severity

Issuescritical = Number of issues with a “critical” severity

It can be seen from the scoring algorithm that the quality of the source text is also considered in the MQM model, because if the source text has problems, it will undoubtedly affect its translation. However, after scrutinizing the source text, no issue was detected, thus it was not necessary to examine the source text in the TQA. Therefore, the source penalties are by definition 0 and do not count for or against the translation’s quality score. The scoring algorithm for the TQA is therefore as follows:

$$TQ = 100 - TP$$

5.3.4 Frequency analysis

As mentioned in Section 4.2.4, because the frequency analysis is a supplementary method, it was decided to use the ten sentences for which the HT and the MT differ most. Hence, only the ten sentences with the lowest BLEU scores were analysed. The strings that were chosen for analysis were those for which an English word or phrase was translated differently in the human translation and post-edited MT. The corpora used for analysis are SogouT[75] (out-of-domain) and OpenSubtitles[76] (in-domain). SogouT is a monolingual (Chinese) corpus, consisting of pages crawled from all domains, and containing 5,463,795 sentences. Its data is collected from 130 million original web pages

[75] http://www.sogou.com/labs/resource/t.php (Accessed: 18 November 2018)
[76] http://opus.nlpl.eu/OpenSubtitles2018.php (Accessed: 18 November 2018)

of various types from the Internet. The corpus was mainly built by Sogou Labs, which is affiliated to the Sogou company, one of China's top three search engines. OpenSubtitles is a collection of parallel corpora of over 60 languages compiled from http://www.opensubtitles.org/, a large database of movie and TV subtitles. While as mentioned in Section 2.2.5.2, OpenSubtitles.org includes lots of fansubbing subtitles, which are of unknown quality (Müller and Volk, 2013), the corpora have been frequently used as datasets for training MT systems (e.g.: Banerjee et al., 2012; Wołk and Marasek, 2015; Sennrich et al., 2016; Toral and Way, 2018). For this research, the researcher used the Chinese data in their ZH-EN corpus, which contains 16,316,804 Chinese sentences.

Two tools were adopted for calculating the frequency occurrence of strings. First, Jieba[77] was used for the segmentation of subtitles. Jieba is a Python Chinese word segmentation module, popularly used by researchers for Chinese word segmentation (e.g.: Ye et al., 2015; Day and Lee, 2016; Zhang et al., 2018). After the segmentation of subtitles, NLTK FreqDist[78] was used for calculating the frequency occurrence of strings. NLTK FreqDist is also a Python module, used for encoding frequency distributions that count the number of occurrences of each outcome of an experiment. It is usually adopted by NLP (Natural Language Processing) researchers for analysing corpora and calculating the frequency of word/phrase (e.g.: Madnani, 2007; Kundu et al., 2015; Garg et al., 2016).

[77] https://github.com/fxsjy/jieba (Accessed: 5 June 2019)
[78] http://www.nltk.org/api/nltk.html?highlight=freqdist (Accessed: 5 June 2019)

5.4 Conclusions

In this chapter, the researcher has elaborated on the research design and its implementation, including a detailed discussion on the pilot study and an introduction of the main experiment. The main experiment bears resemblance to the pilot, but contains a larger sample and one more type of subtitles. In addition, a TQA and frequency analysis were conducted to better understand the results of the main experiment. From the next chapter, the researcher will make a comprehensive analysis of the main experiment.

Chapter 6 Findings and Data Analysis I - Questionnaires

6.1 Introduction

Chapters 6 and 7 present the data analysis for the main experiment carried out at Anhui Normal University and Tongling University in China during October, 2017. This chapter focuses on the data analysis of questionnaires and is structured as follows: Section 6.2 utilises descriptive statistics to present data from the pre-task questionnaire and English test, followed by a clarification of the main research question and hypotheses in Section 6.3. Section 6.4 discusses the data collected from the post-task questionnaire, divided into two sub-sections: comprehension testing and attitude survey.

6.2 Pre-task questionnaire and English test

As Step 1 of the main experiment, this part aims to collect information on the background and the English proficiency level of all the participants, so as to identify the suitable ones for Step 2 (eye-tracking) and Step 3 (post-task questionnaire) as mentioned in Section 5.3.2.

6.2.1 Pre-task questionnaire

As mentioned in Section 5.3.2, the pre-task questionnaire was conducted through 问卷星(www.wjx.cn), a Chinese website providing an online survey service. There were 72 returned questionnaires in total, because six participants submitted the questionnaire twice due to network issues. They agreed with the researcher that the second submission would be considered as the definitive version. Therefore, there were 66 valid questionnaires. However, six participants forgot to complete the pre-task questionnaire first and moved on to Steps 2 and 3 directly. Though they completed Step 1 in the end, inverting the order of the experimental steps may have affected their answers to some questions, more specifically, Questions 9, 11, 13, and 15 in the pre-task questionnaire. Hence, the data for these six participants (P09, P10, P12, P25, P28, and P51) for those questions will be highlighted in the footnotes. According to the responses to the pre-task questionnaire, none of the participants had participated in the MOOC being used for Step 2 before.

It has to be noted that although 66 participants completed the pre-task questionnaire, not all of them answered all questions. In addition, due to the design of the questionnaire, some questions could be skipped based on participants' answers to previous questions. Therefore, the number of respondents varies in the figures in this section, which present the demographic profiles of the participants.

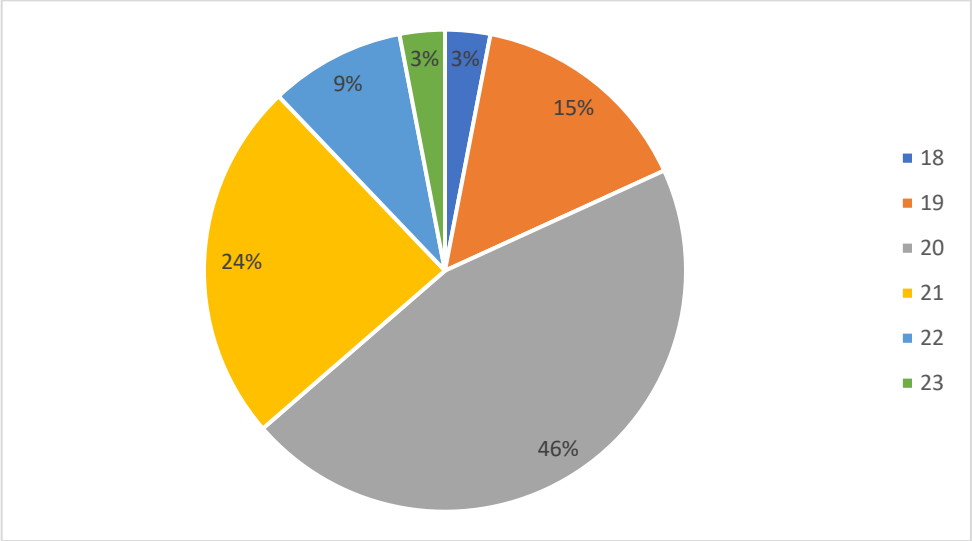

Figure 6.1 Age of participants (n=66)

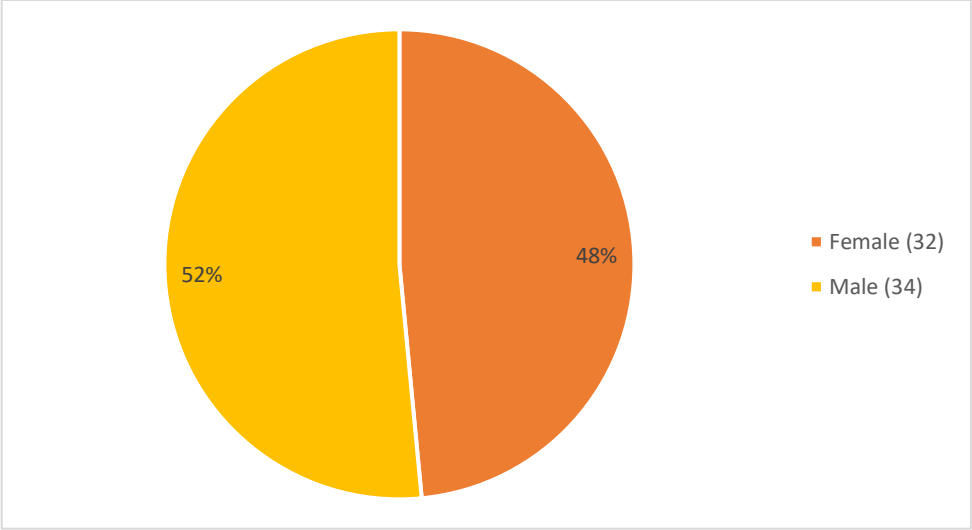

Figure 6.2 Gender balance of participants (n=66)

We can see from Figure 6.1 that the majority of participants were aged 20 (46%). This is perhaps because most participants were Year 3 undergraduates (as shown in Figure 6.3), and normally people go to university in China at the age of 18. Figure 6.2 shows that the gender ratio of participants was nearly balanced: 34 males and 32 females.

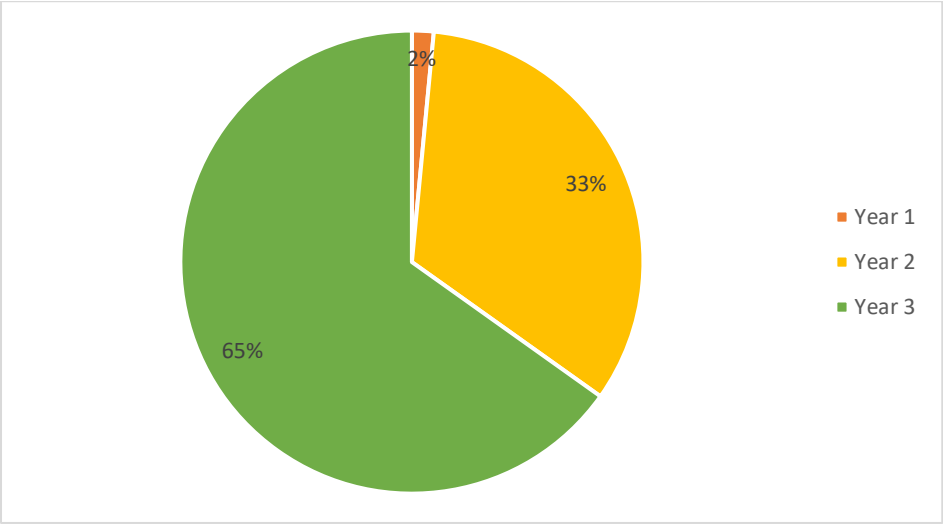

Figure 6.3 Year of participants in university (n=66)

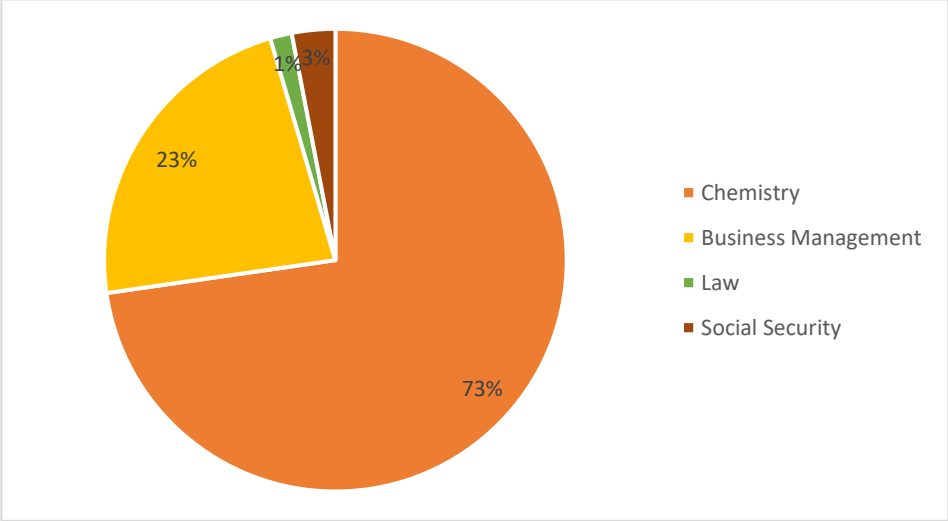

Figure 6.4 Major of participants (n=66)

From Figure 6.3, we can see that Year 3 undergraduates account for the highest proportion (65%). Figure 6.4 shows that as many as 73% of participants came from a Chemistry background. Business students were in second place, accounting for 23% of the participants.

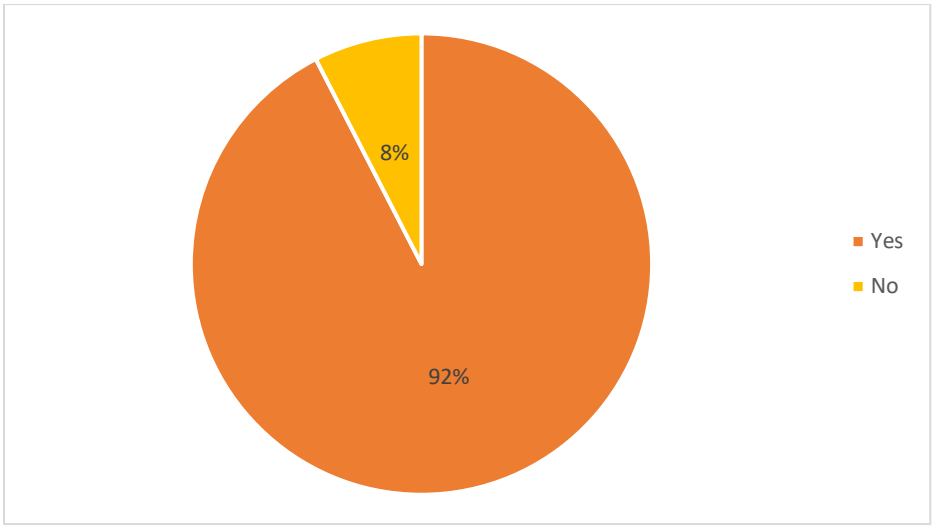

Figure 6.5 Participant has experience of using MT tools (n=66)

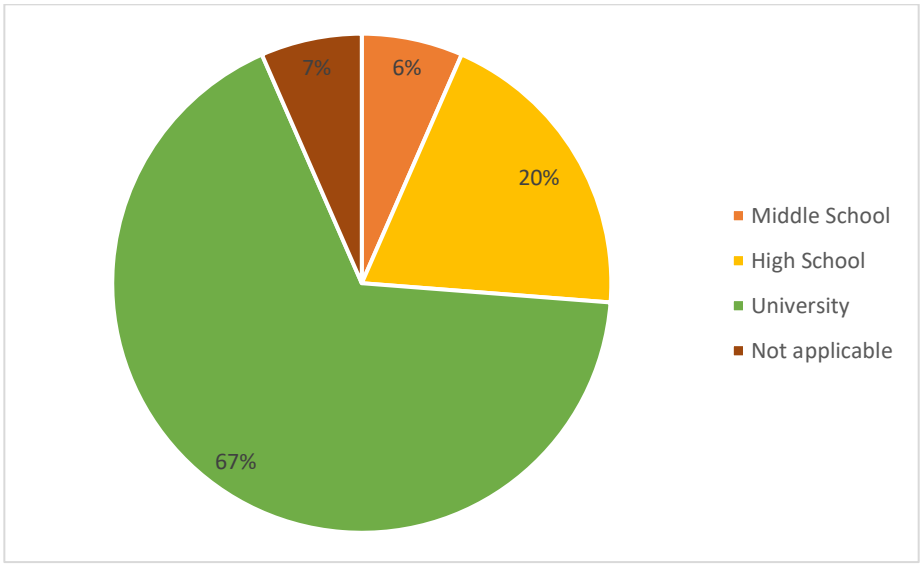

Figure 6.6 First time using MT tools (n=61)

Figure 6.5 shows that 92% of participants had experience of using MT. Since the two universities are typical mid-ranking universities and located in small cities in China, this high percentage may indicate the successful popularization of MT tools in China. From Figure 6.6, it can be seen that most participants (67%) started to use MT tools in university. 20% of participants have been using MT tools since high school. To gain a

deeper understanding of MT usage, participants were asked to clarify the frequency of their use of MT tools. As shown in Figure 6.7, more than half of them (34) used MT every month (if we combine “A lot”, “Daily”, “Often”, “A few times per week” and “A few times per month”). However, 19 participants claimed that they rarely used MT tools, accounting for almost 32% of participants. This merits further investigation. For this purpose, questions related to perceived MT quality and reasons for not using MT were included. See Figures 6.8 and 6.9 for more information.

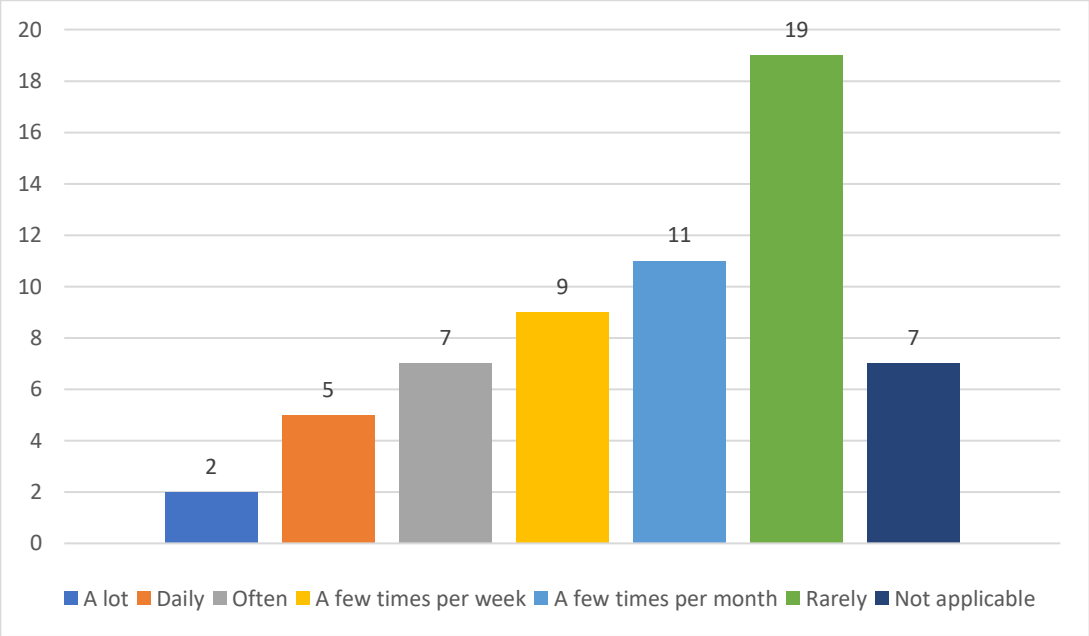

Figure 6.7 Rate of MT usage (n=60)

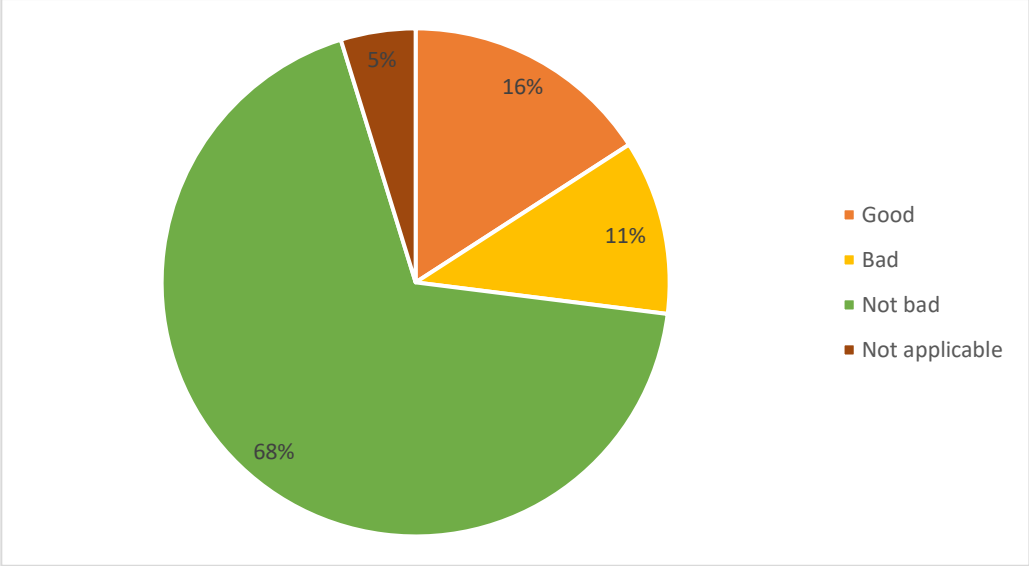

Figure 6.8[79] Quality of MT (n=63)

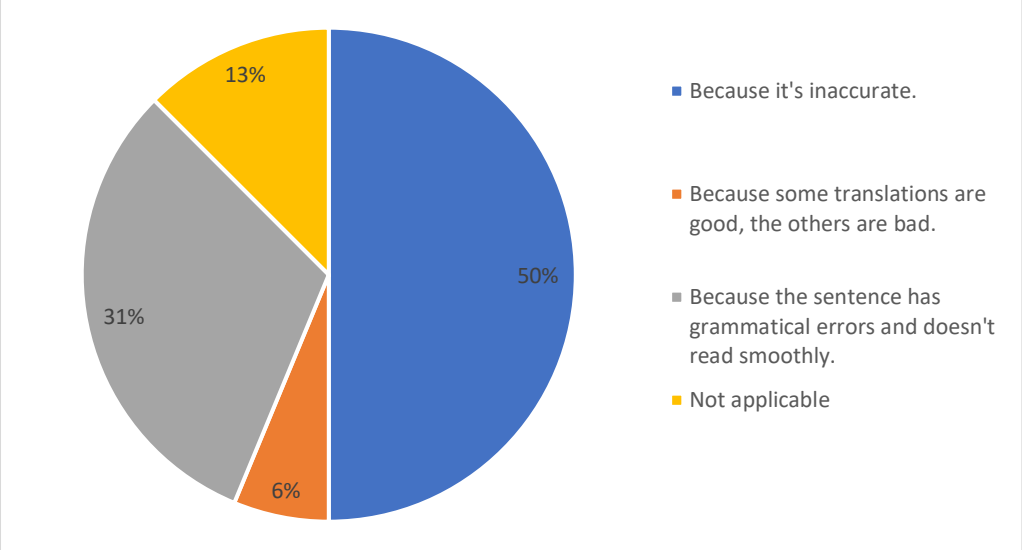

Figure 6.9 Reasons to not use MT (n=32)

16% of participants considered the quality of MT "good", and over 60% of participants considered it "not bad" (Figure 6.8). As mentioned in Section 4.1.1, the answers of the six participants who filled out the questionnaire in the wrong sequence are highlighted in the footnotes. Here, half of them (P25, P28, P12) considered MT quality as "not bad".

[79] Good (P51); Bad (P09); Not bad (P25, P28, P12); Not applicable (P10)

The distribution of answers is similar to that of the main group. Regarding the reasons for not using MT, inaccuracy was identified as the most serious problem, as it accounts for half of the answers represented in Figure 6.9. The second most common reason for not using MT is that sentences have grammatical errors and do not read smoothly (31%).

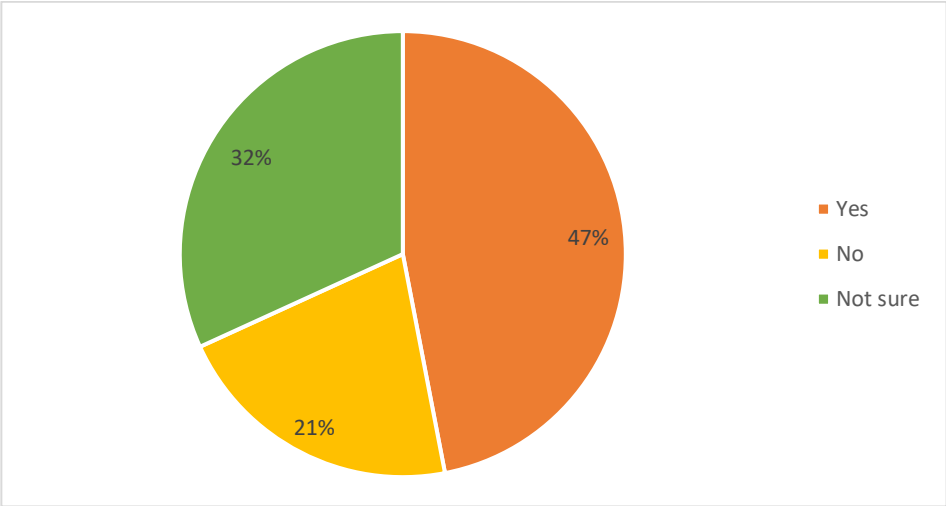

Figure 6.10[80] Participant has experience of watching videos with machine translated subtitles (n=66)

Figure 6.10 shows that 47% of participants had experience of watching videos with machine translated subtitles. 32% of participants were not sure about this. This result is worthy of some discussion. From the perspective of the researcher, the percentage of participants who were unsure should be higher, and the percentage of participants who claimed they had experience should be lower. As a viewer and a former member of a Chinese fansub group, in the researcher's own experience, while many people in China watch subtitled foreign material on TV or DVD/VCD, there are also many people who watch foreign videos online. The copyright of these online videos has long been in an ambiguous situation (see Section 2.2.5.1). Fansub groups translate and subtitle the

[80] Yes (P10, P51); No (P09, P12, P25); Not sure (P28)

videos, release them online, usually on their forum or website, so that people can easily download them for free. Aside from the fansubbed videos, some other videos online are also translated and subtitled. However, their source, translator and subtitler remain unattributed. To investigate it further, a few weeks after the experiment, the author contacted some participants remotely via an online chat tool, asking them why they picked "Yes" to this statement. The researcher learned that they were actually unsure whether they had watched videos with machine translated subtitles before, but sometimes simply assumed that this was the case: "When we're watching foreign movies online, if the subtitles have many mistakes or look illogical, we believe they must be machine translated" in the words of one participant.

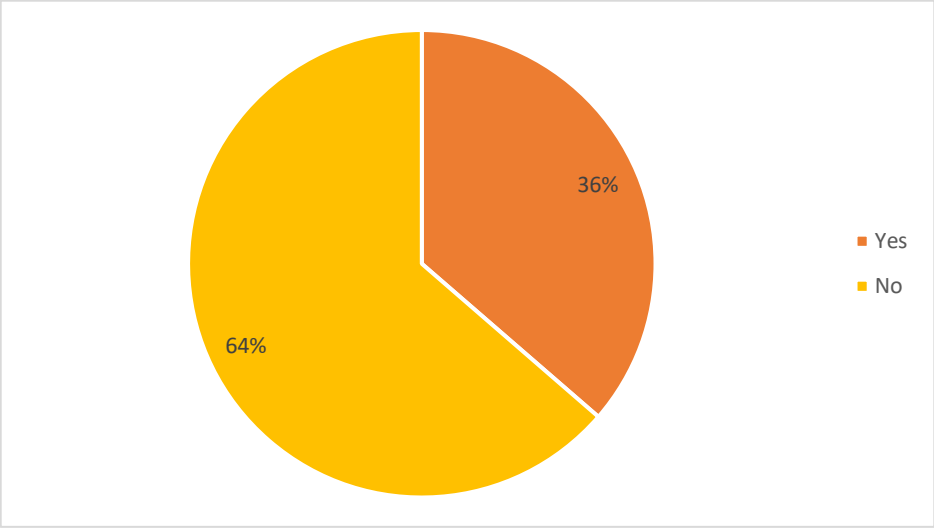

Figure 6.11 Participant has experience of using MOOCs (n=66)

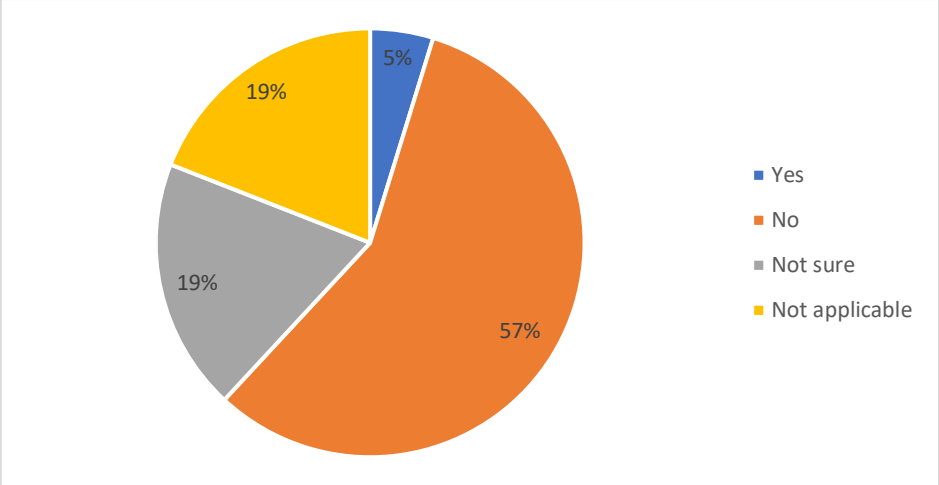

Figure 6.12[81] Participant has experience of watching MOOCs with machine translated subtitles (n=42)

It can be seen from Figure 6.11 that 24 participants (36%) had used MOOCs before. Regarding the question "Do you have experience of watching MOOCs with machine translated subtitles?", Figure 6.12 indicates that most participants had no such experience. Among the 42 participants, 5% said yes, 57% gave a negative answer, and 19% were not sure, while others (19%) skipped the question. Still, the 5% of participants here might have the same issue as the 47% of participants in Figure 6.10. They either did have experience or they were actually unsure. Of the six participants who completed the pre-task questionnaire in Step 1 in the end, none gave a positive answer to this question. Although participants were not told whether the subtitles of the MOOC video were machine translated or not, they would have made their own judgement after watching it. If the six participants thought the subtitles were machine translated, they might have given a positive answer or simply ticked "unsure" to the question even if they had no experience of MT subtitled MOOCs before they watched this one. Here, two participants skipped the question and one participant was unsure. It is unknown

[81] No (P09, P12, P25); Not sure (P28); Not applicable (P10, P51)

whether the responses of these three participants was based on their experience including or excluding watching the current MOOC video.

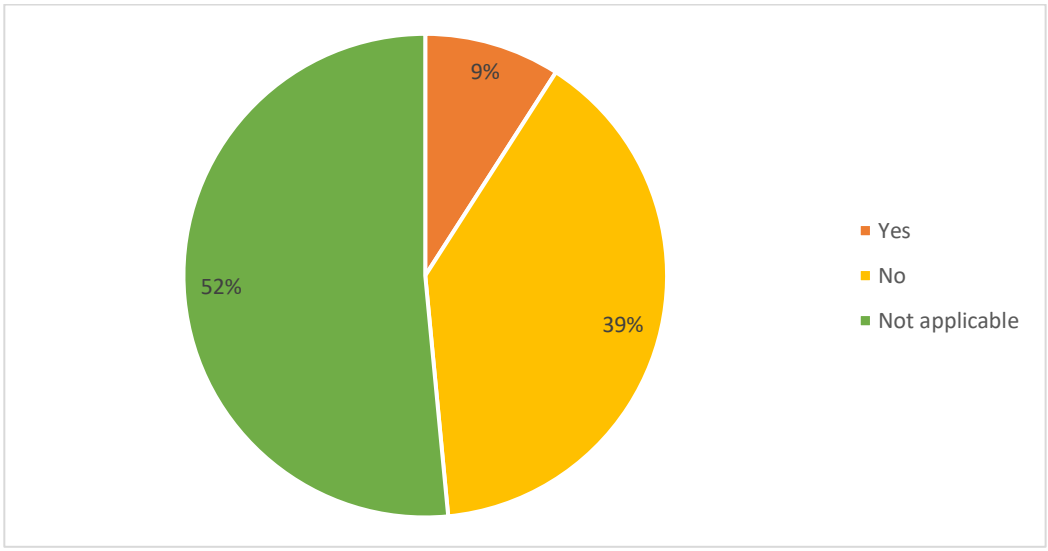

Figure 6.13[82] Participant believes MT can fully transfer the meaning of source language (n=66)

When questioned if participants believed MT could fully transfer the meaning of the source language text, only 9% of participants believed so, 39% of them answered negatively, and 52% of participants skipped the question (Figure 6.13). None of the highlighted six participants gave a positive answer. Four of them answered "No", the other two skipped the question. This result indicates that participants were not confident in the quality of machine translated subtitles, while the results presented in Figure 6.8 show only 11% of participants claimed the MT quality was bad, and the majority of participants (84%) believed the quality was "Not bad" or "Good". It could be the word "fully" that makes the difference. Although most participants felt that MT was

[82] No (P09, P12, P25, P28); Not applicable (P10, P38)

not of poor quality in general, they did not think it could convey the meaning of the source text completely.

In summary, the most typical respondent profile in this research is that of a “20-year-old", “Year-3 undergraduate”, with a “Chemistry background”. As many as 92% of participants had used MT before: 67% of them started using it since starting university, and almost 57% of them used MT every month. Regarding the quality of MT, 84% of participants believed it was “not bad” or “good”. However, contradicting with this point, only a few participants believed MT could fully transfer the meaning of the source language text. For those who had never used MT, their reasons mostly related to “inaccuracy of MT output”. 64% of all participants had experience of taking MOOCs. A questionable result is that 5% of participants claimed they had used MOOCs with machine translated subtitles. Similarly, 47% of all participants claimed they had watched videos with machine translated subtitles, which is also questionable. To the best knowledge of the researcher and based on the interviews with participants afterwards, the ‘true’ percentages for both results are probably lower.

6.2.2 Online English test

The online English proficiency test[83] for identifying English proficiency levels follows the Common European Framework of Reference for Languages.[84] According to this

[83] “General English”, Cambridge Assessment English. http://www.cambridgeenglish.org/test-your-english/general-english/ (Accessed Jan 12, 2018)
[84] “Common European Framework of Reference for Languages”, Council of Europe. https://www.coe.int/en/web/common-european-framework-reference-languages/ (Accessed Mar 28, 2019).

framework, 20 more correct answers out of 25 corresponds to Level C (proficient user). As discussed in Section 5.1.2, generally speaking, the English level of Chinese people is low. Thus, it had been decided that all participants should have a low or medium level of English. Among the 61 participants who completed the full experiment, seven of them had Level C in the English test. They were P26, P34, P36, P38, P49, P51, and P54. Rather than eliminate the data for these seven participants, it was decided to scrutinize their comprehension scores first to see if they scored higher than other participants, because this may give insight into the influence of English proficiency on comprehension. If these participants had a varied performance in comprehension, then it may not be necessary to remove them from the data analysis in the following sections.

Table 6.1 Participants' English Scores

Group PE	English Score	Group RAW	English Score	Group HT	English Score
P01	12	**P25**	12	**P47**	15
P02	18	**P26**	21	**P48**	18
P03	16	**P27**	19	**P49**	21
P04	14	**P28**	8	**P50**	14
P05	15	**P29**	13	**P51**	20
P06	9	**P30**	16	**P52**	19
P07	19	**P31**	12	**P53**	9
P08	14	**P32**	14	**P54**	20
P09	15	**P33**	15	**P55**	11
P10	19	**P34**	22	**P56**	18
P11	9	**P35**	17	**P57**	14
P12	17	**P36**	20	**P58**	19
P13	15	**P37**	16	**P59**	18
P14	15	**P38**	20	**P60**	13
P15	14	**P39**	17	**P61**	18
P16	11	**P40**	15		
P17	11	**P41**	15		
P18	8	**P42**	15		
P19	13	**P43**	11		
P20	7	**P44**	13		
P21	12	**P45**	11		
P22	11	**P46**	12		
P23	13				
P24	10				
Mean	13.21	**Mean**	15.18	**Mean**	16.47
SD	3.34	**SD**	3.63	**SD**	3.58

Mean = 13.96; Median = 14; SD = 3.20 (54 participants)

Mean = 14.72; Median = 15; SD = 3.69 (61 participants)

Table 6.1 shows the English score of participants displayed according to the different groups (PE/RAW/HT). Data on the seven participants who had Level C is underlined in the table. The mean, median and SD is calculated both including and excluding the seven participants. When including the seven participants, the mean of 61 participants' English score is 14.72 (SD = 3.69), the median is 15, corresponding to Level B1 (threshold). When they are excluded, the mean of the remaining 54 participants' English score is 13.96 (SD = 3.20) and the median is 14, still corresponding to Level B1 (threshold).

6.3 Question and hypotheses

The main research question, main hypothesis and sub-hypotheses for the main experiment are as follows:

Main research question: Is there a difference in reception between participants who are offered raw MT subtitles and those who are offered full PE subtitles or HT subtitles?

Main hypothesis: Participants who are offered full PE subtitles or HT subtitles will score higher on our reception metrics compared with those who are offered raw MT subtitles.

Regarding "Response":

Hypothesis 1a: More attention is allocated to the subtitle area (AOI_SUB) when raw MT subtitles are displayed than when full PE subtitles and HT subtitles are displayed.

Hypothesis 1b: More attention is allocated to the image area (AOI_IMA) when full PE subtitles and HT subtitles are displayed than when raw MT subtitles are displayed.

Measured by glances count, fixation count and glance duration (see Section 7.3).

Regarding "Reaction":

Hypothesis 2: The level of comprehension is higher with full PE subtitles and HT subtitles than with raw MT.

Measured by comprehension testing (see Section 6.4.1).

Hypothesis 3: Average fixation duration is shorter when full PE subtitles and HT subtitles are displayed than when RAW subtitles are displayed.

Measured by average fixation duration (see Section 7.3).

Regarding "Repercussion":

Hypothesis 4: Attitudes are more positive among participants shown full PE subtitles or HT subtitles than those shown raw MT subtitles.

Measured by attitude survey (see Section 6.4.2).

6.4 Post-task questionnaire

As mentioned in Section 5.3.2, 61 participants completed the post-task questionnaire. The questionnaire had two parts: comprehension testing and attitude survey (see Appendix E1).

6.4.1 Comprehension testing

The hypothesis that relates to comprehension testing is repeated below:

Regarding "Reaction":

Hypothesis 2: The level of comprehension is higher with full PE subtitles and HT subtitles than with raw MT.

Regarding the relationship between HT subtitles and full PE subtitles, a sub-hypothesis is proposed as below:

Hypothesis 2.1: The level of comprehension is higher with HT subtitles than with full PE subtitles.

It has to be noted that Hypothesis 2.1 is based on the assumption that the quality of HT subtitles is higher than full PE subtitles and raw MT subtitles.

There were 13 questions in this part. The full score is 13 points. As mentioned in Section 5.3.1, the post-task questionnaire (see Appendix E1) for the main experiment is slightly different from the one for the pilot study, but the scoring mechanism remains the same. Table 6.2 presents the comprehension testing score per group. We can see that the highest score of Group PE is higher than the other two groups, and its mean is also higher than that of the others. Regarding the mode of each group, for Group PE, six participants scored 10 and another six scored 11; for Group RAW, seven participants scored 8; and for Group HT, three participants scored 9, three scored 10, and another three scored 11.

Table 6.2 Comprehension testing score per group

	Max	**Min**	**Mode**	**Mean**	**SD**
Group PE (24)	13	6	10, 11	9.58	1.74
Group RAW (22)	11	6	8	8.55	1.37
Group HT (15)	12	6	9, 10, 11	9.47	1.85

Table 6.2 suggests that Group PE performed the best, Group HT performed second best and Group RAW performed the worst, which seems to agree with Hypothesis 2, but disagree with Hypothesis 2.1. To validate this point and to see if the scores of the groups are significantly different from each other, a one-way ANOVA (Analysis of Variance) was carried out.

6.4.1.1 ANOVA

ANOVA is used to test if there is a difference between the means of groups. This test helps to decide whether to reject the null hypothesis or accept the alternative hypothesis. One-way or two-way refers to the number of independent variables in the test. As the name suggests, one-way has one independent variable and two-way has two independent variables. A one-way ANOVA is used for making comparisons between two means from two independent groups. The null hypothesis is that the two means are equal, while the alternative hypothesis is that the two means are unequal to a statistically significant extent. To determine the significance of the result, the p-value is calculated. P-value, or probability value, refers to "the probability of observing an effect given that the null hypothesis is true" (Devore, 2011, p.301). If the p-value is less than a predetermined level, typically represented by the symbol α, the null hypothesis is rejected. In regard to the significance level, a 0.05 level of significance is usually selected in research dealing with behaviour and attitudes. It means the result is unlikely to occur more than 5 times out of 100 at random.

In the case of this study, the independent variable is the mode of production of the subtitles in the MOOC video. There are three groups for this independent variable: post-edited, raw machine translated and human translated. The dependent variable is the comprehension testing score. The null hypothesis (H_0) is that the three population means are equal. The alternative hypothesis (H_1) is to reject H_0.

Hypotheses and α:

H_0: $\mu_{pe} = \mu_{raw} = \mu_{ht}$

H_1: not H_0

$\alpha = 0.05$

Two outliers (P04 and P10, whose score in the comprehension testing – 6, is too low compared to other participants in the group, for their standard deviation values were more than twice the average standard deviation) in Group PE are eliminated from the ANOVA calculation. Although Group HT has fewer participants compared to the other two groups, since the difference only amounts to 7 against the other two groups, ANOVA can still be run for an unequal number of samples (Ross, 2014, p.458). The comprehension score of the seven participants who had Level C in the English test is presented in Table 6.3. Four participants were from Group RAW and three participants were from Group HT. Compared to their group means (Group RAW = 8.55, Group HT = 9.47), it can be seen that their comprehension scores vary a lot. A high English score is not necessarily related to a high comprehension score. For example, P49 scored 21 points in English, but 7 points in comprehension. Therefore, it was decided that the experimental data of the seven participants would not be removed from data analysis in this thesis.

Table 6.3 Comprehension testing score of the seven participants reporting Level C competence

Participant	English score	Comprehension testing score
P26 (RAW)	21	8
P34 (RAW)	22	9
P36 (RAW)	20	8
P38 (RAW)	20	10
P49 (HT)	21	7
P51 (HT)	20	12
P54 (HT)	20	9

Table 6.4[85] ANOVA between Group PE, Group RAW and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Group PE	22	218	9.91	1.99
Group RAW	22	188	8.55	1.88
Group HT	15	142	9.47	3.41

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	21.10	2	10.55	4.58	0.01	3.16
Within Groups	129.01	56	2.30			
Total	150.10	58				

[85] SS: sum of square
df: degree of freedom
MS: mean square
F: test statistic
P-value: significance level

F crit: critical value

Degree of freedom refers to "the number of values in the final calculation of a statistic that are free to vary" (Morien, 2007, p.136). F crit is also called F statistic, a value obtained when running an ANOVA test or a regression analysis. It helps to decide if a group of variables are jointly significant. If the F value is greater than F crit, the null hypothesis can be rejected. As it is only one measure of significance, the p-value should also be taken into consideration.

Table 6.4 presents the ANOVA result between the three groups. The p-value shows the probability of attaining an F this extreme or more if the null hypothesis is true. It can be seen from the table that p-value is smaller than 0.05, so that the null hypothesis can be rejected, which is consistent with the F value being higher than the critical value of F. In other words, there is a statistically significant difference between the comprehension testing score of the three groups.

6.4.1.2 LSD test

Since a significant ANOVA does not show where the difference lies in the data, the LSD (least significant difference) test needs to be conducted for making comparisons between two means from two groups. This test was developed by Fisher in 1935 and can only be used when the null hypothesis is rejected. The idea of the test is to "compute the smallest significant difference (i.e., the LSD) between two means as if these means had been the only means to be compared (i.e., with a t-test)" (Williams and Abdi, 2010, p.1). If the difference is larger than the LSD, then it should be considered a significant result.

The formula for LSD is:

$$LSD = t\sqrt{MSw\ (\frac{1}{N1} + \frac{1}{N2})}$$

Where:

t = critical value from the t-distribution table

MSw = mean square within

N = number of scores used to calculate the means

In the case of this study, the t-critical value for (α = 0.05, dfw = 56) is 2.003.[86] Here, the means from Group PE and Group RAW are compared first. When the given values are inserted into the LSD formula, we obtain the following result:

$$\text{LSD} = 2.003 \sqrt{2.30 \left(\frac{1}{22} + \frac{1}{22}\right)} = 0.916$$

From Table 6.4, it can be seen that the mean of Group PE is 9.91, and the mean of Group RAW is 8.55. Hence, the absolute value of the difference between these two means is 1.36. This value is greater than 0.916, indicating that there is a statistically significant finding between Group PE and Group RAW. Using the same method, the LSD value for Group PE and Group HT is 1.017, and the results of two groups are not significantly different from each other (0.44 < 1.017). The LSD value for Group RAW and Group HT is also 1.017, and the results of them are not significantly different from each other (0.92 < 1.017) either. In summary, the differences between the three groups are as follows:

Comparison 1: Group PE vs. Group RAW: significantly different

Comparison 2: Group RAW vs. Group HT: not significantly different

Comparison 3: Group HT vs. Group PE: not significantly different

[86] "Free student t-value Calculator", Free Statistics Calculators. https://www.danielsoper.com/statcalc/calculator.aspx?id=10 (Accessed: 12 January 2018)

It has to be noted that the quality of post-edited subtitles and human translated subtitles is assumed to be high, and that of raw machine translated subtitles is lower, though not substantially according to the HTER (see Section 5.2.2), which means the result of Comparison 1 and Comparison 3 is reasonable. However, Comparison 2 seems to go against conventional wisdom. As mentioned in Section 5.2.2, the human translation was conducted by an experienced English teacher. Hence, it can reasonably be presumed that the quality of the human translated subtitles is high. In addition, Section 5.2.2 explained that taking human translated subtitles as reference, the BLEU score for the raw machine translated subtitles was found to be at 58.63%, indicating that there is a difference between the two sets of subtitles. Therefore, reasons for the result of Comparison 2 remain unknown. Chapter 7 investigates this issue further, from the perspective of eye-tracking data.

6.4.2 Attitude survey

The hypothesis that relates to the attitude survey is repeated below:

Regarding "Repercussion":

Hypothesis 4: Attitudes are more positive among participants shown full PE subtitles or HT subtitles than those shown raw MT subtitles.

Regarding the relationship between HT subtitles and full PE subtitles, a sub-hypothesis is proposed as below:

Hypothesis 4.1: Attitudes are more positive among participants shown HT subtitles than those shown full PE subtitles.

Same as Hypothesis 2.1, it has to be noted that Hypothesis 4.1 is based on the assumption that the quality of HT subtitles is higher than full PE subtitles and raw MT subtitles.

This part consists of 14 five-point Likert scale questions (see Appendix E1). All statements and their responses are presented from Figure 6.14 to Figure 6.27. By adding up the times that each option was chosen for each question by each group, Table 6.5 shows a clear distribution of different answers for all groups in the form of percentage. The number in parentheses is the exact number of answers.[87] In the survey (see Appendix E1), 12 statements were related to the subtitles as just presented to them, and two statements (S22 and S23) were about their attitude towards MT in general (not the specific machine translated subtitles in this study). It should be noted that throughout the experiment, participants were not informed whether the subtitles were translated by human or machine. However, in the informed consent (see Appendix A1) and pre-task questionnaire (see Appendix C1), "machine translation" and "machine translated subtitles" were both mentioned. It is participants' own judgement whether

[87] For example, if all the participants in Group PE chose 'Strongly agree' for all the 14 statements, then this scale would have been chosen 336 times (24×14). However, it was chosen 113 times by Group PE because they did not 'Strongly agree' with all the statements.

they were watching machine translated subtitles or not. As one participant from Group PE put it, "I did not even think about MT when I was doing the experiment." When asked his opinion about the subtitles, he answered, "I think it's human translated, because in my mind, MT is rather rigid." After asking eight more participants, three of them (Group PE) thought the subtitles were machine translated, others (five from Group PE and one from Group RAW) replied that they had never thought about MT when doing the experiment.

Before focusing on answers to individual groups of statements, an overview of the data is presented in Table 6.5 first. In respect of the option "strongly agree", the number for Group PE is higher than that of the other two. In addition, for all groups, most answers go to "agree", fewest answers go to "disagree" and "strongly disagree", indicating that participants' attitude to the subtitles is not bad. In other words, participants held a positive view towards the machine translated subtitles.

Table 6.5[88] Attitude survey results per group

	Strongly agree	**Agree**	**Neutral**	**Disagree**	**Strongly disagree**
Group PE (24)	33.63% (113)	48.21% (162)	13.39% (45)	4.46% (15)	0.30% (1)
Group RAW (22)	17.21% (53)	57.79% (178)	18.51% (57)	6.49% (20)	0
Group HT (15)	28.16% (58)	38.83% (80)	24.27% (50)	8.74% (18)	0

[88] One participant in Group HT failed to answer Statements 24, 25, 26, and 27. Hence, the number of respondents for the four questions in this group is 14.

For the option “Neutral”, it is not clear what the participants really thought. Hence, by adding up the percentage of “Strongly agree” and “Agree”, the result for the percentage of agreements is Group PE (81.84%) > Group RAW (75%) > Group HT (66.99%), while by adding up the percentage of “Disagree” and “Strongly disagree”, the result for the percentage of disagreements is Group HT (8.74%) > Group RAW (6.49%) > Group PE (4.76%). This means that Group PE was the most satisfied with their subtitles, while Group HT was the most unsatisfied with their subtitles. This result is unexpected and rejects both Hypotheses 4 and 4.1. In addition, this result is not consistent with the comprehension testing score (Group PE > Group HT > Group RAW, see Section 6.4.1), because while Group PE performed the best in both comprehension testing and attitude survey, the relationship between Group RAW and Group HT is inconsistent.

As mentioned in Section 4.2.2.3.1, in the attitude survey, all statements are designed based on the six concepts in TAM and can be categorized as follows:

Perceived usefulness: Statements 14, 15

Perceived ease of use: Statements 16, 17, 18, 19

Perceived enjoyment: Statement 20

Perceived quality: Statement 21

Intention to accept machine translated subtitles: Statements 22, 23

Compensation: Statements 24, 25, 26, 27

1) Perceived usefulness

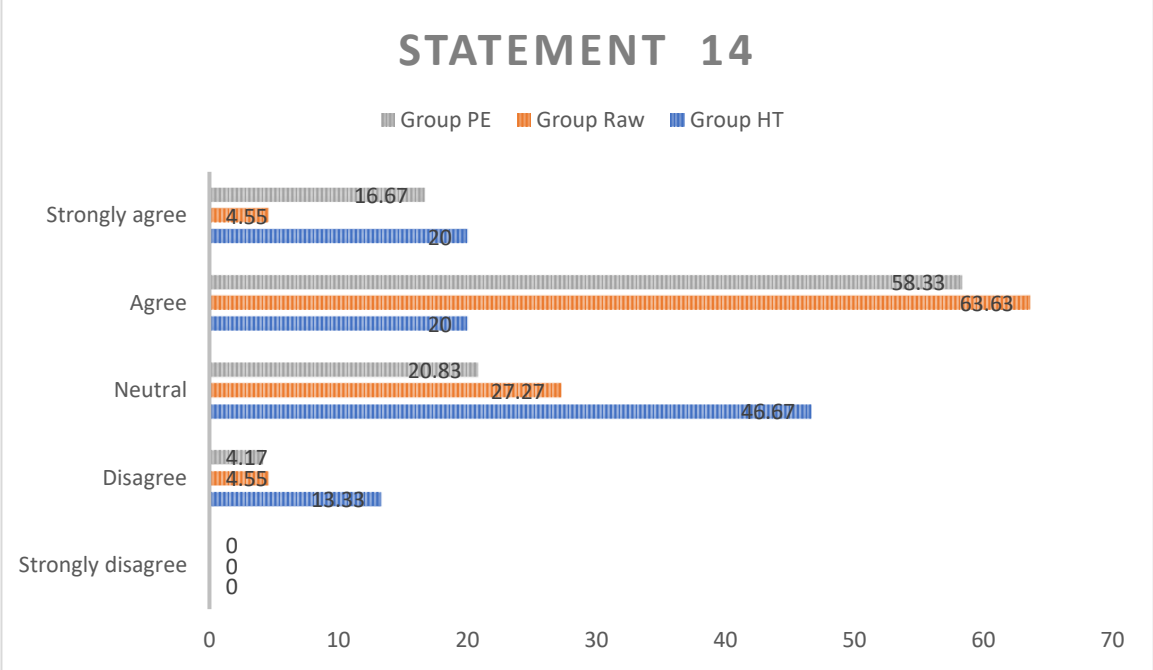

Figure 6.14 Level of agreement with the statement: "The subtitles allow me to fully understand the contents of the MOOC" (percentage of participants)

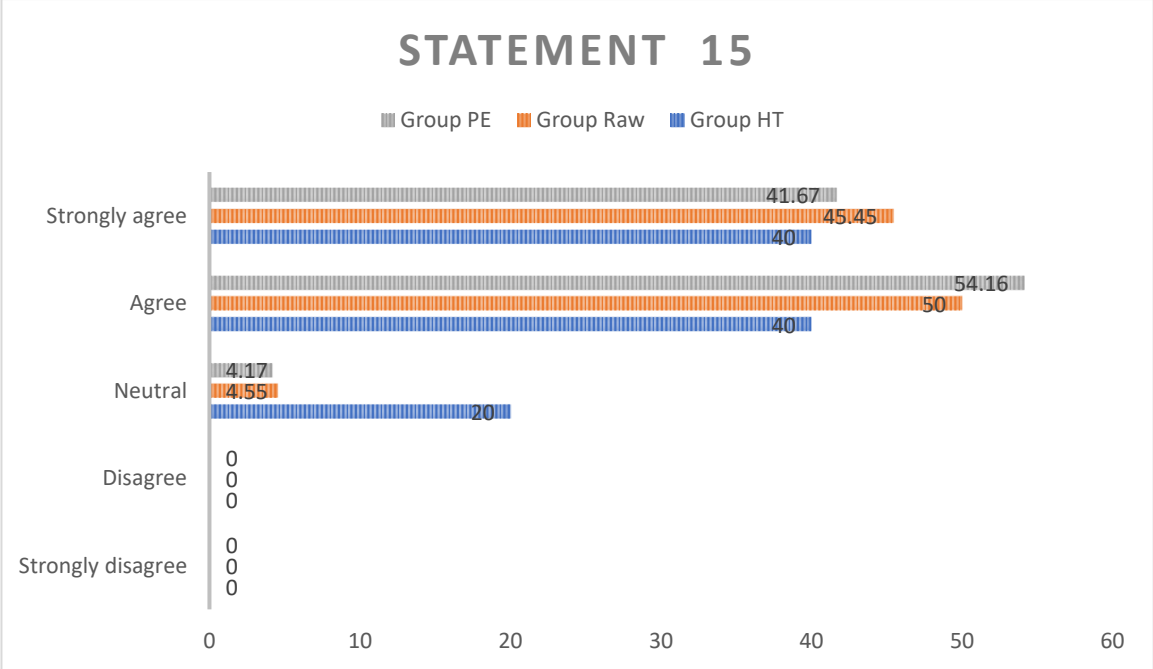

Figure 6.15 Level of agreement with the statement: "The subtitles are useful to me" (percentage of participants)

Based on Figures 6.14 and 6.15, it can be seen that 75% (16.67%+58.33%) of participants in Group PE and 68.18% (4.55%+63.63%) participants in Group RAW perceived they had fully understood the MOOC content by reading subtitles. 95.83% (41.67%+54.16%) of participants in Group PE and 95.45% (45.45%+50%) of participants in Group RAW agreed that the subtitles were useful to them. However, only 40% and 80% of Group HT agreed with the two statements respectively.

2) Perceived ease of use

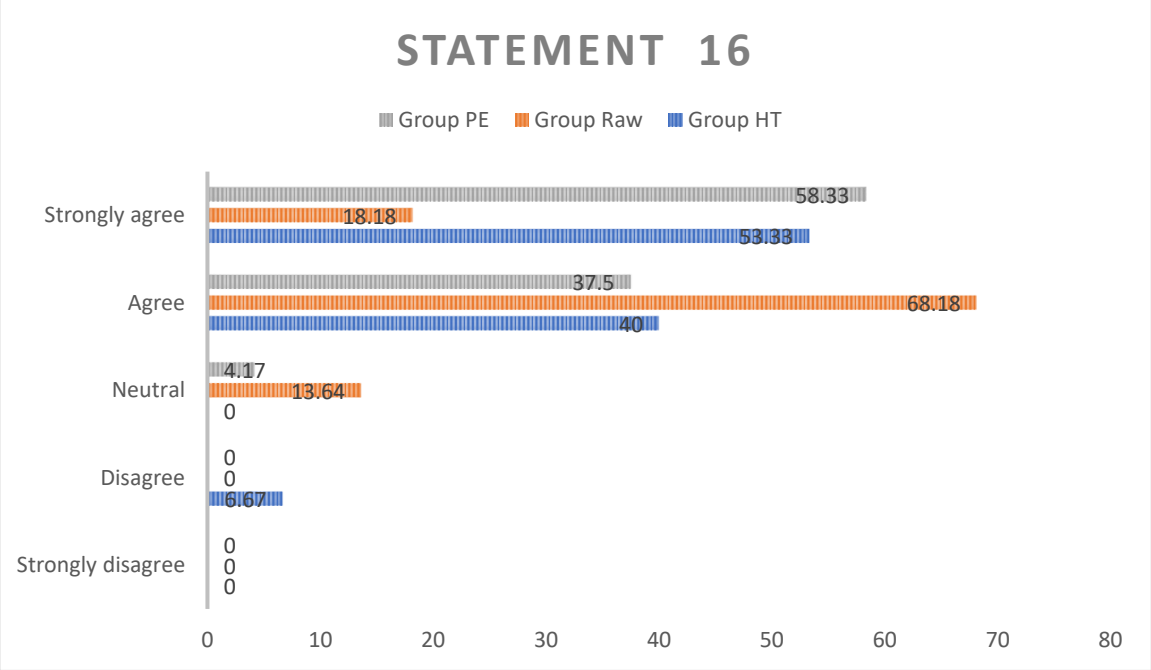

Figure 6.16 Level of agreement with the statement: "The subtitles are easy to understand" (percentage of participants)

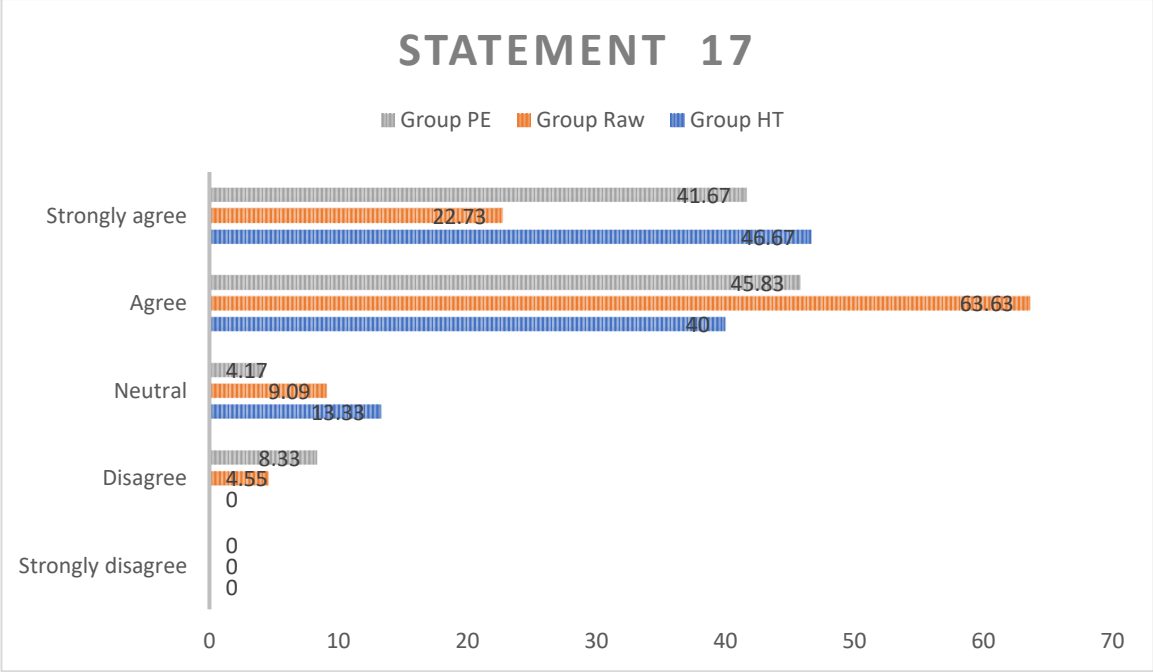

Figure 6.17 Level of agreement with the statement: "Interacting with the subtitles does not require a lot of my mental effort" (percentage of participants)

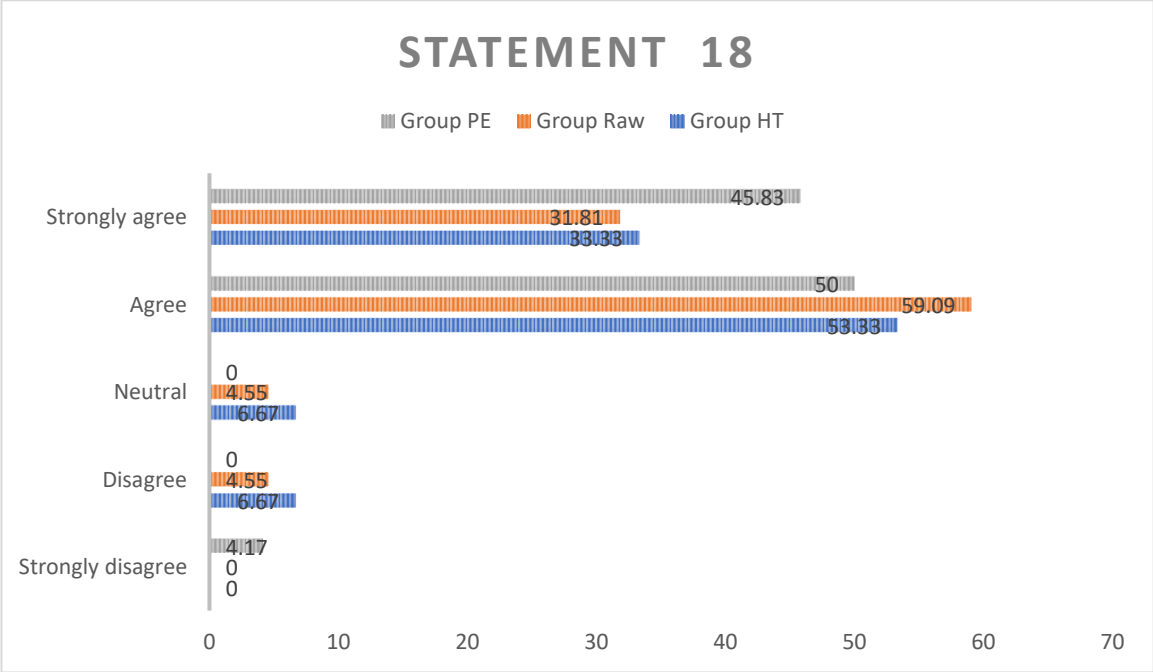

Figure 6.18 Level of agreement with the statement: "I would find it easy to get the information I need from subtitles" (percentage of participants)

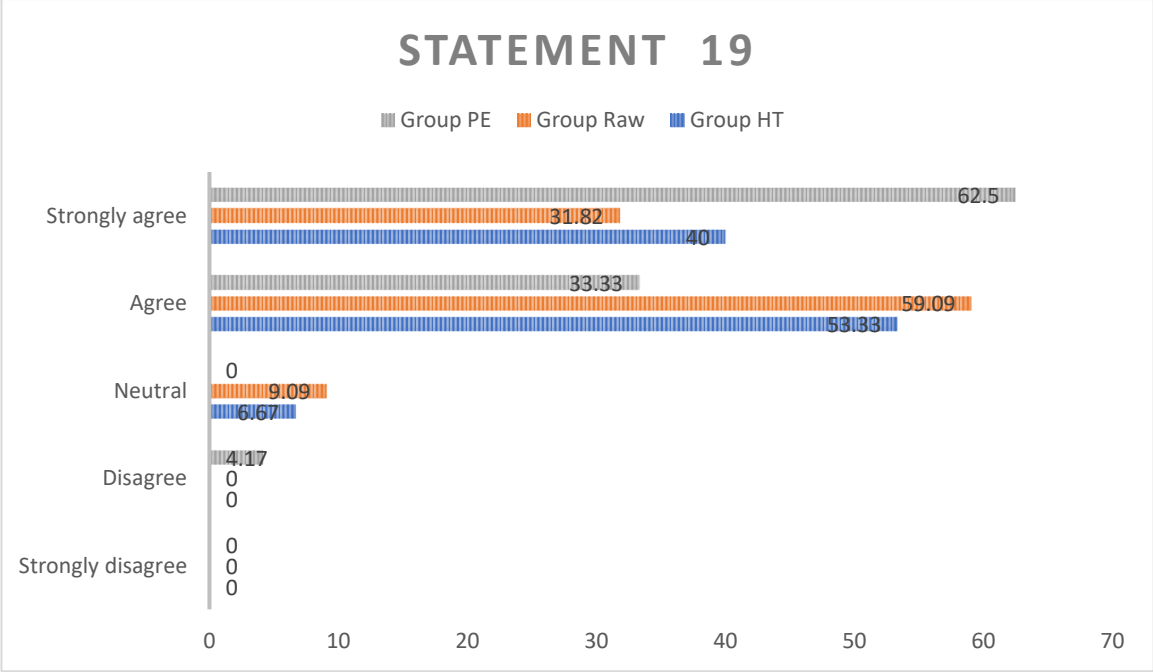

Figure 6.19 Level of agreement with the statement: "The subtitles are clear and understandable" (percentage of participants)

Table 6.6 Percentage of agreements (including "strongly agree" and "agree") on the four statements per group

	S16	S17	S18	S19
Group PE	95.83%	87.50%	95.83%	95.83%
Group RAW	86.36%	86.36%	90.90%	90.91%
Group HT	93.33%	86.67%	86.66%	93.33%

According to Table 6.6, it can be seen that Group PE recorded the highest percentage of agreements on all four statements relating to perceived ease of use, indicating the post-edited subtitles were easy to understand and process. However, Group HT did not consistently score lower here, but scored higher than Group RAW, except on S18. In other words, the perceived usefulness and perceived ease of use of Group PE towards their subtitles is the highest compared to the other two groups. Group HT had generally higher perceived ease of use than Group RAW, which is consistent with the

comprehension testing score (Group PE > Group HT > Group RAW, see Section 6.4.1), while Group RAW had higher perceived usefulness than Group HT, which is opposite to the comprehension testing score.

3) Perceived enjoyment

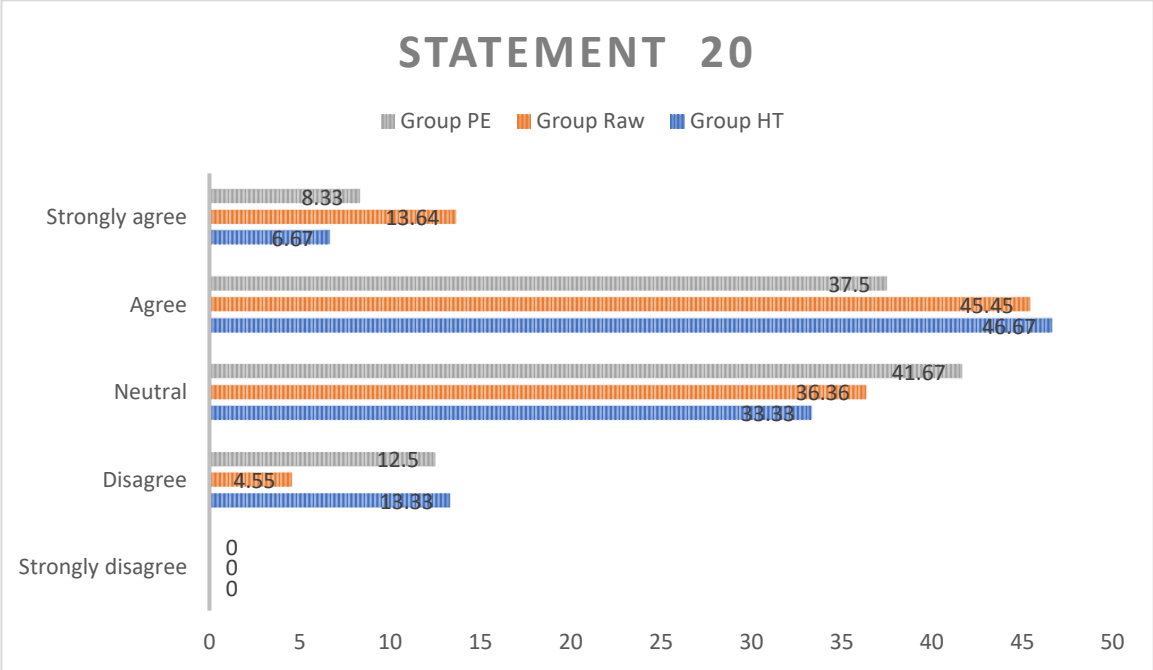

Figure 6.20 Level of agreement with the statement: "I enjoyed reading the subtitles" (percentage of participants)

As for this statement, 45.83% of participants in Group PE, 59.09% of participants in Group RAW and 53.34% of participants in Group HT enjoyed reading subtitles. Although all groups had a low agreement with this statement, it is a surprise that those who read raw machine translated subtitles gave the highest enjoyability rating to reading subtitles.

4) Perceived quality

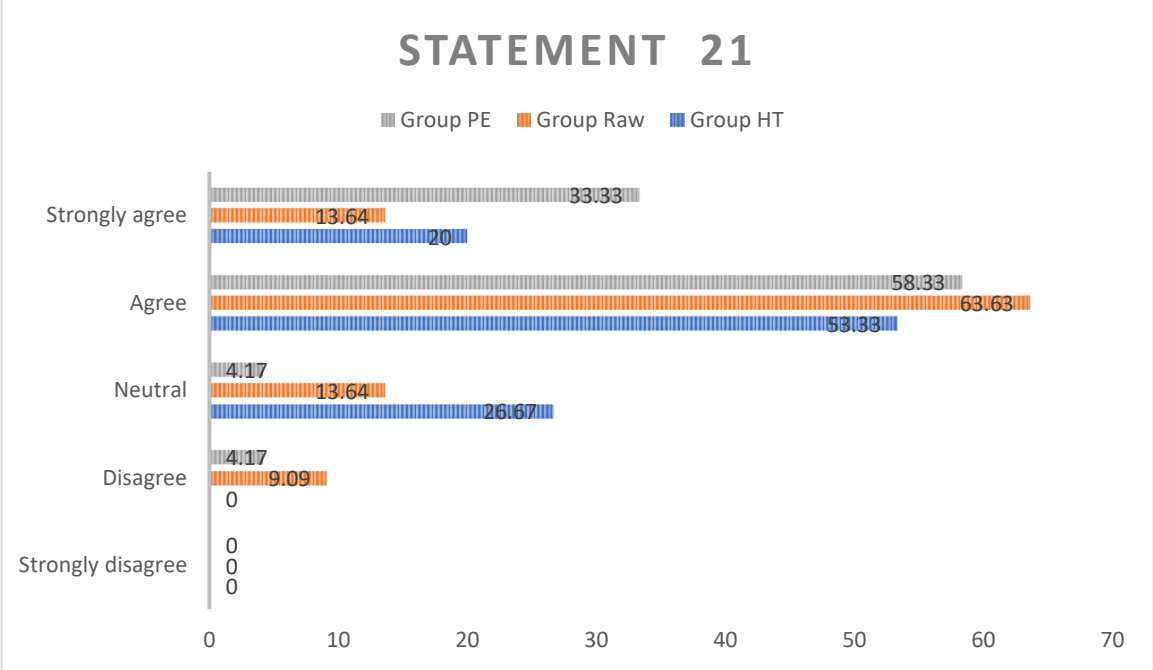

Figure 6.21 Level of agreement with the statement: "I'm satisfied with the subtitles" (percentage of participants)

Regarding this statement, 91.66% of participants in Group PE, 77.27% of participants in Group RAW and 73.33% of participants in Group HT were satisfied with the subtitles.

5) Behavioural intention

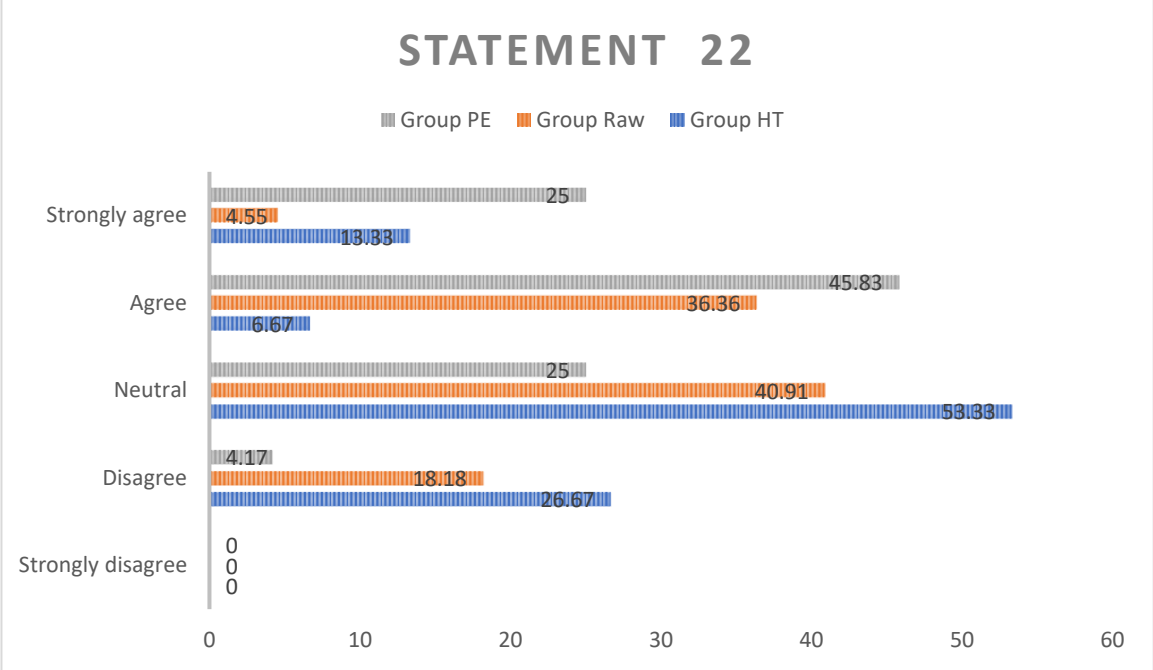

Figure 6.22 Level of agreement with the statement: "If I have a chance, I would use machine translation to translate English subtitles in the future, because I know it will do a good job" (percentage of participants)

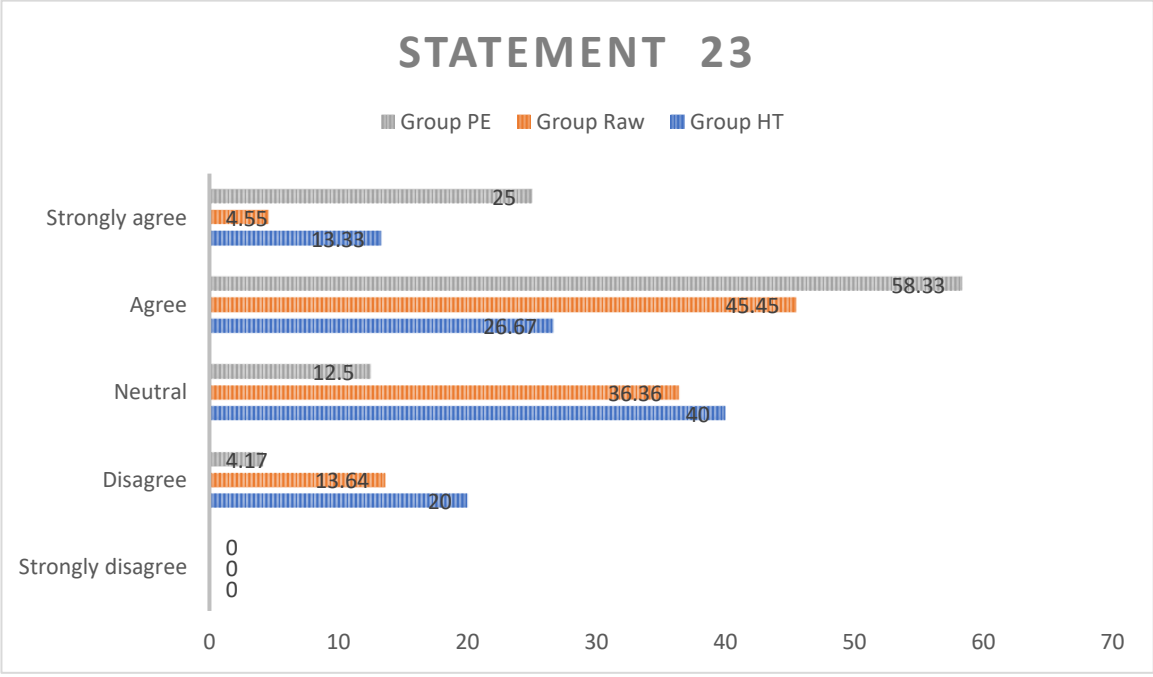

Figure 6.23 Level of agreement with the statement: "I would recommend machine translation to my friends if they need to translate subtitles" (percentage of participants)

According to Figure 6.22, 70.83% of participants in Group PE, 40.91% of participants in Group RAW, and 20% of participants in Group HT would like to use machine translated subtitles in the future. According to Figure 6.23, 83.33% of participants in Group PE, 50% of participants in Group RAW and 40% of participants in Group HT would recommend MT to their friends. Again, it should be noted that participants were not aware if the subtitles they had read were translated by human or machine, but they would have made their own judgement after completing Step 1 and Step 2. What's interesting is that Group PE and Group RAW were more likely to recommend MT, while Group HT was less likely to do so – even though they in theory did not know whether or not they had been exposed to HT or MT.

6) Compensation

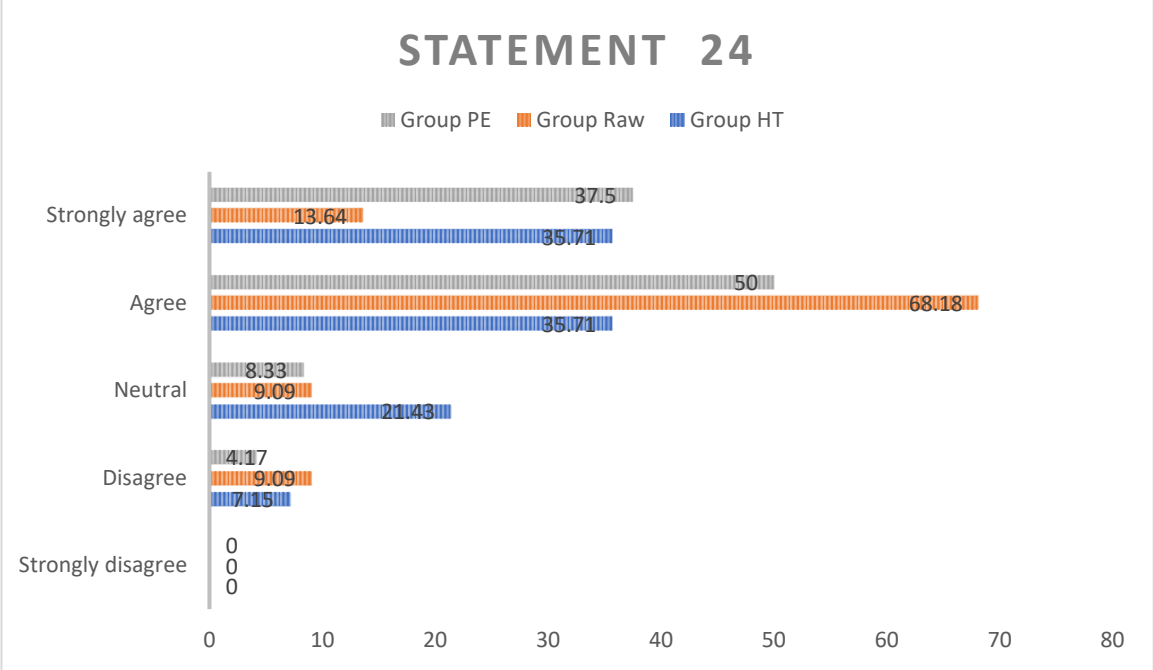

Figure 6.24 Level of agreement with the statement: "I could comprehend the subtitles if there was no one around to tell me what to do as I go" (percentage of participants)

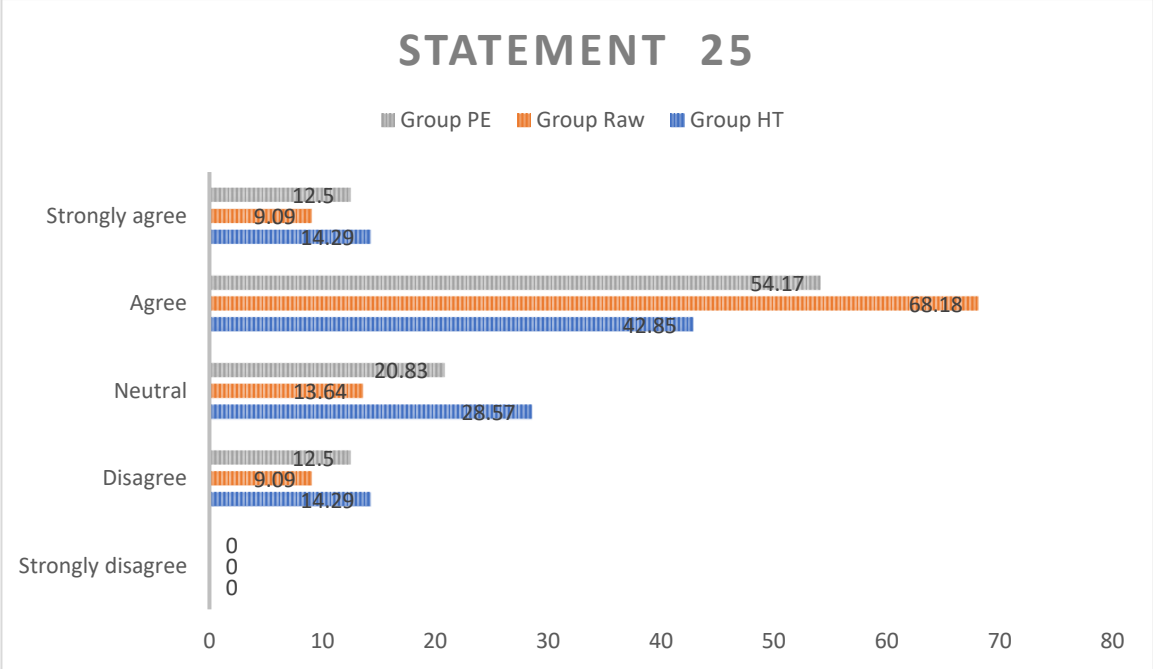

Figure 6.25 Level of agreement with the statement: "I could comprehend the subtitles if I could call someone for help if I got stuck" (percentage of participants)

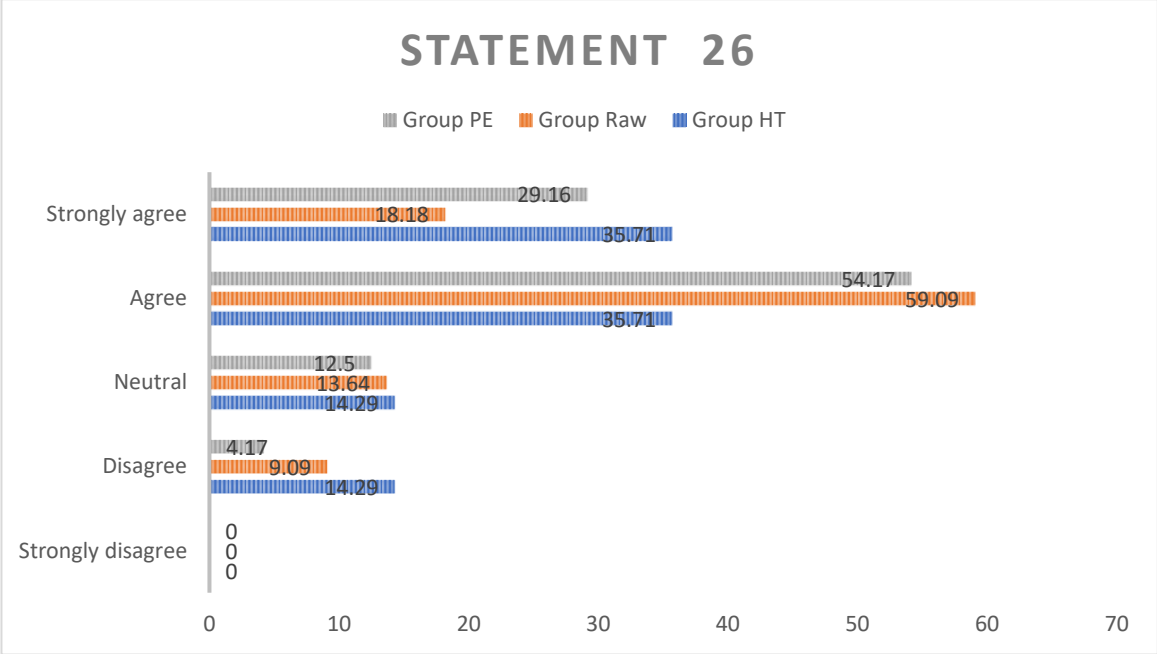

Figure 6.26 Level of agreement with the statement: "I could comprehend the subtitles if I have a lot of time" (percentage of participants)

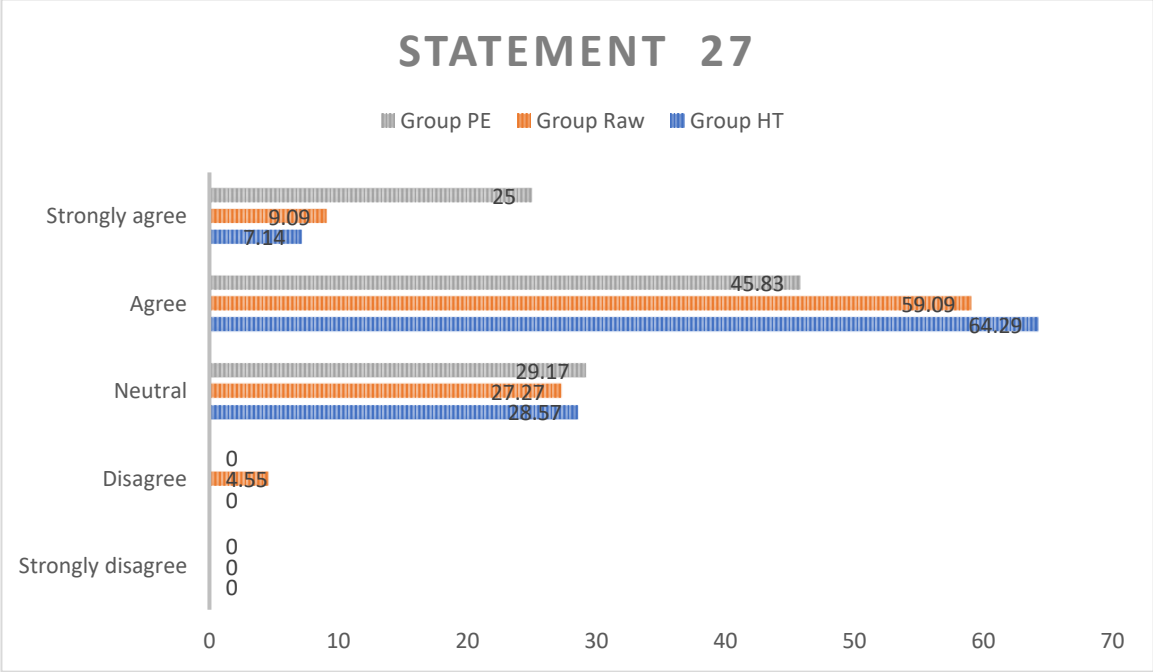

Figure 6.27 Level of agreement with the statement: "I could comprehend the subtitles if I had just the built-in help facility (i.e.: online dictionary) for assistance" (percentage of participants)

Table 6.7 Percentage of agreements (including "strongly agree" and "agree") on the four statements in each group

	S24	S25	S26	S27
Group PE	87.50%	66.67%	83.33%	70.83%
Group RAW	81.82%	77.27%	77.27%	68.18%
Group HT	71.42%	57.14%	71.42%	71.43%

In Table 6.7, except S27, the percentage of agreements with the other three statements for Group HT is the lowest. Besides, none of the percentages is above 90% for any group, indicating that in regard to compensation, participants did not hold a very positive opinion (just broadly speaking, e.g.: 87.5% for Group PE is pretty positive to S24).

In summary, with the exception of Statements 20, 25 and 27, post-edited subtitles outperformed the other two types of subtitles for all questions. However, to the statement "I enjoy reading subtitles" (S20), it is surprising that the percentage of agreements for Group PE was the lowest and that for Group RAW was the highest. Looking at the number of times each of the three groups had the lowest mean score for a particular statement, it emerges that Group PE had the lowest score on one occasion (S20), Group RAW on four occasions (S16, S17, S19 and S27), and Group HT on nine occasions (S14, S15, S18, S21, S22, S23, S24, S25 and S26). Similarly to what Comparison 2 shows in Section 6.4.1.2, this result is undoubtedly against expectation, mostly because human translated subtitles received the most negative feedback while they are normally expected to have the highest quality. The fact that Group HT had the fewest participants is conjectured to be the reason for this result. However, this point requires more investigation when analyzing the eye-tracking data for Group HT.

Chapter 7 Findings and Data Analysis II - Eye-tracking Data

7.1 Introduction

This chapter revolves around the eye-tracking data obtained from Step 2 of the experiment (see Section 5.3.2). Section 7.2 focuses on explaining what valid eye-tracking data is. This is followed by hypothesis testing in Section 7.3 and additional remarks about outliers in Section 7.4. Section 7.5 presents a discussion about the results of the eye-tracking data.

7.2 Valid data

As mentioned in Section 5.3.2, there were 61 participants who completed the full experiment. However, taking their tracking ratio into account, almost half of them had a tracking ratio below 70%. According to O'Brien (2010, p. 257), 70% could be seen as the lowest limit threshold, and participants whose tracking ratio is below 70% should be removed from the study. Therefore, it has been decided that valid eye-tracking data for this research should have a tracking ratio above 70%. As Blignaut and Wium (2014) suggest, the trackability of Asian participants' eye-tracking data can be expected to be lower than for other ethnicities, and this fact needs to be taken into account when designing eye-tracking research. It is assumed that this has to do with the darkness of the Asian eye and eye tracking manufacturers have tried to compensate by having

dark/bright pupil tracking, as described by Tobii[89]. Despite these advances, there may still be issues with tracking darker eyes. After eliminating those who had invalid eye-tracking data, there were 12 participants in Group PE, 14 participants in Group RAW and 13 participants in Group HT.

Vision is produced by the brain and the eyes working together. When one eye fails to work properly with the brain, the vision of this eye will be reduced, and the brain will favour the other eye. This condition is called amblyopia, or lazy eye[90]. The eye that the brain favors is the better eye. In this experiment, the data for both eyes was exported from SMI BeGaze. Two steps were conducted in order to find the best eye for analysis. Firstly, the calibration data for each participant was examined, especially the left eye deviation and right eye deviation, as exemplified in Figure 7.1. During calibration, participants were required to look at a moving red dot on the screen. The eye for which less deviation was reported fixated on the dot better. Secondly, the researcher checked the recordings in BeGaze with an overlay of eye-tracking data for ten seconds and decided which of the two eyes was fixating on the subtitle area, as shown in Figures 7.2 and 7.3 (the right eye was fixating more on the subtitle area for the participant in question). The better eye of each participant is clearly indicated in Table 7.1.

[89] https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/what-is-dark-and-bright-pupil-tracking/ (Accessed: 9 October 2019)
[90] https://nei.nih.gov/health/amblyopia/amblyopia_guide (Accessed: 13 March 2018)

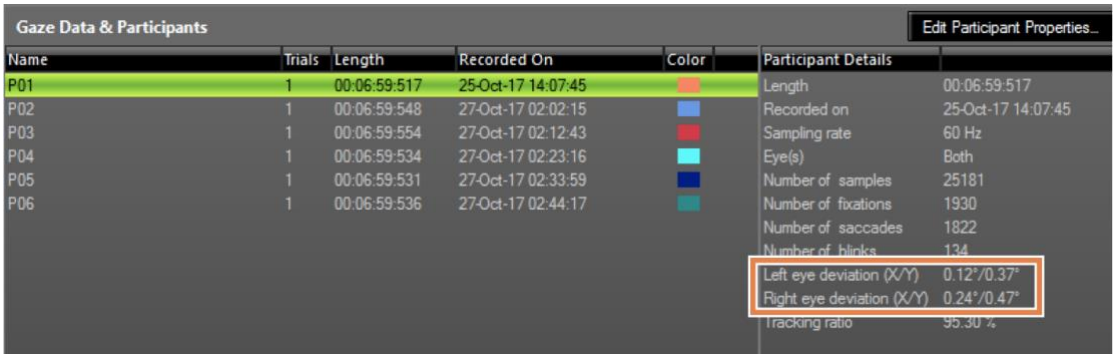

Figure 7.1 Eye deviation data (left eye deviation is smaller than right eye deviation)

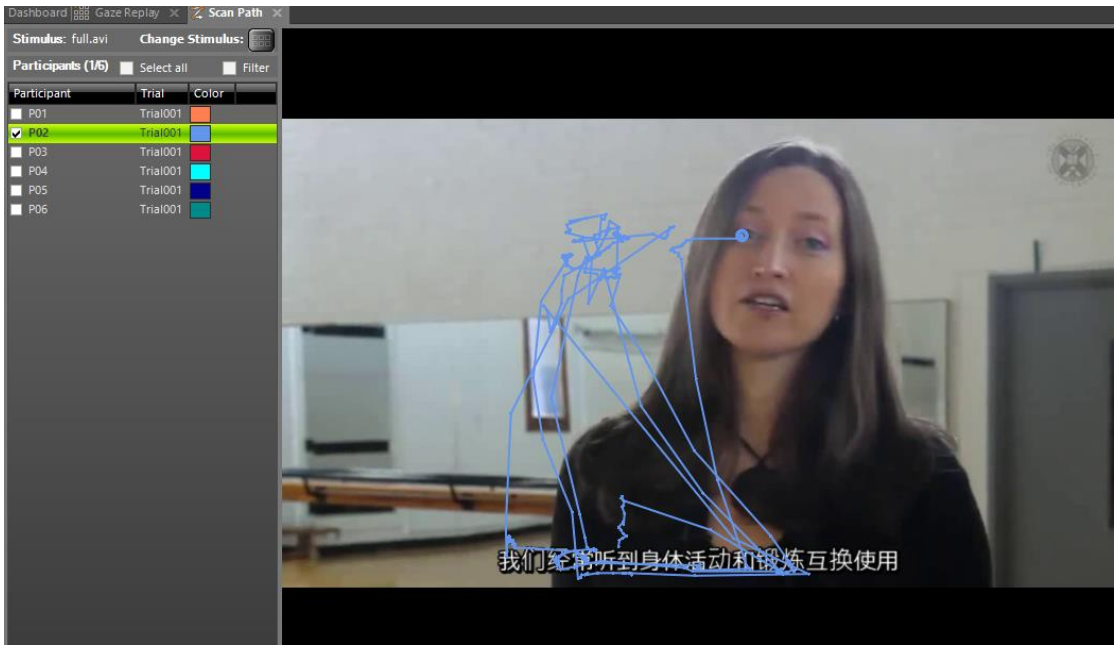

Figure 7.2 Scan path of right eye

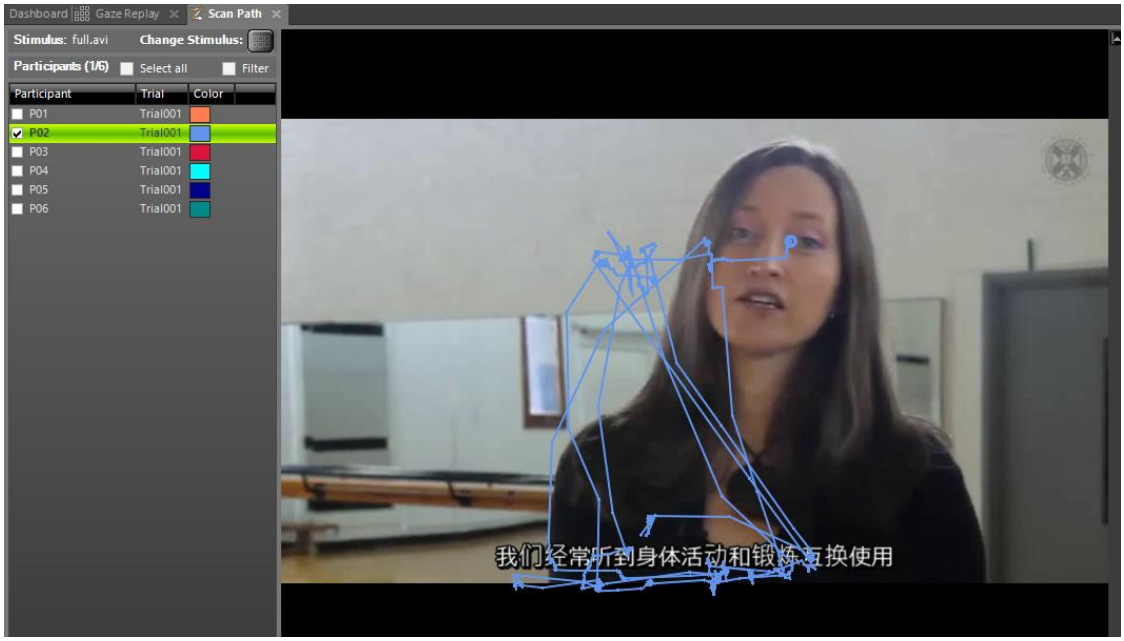

Figure 7.3 Scan path of left eye

Table 7.1 Tracking ratio of participants from each group (R for 'right eye', L for 'left eye')

	Group PE		Group RAW		Group HT
P01	93.83R	P25	85.13R	P47	97.52L
P02	91.49L	P26	82.18L	P48	96.12R
P04	97.93L	P28	86.95R	P51	95.76L
P06	90.15R	P29	77.18L	P52	80.26R
P09	80.18R	P30	91.40L	P53	75.46L
P10	97.71L	P33	90.01R	P54	86.11R
P11	97.84L	P35	90.98L	P55	97.33R
P12	90.01L	P36	95.20L	P56	76.13L
P14	73.76R	P37	95.33R	P57	72.52R
P17	78.53R	P38	98.42R	P58	89.71L
P23	96.67L	P39	91.85R	P59	71.17L
P24	81.80L	P40	85.47R	P60	89.44L
		P43	84.58R	P61	94.81R
		P46	92.09R		

7.3 Hypotheses and testing

7.3.1 Hypotheses related to eye-tracking

The main hypothesis and sub-hypotheses of this research have been clarified in Section 6.3. The hypotheses that related to eye-tracking are repeated below:

Regarding "Response":

Hypothesis 1a: More attention is allocated to the subtitle area (AOI_SUB) when raw MT subtitles are displayed than when full PE subtitles and HT subtitles are displayed.

Hypothesis 1b: More attention is allocated to the image area (AOI_IMA) when full PE subtitles and HT subtitles are displayed than when raw MT subtitles are displayed.

Regarding "Reaction":

Hypothesis 3: Average fixation duration is shorter when full PE subtitles and HT subtitles are displayed than when RAW subtitles are displayed.

All the hypotheses were built upon the premise that the quality of raw MT subtitles was lower than PE subtitles, which was lower than HT subtitles, and because of this, PE subtitles and HT subtitles would be easier for participants to process than raw MT subtitles. It has to be noted that the main focus of this study is to compare raw MT with post-edited translation and human translation. However, a comparison between post-

edited translation and human translation was also conducted though we assume that there are fewer quality differences between these two.

The MOOC video was less than seven minutes long. According to the data exported from the SMI eye-tracker, the export end trial time (visible time) was 419,000 ms. Two AOIs were defined on the video for analysis: subtitle area (AOI_SUB) and image area (AOI_IMA). Regarding the size of each AOI, AOI_SUB was 106559 px (coverage 20.2%) and AOI_IMA was 394279 px (coverage 74.9%). More eye-tracking data analysis will be carried out in combination with hypothesis testing in the following sections.

7.3.2 Hypothesis testing

It is expected that less demand on attention is required if the subtitles are easy to process. Hypotheses 1a and 1b assume that the better the quality of subtitles, the more attention participants would pay to the image area, and the less they would pay - or have to pay - to the subtitle area.

Hypothesis 1a: More attention is allocated to the subtitle area (AOI_SUB) when raw MT subtitles are displayed than when PE subtitles and HT subtitles are displayed.

This hypothesis was measured by the glances count in the subtitle AOI of each video. The higher the number of visits in AOI_SUB, the more attention participants were deemed to give to the subtitles. Thus, two sub-hypotheses have been proposed as follows:

Hypothesis 1a.1: Group RAW > Group PE and Group HT

Hypothesis 1a.2: Group PE > Group HT

Table 7.2 Glances Count in Subtitle AOI of Each Group

	Group PE	Group RAW		Group HT	
P01	150	P25	163	P47	119
P02	163	P26	151	P48	119
P04	107	P28	175	P51	197
P06	144	P29	62	P52	24
P09	121	P30	119	P53	152
P10	178	P33	115	P54	102
P11	132	P35	128	P55	117
P12	135	P36	140	P56	138
P14	48	P37	186	P57	106
P17	186	P38	132	P58	163
P23	171	P39	173	P59	139
P24	79	P40	129	P60	112
		P43	202	P61	145
		P46	105		
Mean	134.50		141.43		125.61
SD	41.04		36.79		40.32

According to the means in Table 7.2, Group RAW had a higher number of visits to AOI_SUB (mean = 141.43) than Group PE (mean = 134.50) and Group HT (125.61), thereby supporting Hypothesis 1a.1. The mean of Group PE was higher than that of Group HT, which supports Hypothesis 1a.2.

In addition, Table 7.2 shows that the data for P14 in Group PE, P29 in Group RAW and P52 in Group HT are outliers. Hence, excluding the three outliers, the means for Group PE, Group RAW and Group HT are 142.36 (SD = 32.19), 147.54 (SD = 30.01) and 134.08 (SD = 27.51) respectively. This result still supports Hypothesis 1a. The three outliers are removed in the following analysis. More discussion on the three outliers will be presented in Section 7.4.

One-way ANOVA is carried out here to find out whether there is a significant difference between the glances count (AOI_SUB) for the three group pairs: Group RAW vs. Group PE, Group RAW vs. Group HT, and Group PE vs. Group HT. This type of one-way ANOVA is carried out for all measures in the following sections. It has to be noted that there are limitations of multiple testing ANOVA with the same samples, it may increase the probability of achieving significant results or lead to Type 1 error (Vasey and Thayer, 1987; Goldman, 2008; Kim, 2017). Meanwhile, other testing methods considered for use, such as MANOVA (multivariate analysis of variance) also have their limitations, for example, the results obtained from MANOVA may be more ambiguous than those of ANOVA (Tavakoli and Gerami, 2013), because the effects that the multiple variables have on each other can be significant, thus may be confused with the effects of independent variables. However, ANOVA does not have this problem because only one dependent variable is used.

1) Group RAW vs. Group PE

Table 7.3 ANOVA for Glances Count (AOI_SUB) of Group RAW and Group PE

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW sub	13	1918	147.54	900.44
PE sub	11	1566	142.36	1036.45

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	159.56	1	159.56	0.17	0.69	4.30
Within Groups	21169.78	22	962.26			
Total	21329.33	23				

2) Group RAW vs. Group HT

Table 7.4 ANOVA for Glances Count (AOI_SUB) of Group RAW and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW sub	13	1918	147.54	900.44
HT sub	12	1609	134.08	756.99

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1129.69	1	1129.69	1.36	0.26	4.28
Within Groups	19132.15	23	831.83			
Total	20261.84	24				

3) Group PE vs. Group HT

Table 7.5 ANOVA for Glances Count (AOI_SUB) of Group PE and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
PE sub	11	1566	142.36	1036.45
HT sub	12	1609	134.08	756.99

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	393.49	1	393.49	0.44	0.51	4.32
Within Groups	18691.46	21	890.07			
Total	19084.96	22				

It can be seen that the p-value is greater than 0.05 and F is smaller than F crit in Tables 7.3, 7.4 and 7.5, indicating that there is no statistically significant difference between the glances count in AOI_SUB of the three group pairs.

Hypothesis 1b: More attention is allocated to the image area (AOI_IMA) when full PE subtitles and HT subtitles are displayed than when raw MT subtitles are displayed.

For this hypothesis, a comparison of glances count in the image AOI of each recording is conducted. As for Hypothesis 1b, two sub-hypotheses have been proposed as follows:

Hypothesis 1b.1: Group RAW < Group PE and Group HT

Hypothesis 1b.2: Group PE < Group HT

Table 7.6 Glances Count for Image AOI of Each Group

	Group PE		Group RAW		Group HT
P01	140	P25	169	P47	118
P02	146	P26	143	P48	119
P04	109	P28	175	P51	163
P06	146	P29	_	P52	_
P09	124	P30	89	P53	142
P10	184	P33	115	P54	168
P11	132	P35	126	P55	114
P12	135	P36	142	P56	139
P14	_	P37	187	P57	83
P17	168	P38	131	P58	168
P23	173	P39	162	P59	137
P24	81	P40	127	P60	104
		P43	157	P61	146
		P46	103		
Mean	139.82		140.46		133.42
SD	29.44		29.01		26.58

In Table 7.6, the mean of each group is as follows: Group RAW (140.46) > Group PE (139.82) > Group HT (133.42). This result is in contradiction to both Hypothesis 1b.1 and Hypothesis 1b.2, but is consistent with the results of the pilot study (see Section 5.2.4).

1) Group RAW vs. Group PE

Table 7.7 ANOVA for Glances Count (AOI_IMA) of Group RAW and Group PE

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW ima	13	1826	140.46	841.60
PE ima	11	1538	139.82	866.76

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2.47	1	2.47	0.003	0.96	4.30
Within Groups	18766.87	22	853.04			
Total	18769.33	23				

2) Group RAW vs. Group HT

Table 7.8 ANOVA for Glances Count (AOI_IMA) of Group RAW and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW ima	13	1826	140.46	841.60
HT ima	12	1601	133.42	706.63

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	309.69	1	309.69	0.40	0.53	4.28
Within Groups	17872.15	23	777.05			
Total	18181.84	24				

3) Group PE vs. Group HT

Table 7.9 ANOVA for Glances Count (AOI_IMA) of Group PE and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
PE ima	11	1538	139.82	866.76
HT ima	12	1601	133.42	706.63

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	235.19	1	235.19	0.30	0.59	4.32
Within Groups	16440.55	21	782.88			
Total	16675.74	22				

As with the result of the glances count in AOI_SUB, in Tables 7.7, 7.8 and 7.9, the p-value is greater than 0.05 and F is smaller than F crit, which means there is no statistically significant difference between the glances count in AOI_IMA of the three group pairs.

To gain a deeper understanding of this, the additional eye tracking measures of fixation count and glance duration for both AOIs for each group is shown in Table 7.10 and Table 7.17. Similarly to glances count, four sub-hypotheses for fixation count are proposed as follows:

Hypothesis 1c.1: AOI_SUB: Group RAW > Group PE and Group HT

Hypothesis 1c.2: AOI_SUB: Group PE > Group HT

Hypothesis 1d.1: AOI_IMA: Group RAW < Group PE and Group HT

Hypothesis 1d.2: AOI_IMA: Group PE < Group HT

Table 7.10 Fixation Count for Both AOIs for Each Group

Group PE			Group RAW			Group HT		
	AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA
P01	758	492	P25	670	501	P47	679	367
P02	959	370	P26	636	431	P48	334	509
P04	800	270	P28	666	416	P51	687	536
P06	464	573	P29	_	_	P52	_	_
P09	390	468	P30	809	206	P53	372	458
P10	656	637	P33	873	275	P54	176	530
P11	1051	331	P35	712	289	P55	930	292
P12	885	300	P36	660	450	P56	810	411
P14	_	_	P37	617	495	P57	520	174
P17	733	450	P38	513	287	P58	454	333
P23	829	484	P39	712	383	P59	353	692
P24	764	186	P40	722	297	P60	531	246
			P43	601	331	P61	359	413
			P46	797	208			
Mean	753.55	414.64		691.38	351.46		517.08	413.42
SD	195.42	135.92		95.93	101.32		221.23	143.36

Table 7.10 shows that regarding fixation count, Group PE had the highest value in AOI_SUB (753.55), which contradicts Hypothesis 1c.1. Group PE had a higher value than Group HT (517.08), which supports Hypothesis 1c.2. Regarding the image area, Group

RAW had the lowest value (351.46), which supports Hypothesis 1d.1. However, Group PE (414.64) had a slightly higher value than Group HT (413.42).

1) Group RAW vs. Group PE

Table 7.11 ANOVA for Fixation Count (AOI_SUB) of Group RAW and Group PE

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW sub	13	8988	691.38	9203.09
PE sub	11	8289	753.55	38189.07

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	23022.82	1	23022.82	1.03	0.32	4.30
Within Groups	492327.80	22	22378.54			
Total	515350.63	23				

Table 7.12 ANOVA for Fixation Count (AOI_IMA) of Group RAW and Group PE

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW ima	13	4569	351.46	10265.77
PE ima	11	4561	414.64	18474.25

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	23780.06	1	23780.06	1.70	0.21	4.30
Within Groups	307931.78	22	13997.90			
Total	331711.83	23				

In both Tables 7.11 and 7.12, the p-value is greater than 0.05 and F is smaller than F crit, demonstrating no statistically significant difference for fixation count in both areas of Group RAW and Group PE.

2) Group RAW vs. Group HT

Table 7.13 ANOVA for Fixation Count (AOI_SUB) of Group RAW and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW sub	13	8988	691.38	9203.09
HT sub	12	6205	517.08	48944.63

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	189577.05	1	189577.05	6.72	0.02	4.28
Within Groups	648827.99	23	28209.91			
Total	838405.04	24				

Table 7.14 ANOVA for Fixation Count (AOI_IMA) of Group RAW and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW ima	13	4569	351.46	10265.77
HT ima	12	4961	413.42	20551.72

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	23951.85	1	23951.85	1.58	0.22	4.28
Within Groups	349258.15	23	15185.14			
Total	373210	24				

Table 7.13 shows that the p-value is smaller than 0.05 and F is greater than F crit, while Table 7.14 shows just the opposite, indicating that there is a statistically significant difference between the fixation count in AOI_SUB of Group RAW and Group HT, but not in AOI_IMA.

3）Group PE vs. Group HT

Table 7.15 ANOVA for Fixation Count (AOI_SUB) of Group PE and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
PE sub	11	8289	753.55	38189.07
HT sub	12	6205	517.08	48944.63

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	320899.66	1	320899.66	7.32	0.01	4.32
Within Groups	920281.64	21	43822.94			
Total	1241181.30	22				

Table 7.16 ANOVA for Fixation Count (AOI_IMA) of Group PE and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
PE ima	11	4561	414.64	18474.25
HT ima	12	4961	413.42	20551.72

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8.54	1	8.54	0.00044	0.98	4.32
Within Groups	410811.46	21	19562.45			
Total	410820	22				

Table 7.15 shows that p-value is smaller than 0.05 and F is higher than F crit, which means the fixation count in AOI_SUB of Group PE is statistically different from that of Group HT. Meanwhile, Table 7.16 shows that p-value is higher than 0.05 and F is smaller than F crit, indicating that the fixation count in AOI_IMA of Group PE is not statistically different from that of Group HT.

In summary, the only two statistically significant differences for the fixation count measure were for the subtitle AOI RAW vs HT and PE vs HT, with PE subtitles having the highest fixation count on average.

Table 7.17 Glance Duration [s] in both AOIs of Each Group

	Group PE			Group RAW			Group HT	
	AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA
P01	207.20	174.27	P25	194.36	139.90	P47	240.44	156.36
P02	264.03	120.87	P26	189.18	141.69	P48	120.34	276.36
P04	274.53	124.24	P28	183.58	162.40	P51	211.62	174.85
P06	131.73	236.04	P29	_	_	P52	_	_
P09	109.79	199.47	P30	287.82	79.25	P53	101.88	148.89
P10	174.45	230.13	P33	296.80	85.02	P54	44.53	149.88
P11	318.21	97.08	P35	264.83	115.54	P55	311.00	92.68
P12	271.81	126.92	P36	226.88	163.95	P56	210.14	138.31
P14	_	_	P37	186.88	205.32	P57	203.89	73.74
P17	172.17	108.34	P38	247.07	165.10	P58	175.12	192.74
P23	227.81	172.67	P39	226.06	148.74	P59	85.87	253.84
P24	279.04	91.61	P40	230.58	120.87	P60	235.81	167.04
			P43	193.66	131.12	P61	120.55	268.46
			P46	311.13	70.07			
Mean	220.98	152.88		233.76	133.00		171.77	174.43
SD	67.26	52.22		44.73	38.80		77.47	64.41

As for glance duration, four hypotheses have been proposed as follows:

Hypothesis 1e.1: AOI_SUB: Group RAW > Group PE and Group HT

Hypothesis 1e.2: AOI_SUB: Group PE > Group HT

Hypothesis 1f.1: AOI_IMA: Group RAW < Group PE and HT

Hypothesis 1f.2: AOI_IMA: Group PE < Group HT

Table 7.17 shows that regarding mean glance duration, Group RAW had the highest value in AOI_SUB (233.76), and the lowest value in AOI_IMA (133.00), which supports Hypothesis 1e.1 and Hypothesis 1f.1. Group HT had the highest value in AOI_IMA

(174.43), and the lowest value in AOI_SUB (171.77), which supports Hypotheses 1e.2, 1f.1 and 1f.2.

1) Group RAW vs. Group PE

Table 7.18 ANOVA for Glance Duration (AOI_SUB) of Group RAW and Group PE

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW sub	13	3038.83	233.76	2000.45
PE sub	11	2430.77	220.98	4524.20

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	972.72	1	972.72	0.31	0.58	4.30
Within Groups	69247.40	22	3147.61			
Total	70220.11	23				

Table 7.19 ANOVA for Glance Duration (AOI_IMA) of Group RAW and Group PE

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW ima	13	1728.97	133.00	1505.05
PE ima	11	1681.64	152.88	2727.32

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2354.50	1	2354.50	1.14	0.30	4.30
Within Groups	45333.81	22	2060.63			
Total	47688.31	23				

Tables 7.18 and 7.19 show that the p-value is greater than 0.05 and F is smaller than F crit. Therefore, the glance duration in both areas of Group RAW is not statistically different from that of Group PE.

2) Group RAW vs. Group HT

Table 7.20 ANOVA for Glance Duration (AOI_SUB) of Group RAW and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW sub	13	3038.83	233.76	2000.45
HT sub	12	2061.19	171.77	6002.44

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	23979.07	1	23979.07	6.13	0.02	4.28
Within Groups	90032.25	23	3914.45			
Total	114011.32	24				

Table 7.21 ANOVA for Glance Duration (AOI_IMA) of Group RAW and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW ima	13	1728.97	133.00	1505.05
HT ima	12	2093.15	174.43	4148.64

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	10711.38	1	10711.38	3.87	0.06	4.28
Within Groups	63695.65	23	2769.38			
Total	74407.03	24				

Table 7.20 shows that the p-value is smaller than 0.05 and F is higher than F crit, while Table 7.21 shows just the opposite, which means that the glance duration in AOI_SUB of Group RAW is statistically different from that of Group HT, but not in AOI_IMA.

3) Group PE vs. Group HT

Table 7.22 ANOVA for Glance Duration (AOI_SUB) of Group PE and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
PE sub	11	2430.77	220.98	4524.20
HT sub	12	2061.19	171.77	6002.44

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	13899.86	1	13899.86	2.62	0.12	4.32
Within Groups	111268.86	21	5298.52			
Total	125168.71	22				

Table 7.23 ANOVA for Glance Duration (AOI_IMA) of Group PE and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
PE ima	11	1681.64	152.88	2727.32
HT ima	12	2093.15	174.43	4148.64

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2665.96	1	2665.96	0.77	0.39	4.32
Within Groups	72908.30	21	3471.82			
Total	75574.26	22				

In both Tables 7.22 and 7.23, p-value is greater than 0.05 and F is smaller than F crit, which means that there is no significant difference between the glance duration in both areas of Group PE and that of Group HT.

In summary, the only statistically significant difference for the glance duration measure lied in the subtitle AOI RAW vs HT, with RAW subtitles having the highest glance duration on average, and RAW image having the lowest.

Regarding "reaction":

Hypothesis 3: Average fixation duration is shorter when full PE subtitles and HT subtitles are displayed than when RAW subtitles are displayed.

Table 7.24 Average Fixation Duration [ms] in Both AOIs of Each Group

	Group PE			Group RAW			Group HT	
	AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA
P01	242	319.7	P25	264.2	227.6	P47	319.9	390.1
P02	250.4	285.9	P26	270.2	285.1	P48	333.7	517.9
P04	318.7	426	P28	250.7	359	P51	202.7	239.9
P06	250.3	376.3	P29	_	_	P52	_	_
P09	235.6	388.7	P30	320.5	336.4	P53	236.7	277.1
P10	230.4	311.7	P33	321.2	275.3	P54	209.8	228.1
P11	272	248.6	P35	347.5	364.9	P55	304.9	276.2
P12	283	383.7	P36	315.9	331.4	P56	234.8	300.2
P14	_	_	P37	273.3	378.4	P57	371.7	401.1
P17	177.2	180.9	P38	461	543.9	P58	362.6	551.7
P23	244.5	321.5	P39	292.3	358.7	P59	215.7	333
P24	344.5	450.9	P40	297.4	374.3	P60	422.5	646
			P43	302.9	368.2	P61	310.1	622
			P46	367.2	306.1			
Mean	258.94	335.81		314.18	346.87		293.76	398.61
SD	45.03	79.53		55.22	74.61		72.75	149.80

It can be seen from Table 7.24 that regarding average fixation duration, Group RAW had the highest value in AOI_SUB (314.18), while Group PE had the lowest (258.94). Group HT had the highest value in AOI_IMA (398.61), while Group PE had the lowest (335.81).

Hypothesis 3a.1: AOI_SUB: Group RAW > Group PE and Group HT

(Hypothesis supported)

Hypothesis 3a.2: AOI_SUB: Group PE > Group HT

(Result: Group PE < Group HT)

Hypothesis 3b.1: AOI_IMA: Group RAW > Group PE and Group HT

(Result: Group HT > Group RAW > Group PE)

Hypothesis 3b.2: AOI_IMA: Group PE > Group HT

(Result: Group PE < Group HT)

1) Group RAW vs. Group PE

Table 7.25 ANOVA for Average Fixation Duration (AOI_SUB) of Group RAW and Group PE

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW sub	13	4084.3	314.18	3049.19
PE sub	11	2848.3	258.94	2027.37

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	18181.97	1	18181.97	7.03	0.01	4.30
Within Groups	56864.05	22	2584.73			
Total	75046.02	23				

Table 7.26 ANOVA for Average Fixation Duration (AOI_IMA) of Group RAW and Group PE

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW ima	13	4509.3	346.87	5567.40
PE ima	11	3693.9	335.81	6324.95

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	728.86	1	728.86	0.12	0.73	4.30
Within Groups	130058.26	22	5911.74			
Total	130787.12	23				

Table 7.25 shows that the p-value is smaller than 0.05 and F is higher than F crit, while Table 7.26 shows just the opposite, which means that the average fixation duration in AOI_SUB of Group RAW is statistically different from that of Group PE, but not in AOI_IMA.

2) Group RAW vs. Group HT

Table 7.27 ANOVA for Average Fixation Duration (AOI_SUB) of Group RAW and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW sub	13	4084.3	314.18	3049.19
HT sub	12	3525.1	293.76	5293.08

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2601.57	1	2601.57	0.63	0.44	4.28
Within Groups	94814.17	23	4122.36			
Total	97415.75	24				

Table 7.28 ANOVA for Average Fixation Duration (AOI_IMA) of Group RAW and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
RAW ima	13	4509.3	346.87	5567.40
HT ima	12	4783.3	398.61	22440.98

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	16704.07	1	16704.07	1.23	0.28	4.28
Within Groups	313659.56	23	13637.37			
Total	330363.63	24				

In both Tables 7.27 and 7.28, the p-value is greater than 0.05 and F is smaller than F crit, which means that the average fixation duration in both areas of Group RAW is not statistically different from that of Group HT.

3) Group PE vs. Group HT

Table 7.29 ANOVA for Average Fixation Duration (AOI_SUB) of Group PE and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
PE sub	11	2848.3	258.94	2027.37
HT sub	12	3525.1	293.76	5293.08

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6959.09	1	6959.09	1.86	0.19	4.32
Within Groups	78497.61	21	3737.98			
Total	85456.71	22				

Table 7.30 ANOVA for Average Fixation Duration (AOI_IMA) of Group PE and Group HT

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
PE ima	11	3693.9	335.81	6324.95
HT ima	12	4783.3	398.61	22440.98

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	22633.67	1	22633.67	1.53	0.23	4.32
Within Groups	310100.28	21	14766.68			
Total	332733.94	22				

We can see that the p-value is greater than 0.05 and F is smaller than F crit in both Table 7.29 and Table 7.30, indicating that there is no significant difference between the average fixation duration in both areas of the two groups.

In summary, the only statistically significant difference for the average fixation duration measure were for the subtitle AOI RAW vs PE, with HT image having the highest average fixation duration on average and PE subtitles having the lowest in both areas.

7.4 Outliers

As mentioned in testing Hypothesis 1a, regarding the data of glances count in AOI_SUB, P14 in Group PE, P29 in Group RAW and P52 in Group HT are outliers. In order to explain why they had such low visits to the subtitles, more analysis of the results of their experimental data (Table 7.31) is made in this section.

Table 7.31 Measurement data of P14, P29 and P52

	Glances Count		Fixation Count		Glance Duration [s]		Average Fixation Duration [ms]		Comprehension Testing Score (Full score: 13)	English Score (Full score: 25)	Tracking Ratio [%]
	SUB	IMA	SUB	IMA	SUB	IMA	SUB	IMA			
P14 (PE)	48	50	99	364	20.62	272.63	188.1	714.4	11	15	73.76
P29 (RAW)	62	56	642	88	328.82	29.79	494.9	301	7	13	77.18
P52 (HT)	24	98	67	469	20.46	166.52	158.2	164.3	12	19	80.26

Table 7.31 shows that a majority of the eye-tracking data for the three participants are outliers, which is highlighted in red. However, the unusual eye-tracking data does not seem to be accompanied by a low comprehension score. Both P14 and P52 scored well in comprehension testing, indicating that their understanding of the video was not impacted by their unusual activity for the two AOIs. Their comprehension scores are consistent with their English scores. From this fact, it can be inferred provisionally that participants' comprehension of the video content was affected by their English

proficiency level, rather than their visits to different AOIs. This provisional conclusion will be discussed in the following section.

When watching the recordings of the three participants, it has been found that P14 was not actually 'reading' the subtitles. His eyes were not moving from left to right on the subtitles, but mostly just glimpsed the latter part of the subtitles. Besides, the eye-tracker lost track of his eyes several times. As for P29, he was reading the subtitles from left to right. It can be seen obviously from the recording that his eyes were focusing on AOI_SUB most of the time, which is also reflected in his fixation count and glance duration data. This is perhaps because his English level was low, thus he had to focus on the subtitles. As with P14, when reading the subtitles, P52 focused on the latter part mostly. In the beginning of the video, she often looked at the subtitle area, while in the middle and latter part of the video, she looked at the image area more. This is also shown in her data of glances count, fixation count and glance duration. This is perhaps because her English level was relatively high, thus she focused more on the subtitle area in the beginning to get a general idea about the video. When she became used to it, she focused more on the image area. In short, it seems that all the three participants adopted unusual viewing strategies, which may not be enough, but at least help to explain the outlier data.

7.5 Discussion

Table 7.32 shows the hypothesized relationships between the three groups for the comprehension testing and attitude survey. The symbol "√" means the result supports

the corresponding hypothesis, and the symbol “x” means the result does not support the corresponding hypothesis. We can see that most of the hypotheses are rejected by the results. As mentioned in Section 6.4, Group PE performed the best in both comprehension testing and attitude survey, while the relationship between Group RAW and Group HT is inconsistent in the two tasks. It has to be noted that the attitude survey of all groups reveals that most participants had a positive attitude towards their subtitles (see Table 6.5).

Table 7.32 Results of comprehension testing and attitude survey

	Comprehension testing	Attitude survey
RAW vs. PE and HT	< √	< x
PE vs. HT	< x	< x

Table 7.33 presents a summary for all results of ANOVA and means. The letter “Y” means there is a statistically significant difference between the two groups, while “N” means the opposite. The hypothesized relationships between the means of the three groups for different measures are also presented in the table. The symbol “√” means the result supports the corresponding hypothesis, and the symbol “x” means the result does not support the corresponding hypothesis.

Table 7.33 Summary for results of ANOVA and means

		AOI	Glances Count	Fixation Count	Glance Duration	Average Fixation Duration
ANOVA	RAW vs. PE	SUB	N	N	N	Y
		IMA	N	N	N	N
	RAW vs. HT	SUB	N	Y	Y	N
		IMA	N	N	N	N
	PE vs. HT	SUB	N	Y	N	N
		IMA	N	N	N	N
Means	RAW vs. PE and HT	SUB	> √	> x	> √	> √
		IMA	< x	< √	< √	> x
	PE vs. HT	SUB	> √	> √	> √	> x
		IMA	< x	< x	< √	> x

In a word, the researcher tests 20 hypotheses in total, 16 based on eye-tracking data and four on questionnaire data. Table 7.33 shows that among the 16 hypotheses, seven are not supported by the results (three in RAW vs. PE and HT, four in PE vs. HT). Only one of the remaining four hypotheses is supported by the results (see Table 7.32). According to the ANOVA results, there are no statistically significant differences between the three group pairs on most measures.

In regard to Group RAW vs. Group PE, the only statistically significant difference lies in the average fixation duration in AOI_SUB. The mean of average fixation duration in AOI_SUB for Group PE is 258.94, while that for Group RAW is 314.18. According to the means, most hypotheses are supported by the results, except the glance count in AOI_IMA and the fixation count in AOI_SUB. As mentioned in Section 6.4.1.2, there is a statistically significant difference between the comprehension testing score of the two

groups, and Group PE outperformed Group RAW at both comprehension test and attitude survey. Relating all the results to the three 'R's in the reception model, we can see that Group PE performed better in 'Reaction' and 'Repercussion' than Group RAW. In regard to 'Response', Group PE performed partially better than Group RAW. Therefore, it can be concluded that, overall, participants who were offered full PE subtitles scored better on the reception metrics than those who were offered raw MT subtitles.

Group RAW and Group HT had statistically significant differences in only two of the measures. Group HT outperformed Group RAW in comprehension testing, but their scores are not significantly different (see Sections 6.4.1 and 6.4.1.2). The attitude survey result suggests that Group RAW had a better attitude towards the subtitles than Group HT (see Section 6.4.2), which goes against the hypothesis.

Group PE and Group HT return no statistically significant differences in most cases, the only exception is the fixation count in AOI_SUB. According to the means, Group PE outperformed Group HT in half of the measures. In addition, Group PE scored better than Group HT in comprehension testing, though the difference was not significant (see Section 6.4.1.2). Also, Group PE had a better attitude towards the subtitles than Group HT (see Section 6.4.2). Undoubtedly, this result confounds expectations.

It can be seen that Group HT, which was expected to have higher reception metrics than Group PE and Group RAW, returned some results that do not support hypotheses. It has to be emphasised that all the hypotheses were built upon the premise that the quality

of the subtitles would increase as we go from RAW to PEMT to HT. This premise was built on the intuition that human translation would outperform MT with PE, if the human translator is a good one. It has to be noted that the human translator is an English teacher rather than professional translator, and there is of course no guarantee that non-professional human translation can reach a quality level equal to that of a professional. For further analysis, a quality assessment for full PE subtitles and HT subtitles is imperative.

In addition, as mentioned in Section 7.4, a provisional conclusion is that participants' comprehension of the video content was affected by their English proficiency level rather than their visits to different AOIs. Hence, it has been decided to test if this conclusion applies to the unexpected results of the three groups. More specifically, the researcher would like to find out if participants' comprehension level was actually correlated to their English level. Table 7.34 presents the English scores and comprehension scores of all participants.

Table 7.34 English scores (full score: 25) and comprehension scores (full score: 13) of all participants (n=61)

Group PE	English Score	Comprehension Score	Group RAW	English Score	Comprehension Score	Group HT	English Score	Comprehension Score
P01	12	11	**P25**	12	7	**P47**	15	9
P02	18	12	**P26**	21	8	**P48**	18	11
P03	16	11	**P27**	19	11	**P49**	21	7
P04	14	6	**P28**	8	9	**P50**	14	10
P05	15	11	**P29**	13	11	**P51**	20	12
P06	9	13	**P30**	16	7	**P52**	19	8
P07	19	9	**P31**	12	10	**P53**	9	6
P08	14	8	**P32**	14	8	**P54**	20	9
P09	15	10	**P33**	15	10	**P55**	11	10
P10	19	6	**P34**	22	9	**P56**	18	9
P11	9	9	**P35**	17	7	**P57**	14	12
P12	17	8	**P36**	20	8	**P58**	19	11
P13	15	8	**P37**	16	8	**P59**	18	10
P14	15	10	**P38**	20	10	**P60**	13	7
P15	14	11	**P39**	17	9	**P61**	18	11
P16	11	10	**P40**	15	9			
P17	11	10	**P41**	15	7			
P18	8	11	**P42**	15	8			
P19	13	9	**P43**	11	6			
P20	7	11	**P44**	13	8			
P21	12	10	**P45**	11	10			
P22	11	8	**P46**	12	8			
P23	13	10						
P24	10	8						

Putting the English scores and comprehension scores of all participants into Excel and calculating the Pearson's correlation coefficient between them, the result turned out to

be -0.048, which was close to zero, indicating that participants' English level and comprehension level were not correlated with each other.

Regarding Hypotheses 1b.1 and 1b.2, they are rejected by the results of glances count in AOI_IMA. The reason could be that the raw MT subtitles had a lower quality than the other two types, so participants had to keep visiting the image area for help, while the image could not provide enough information for them to comprehend the MOOC content, they also had to keep visiting the subtitles. This point is verified by the fact that Group RAW had the lowest number of fixation counts in both AOIs, but they had longer fixation durations, as is shown in the results for average fixation duration. In regard to Group PE and Group HT, results show that Group PE had a higher number of glances count in both AOIs. As discussed in Section 7.3.1, the premise for our hypotheses is that the quality of full PE subtitles is lower than HT subtitles. If we use the same way of thinking when explaining the results of Group RAW, Group PE should have a lower fixation count in both areas than Group HT, and longer average fixation duration in both areas than Group HT. However, the result was just the opposite. Group PE had a higher fixation count in both areas than Group HT, and shorter average fixation duration in both areas than Group HT. Therefore, although Group PE kept visiting both areas, they did not spend much effort processing the subtitles or image, plus this group performed the best in comprehension testing (see Section 6.4.1.1), possibly demonstrating the high quality of full PE subtitles.

The researcher speculated that if the number of Chinese characters in the subtitles provided for Group PE was more, it would lead to a higher fixation count. As mentioned

in Section 5.3.1, the number of full PE subtitles was 138 and that of HT subtitles was 141. After eliminating the last three lines of HT subtitles and comparing the number of characters for each of the 138 lines for the two groups, it was found that 56 lines of full PE subtitles had more characters than HT subtitles, and 51 full PE subtitles had fewer characters than HT subtitles. Table 7.35 shows the ANOVA result for the number of characters of full PE subtitles and HT subtitles. It can be seen that p-value is greater than 0.05, and F is smaller than F crit, thus there is no statistically significant difference between the two types of subtitles. The mean character count for full PE subtitles is 11.02 (SD = 4.67), while that for HT subtitles is 10.99 (SD = 4.62). This tiny difference cannot lead to the conclusion that full PE subtitles had more characters in each line than HT subtitles. Hence, the researcher's speculation cannot be confirmed by this.

Table 7.35 ANOVA for the number of Chinese characters of PE subtitles and HT subtitles

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
PE	138	1521	11.02	21.79
HT	138	1516	10.99	21.31

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.09	1	0.09	0.0042	0.95	3.88
Within Groups	5904.91	274	21.55			
Total	5905.00	275				

On the whole, results show that participants who were offered full PE subtitles scored better on the reception metrics than those who were offered raw MT subtitles, but not significantly. As mentioned in Section 5.2.2, the quality of the raw MT subtitles was relatively good. In regard to the participants who were offered HT subtitles, they did not perform better than the other two groups, which is certainly not in support of the hypotheses. It has to be emphasized that all the hypotheses were built upon the premise that the quality of raw MT subtitles was lower than full PE subtitles, which was lower than HT subtitles. If the quality of PE and HT subtitles was about the same, then it might help explain the unexpected results on the reception metrics. Therefore, in order to verify the premise, a quality assessment for HT subtitles and PE subtitles was conducted and is presented in the following chapter. Notwithstanding the unexpected results, most participants held a positive attitude towards the subtitles regardless of their type, and most participants in all groups did well in the comprehension testing. This means MT can help in translating English subtitles for MOOCs into Chinese, and may contribute to the development of MOOCs in China in the long term.

While the current Chapter has concentrated on participants' reception of subtitles, based on their behaviour when presented with those subtitles, Chapter 8 thus looks in greater detail at the stimulus itself. In other words, Chapter 8 drills down into the subtitles as translations (in the TQA) but also as examples of Chinese (in an associated frequency analysis), in a bid to find explanations for the findings in this Chapter.

Chapter 8 Findings and Data Analysis III – TQA and Frequency Analysis

8.1 Introduction

As mentioned in Section 4.2, TQA and frequency analysis are supplementary methods in this research. They were used to explore the difference between human translated subtitles and post-edited machine translated subtitles. In this chapter, Section 8.2 focuses on TQA and Section 8.3 on frequency analysis.

8.2 TQA

This chapter includes an introduction to the TQA procedure, a description and a discussion of the results, as well as references to the inter-annotator agreements.

8.2.1 TQA procedure

The ten eligible participants read and agreed with the plain language statement and informed consent form for this research. After that, they completed an online questionnaire that elicited demographic information. Finally, they were asked to read the guidelines first and then conduct the quality assessment for subtitles in Excel (see Figure 5.2). The guidelines included the definitions of all metrics and severity levels with

examples, and steps for conducting QA (see Appendix F1). Each evaluator saw one HT subtitle and one PE subtitle for the same source text. As mentioned before, the subtitles were presented in randomized order, thus the participant would not know which one was HT or PE. For each sentence, they were asked to locate any errors, identify its error type (see Table 5.9 for the ten error types) and severity level (minor, major, critical) from the drop-down menu, and tick the better translation of the two. They could also add comments. Participants who completed the full assessment were reimbursed with vouchers (approximately 20 euro).

Table 8.1 English score of the eligible evaluators

Evaluator	English score (Full score: 25)
P01	22
P02	21
P03	23
P04	22
P05	21
P06	19
P07	19
P08	20
P09	24
P10	20

All ten evaluators are women. Figures 8.1, 8.2 and 8.3 show the demographic profile of the evaluators. Six of them were aged 22. Also, six were 1st year Master students who were majoring in Mass Media Translation. As for these six Master students, half of them hold a Bachelor's degree in English and the other half hold a Bachelor's degree in Communications. Regarding the other four evaluators who were final year undergraduate students, only one of them did not major in English (International

Business), but her English score was the highest (24 points). Thus after some guidance, it was presumed that she was capable of completing the quality assessment.

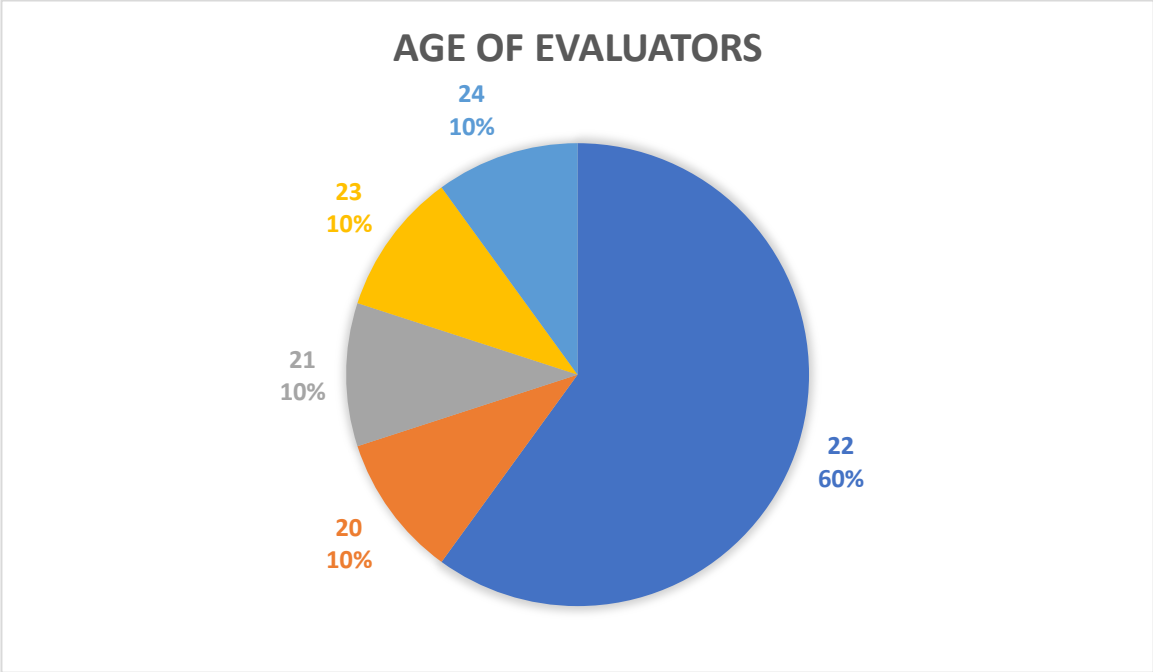

Figure 8.1 Age of evaluators

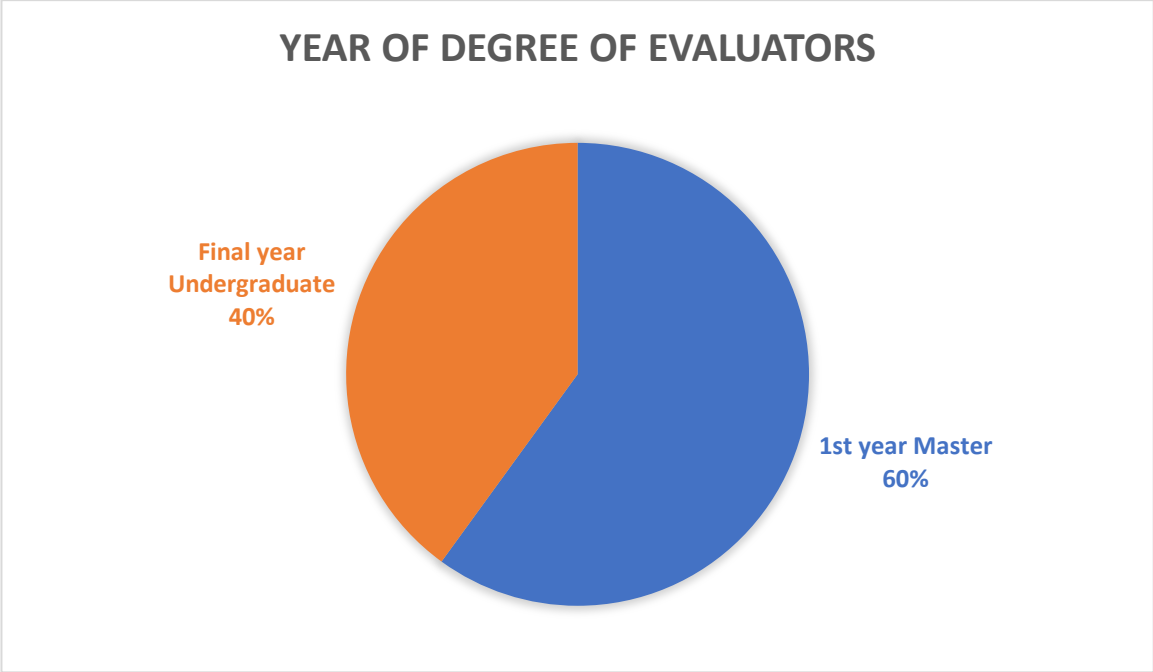

Figure 8.2 Year of degree of evaluators

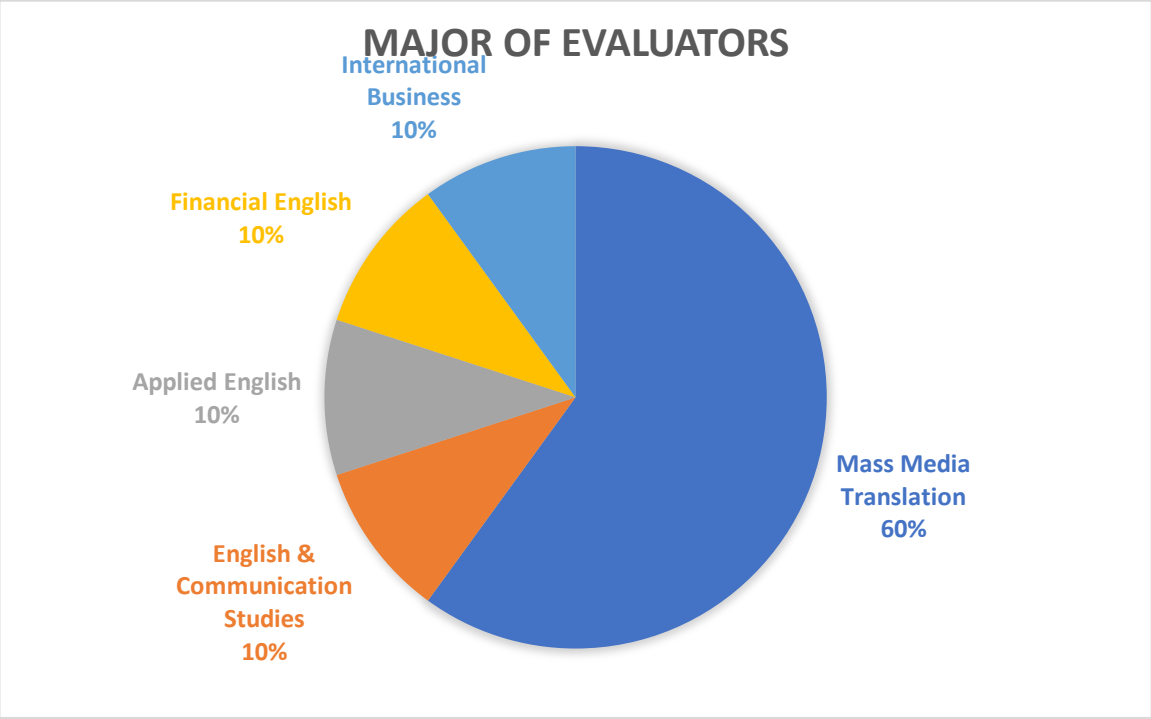

Figure 8.3 Major of evaluators

8.2.2 Inter-annotator agreement

Considering that evaluation is a subjective process, and the ten evaluators come from different backgrounds, before conducting the data analysis of their completed QA files, the inter-annotator agreement (IAA), needs to be determined.

Microsoft Excel was used for calculating Fleiss' kappa in this research. The result is 0.2413, which means the ten evaluators had fair agreement with each other. Results show that IAA is greater for PE than for HT, because for evaluating HT subtitles, $\kappa = 0.197$, corresponding to slight agreement, while for PE subtitles, $\kappa = 0.2856$, corresponding to fair agreement. It has been decided to use the average score of the two as the final score.

Given the complexity of the problem and relatively low professional level of the novices, this low number was to be expected. Nevertheless, it has to be admitted that the inter-annotator agreement for this quality assessment is rather low.

8.2.3 Results

Table 8.2 shows the QA scores given and errors identified by the ten evaluators. Regarding the QA scores, the scoring algorithm has already been clarified in Section 5.3.3. An example (P01) is presented here as a further explanation:

Because:

TQ = 100 – TP

TP =

$$\frac{\text{Issues}_{\text{minor}} + \text{Issues}_{\text{major}} \times \text{SeverityMultiplier}_{\text{major}} + \text{Issues}_{\text{critical}} \times \text{SeverityMultiplier}_{\text{critical}}}{\text{Word count (target)}}$$

Therefore:

$$\text{QA score of P01 (HT)} = 100 - \frac{70 + (10 \times 10) + (5 \times 100)}{1685} = 99.6024$$

Note: the number of words for HT subtitles is 1685, and that for PE subtitles is 1667 (including individual Chinese characters and punctuation).

Table 8.2 QA scores given and errors identified by the ten evaluators

	TYPE OF SUBTITLES	SCORE	ERROR TYPE									
			Addition	Mistranslation	Overly-literal	Omission	Grammar	Spelling	Style	Locale-convention	Terminology	Others
P01	**HT** 55	99.602	12	3	25	16	1	1	—	24	1	1
	PE 49	99.713	7	4	37	8	2	—	—	15	—	—
P02	**HT** 53	99.852	12	2	20	18	1	2	—	23	—	2
	PE 45	99.862	3	4	22	10	2	—	1	18	—	—
P03	**HT** 23	99.975	2	1	11	—	—	—	3	8	—	—
	PE 34	99.895	1	2	26	1	—	—	1	6	—	3
P04	**HT** 11	99.995	—	—	5	1	1	—	—	4	—	—
	PE 14	99.987	—	—	10	—	—	—	1	4	—	—
P05	**HT** 18	99.969	2	5	4	3	—	—	—	—	1	2
	PE 12	99.992	—	2	2	2	1	—	1	—	2	3
P06	**HT** 15	99.954	—	3	1	3	2	1	—	5	—	—
	PE 21	99.831	1	2	6	2	2	—	—	5	2	1
P07	**HT** 08	99.995	—	—	5	3	—	—	—	—	—	—
	PE 14	99.992	1	1	10	—	2	—	—	—	—	—
P08	**HT** 23	99.763	2	5	11	4	10	—	—	—	—	—
	PE 18	99.854	—	5	7	—	5	—	—	—	—	1
P09	**HT** 21	99.667	—	7	12	2	9	3	2	—	4	—
	PE 27	99.967	—	1	13	3	5	2	1	—	1	1
P10	**HT** 32	99.901	1	1	19	3	—	—	—	5	2	1
	PE 32	99.876	1	2	21	—	2	—	—	10	—	1

In Table 8.2, the top one and two error types are highlighted in red and green respectively. We can see that “Overly-literal” is the most common error type for both subtitles, and “Locale-convention” is the second most common one, while “Spelling”, “Style” and “Terminology” are the three least common error types.

HT subtitles have been given a higher score (not significantly) than PE subtitles in half of the cases (i.e. P03 (99.975 > 99.895), P04 (99.995 > 99.987), P06 (99.954 > 99.831), P07 (99.995 > 99.992) and P10 (99.901 > 99.876), while the opposite is true for the other half. In the column “Type of subtitles”, the numbers in the cells refer to the number of sentences in the subtitles that have errors. Based on this number, it can be seen that for four evaluators (P01, P02, P05, and P08), more HT subtitles have errors, while for five evaluators (P03, P04, P06, P07, and P09), more PE subtitles have errors. When comparing this result with the score, they are mostly consistent with each other except in two cases, P09 (21 < 27, 99.667 < 99.967) and P10 (32 = 32, 99.901 > 99.876). For P09, more sentences in PE subtitles have errors, but the score for PE subtitles is higher. For P10, in both types of subtitles, 32 sentences have errors, but the score for HT subtitles is higher. After scrutinizing the data of the two evaluators, five errors in HT subtitles were annotated as “critical” by P09, while there was no “critical” error in PE subtitles, which impacted the score. In the case of P10, eight errors in PE subtitles were annotated as “major” while for HT subtitles, only four errors were “major”, the number for other severity levels were more or less the same.

Rather than looking at the scores on an individual evaluator basis, a group level comparison might give us more insight. The mean and median QA score for both subtitles are presented in Table 8.3.

Table 8.3 Mean and median QA score for HT and PE subtitles

QA scores	Mean	Median
HT	99.867	99.928
PE	99.897	99.886

The mean score for HT subtitles is very slightly lower than for PE subtitles, while it is the opposite for the median score. This indicates that the distribution of the score given by the ten evaluators to HT subtitles is more uneven than what they gave to PE subtitles.

8.2.4 Discussion

According to the analysis of the data collected from the main experiment (see Section 7.5), it has been found that participants who were offered full PE subtitles scored almost the same on our reception metrics as those who were offered HT subtitles based on eye-tracking, comprehension testing and an attitude survey. As a way to investigate further, a quality assessment for both subtitles was carried out by ten novice evaluators. Results show that when comparing the QA scores, HT subtitles have been given a higher score than PE subtitles by half the evaluators; when comparing the number of sentences with errors, more HT subtitles have errors than PE subtitles according to four evaluators, and more PE subtitles have errors than HT subtitles according to five evaluators. The slight

difference between the mean score of HT and PE subtitles cannot infer anything. All in all, it seems that there is effectively no difference in quality between the HT and PE subtitles, which on one hand contradicts the premise of the hypotheses (as mentioned in Section 7.3.1, the premise of all the hypotheses in this research is that the quality of raw MT subtitles was lower than PE subtitles, which was lower than HT subtitles.), but on the other hand reinforces the results of the eye-tracking and questionnaires.

8.3 Frequency analysis

In this section I present a frequency analysis of the ten sentences with the lowest BLEU scores according to the analysis presented in Section 5.2.2. As mentioned in Section 4.2.4, the frequency analysis is just a supplementary method for further understanding the TQA results, and the lower the BLEU score, the greater the difference between raw MT and the reference translation. Therefore, only ten sentences with the greatest difference between PE and HT were compared in order to obtain the most obvious results. For ease of comparison, corresponding strings under analysis in the original, human translated and post-edited MT subtitles are colour coded. As mentioned in Section 4.2.4, the strings that are chosen for analysis are those for which an English word or phrase is translated differently in the human translation and post-edited MT and plays an important role in differentiating the two translations. It has to be noted that the segmentation for this analysis is based on the English word or phrase, rather than on Chinese n-grams, the length of which can differ. For example, in Sentence No.81, the PE output for “move about” is “动一动” (1-gram), while the HT output is “四处走走” (2-gram).

The number in parentheses beside each highlighted string refers to the number of occurrences of that string in the SogouT or OpenSubtitles corpus. The relationship between the number of occurrences of the string in the HT subtitles and the PE subtitles is shown in the "Notes" Table that accompanies each example. The number in the column for each selected string (pink, yellow or blue) indicates how many times more frequent the larger value is than the smaller value, which presents a clearer picture of their difference. For example, in Sentence No.50 (Example 8.1), for the string "would also like", the number of occurrences of the PE output in SogouT corpus is 8632 while that for the HT output is 218. 8632 ÷ 218 ≈ 39.60, so the occurrence of the PE output is 39.60 times more than that of the HT output for the string "would also like" in SogouT corpus. When comparing the relationship between the corresponding outputs in the two conditions (HT vs PE), if the number of occurrences for the string in either corpus is 0, we indicate that the ratio does not apply, but give the absolute difference between the two scores (e.g. Sentence No. 81 in OpenSubtitles in Example 8.2); if the number of occurrences for the string in both outputs is 0, then we take their relationship as 0 (e.g. Sentence No. 48 in Example 8.5).

Example 8.1

Sentence No. 50 (BLEU score = 0.1685):

SogouT

ENG: We would also like them to have strength building activities and to minimize the time they spend sitting.

PE: 我们也希望（8632）他们进行力量训练，尽量减少（1366）他们坐着（3655）的时间。

HT: 我们还想要（218）他们进行力量训练，并且让他们坐着不动（70）的时间减到最少（28）。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 39.60	> 48.79	> 52.21

OpenSubtitles

ENG: We would also like them to have strength building activities and to minimize the time they spend sitting.

PE: 我们也希望（1099）他们进行力量训练，尽量减少（28）他们坐着（1942）的时间。

HT: 我们还想要（644）他们进行力量训练，并且让他们坐着不动（21）的时间减到最少（5）。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 1.71	> 5.6	> 92.48

Example 8.2

Sentence No. 81 (BLEU score = 0.1893):

SogouT

ENG: Can we break up longer periods of sitting in the workplace and get up and move about more?

PE: 我们可不可以不要总是(48761)坐在工作场所(685)，多起来(293)动一动(83)？

HT: 我们可以不要长时间(15282)的坐在工作间里(16)，经常站起来(5)并且四处走走(21)么？

Notes:

	Pink	Yellow	Blue	Green
PE vs. HT	> 3.19	> 42.81	> 58.60	> 3.95

OpenSubtitles

ENG: Can we break up longer periods of sitting in the workplace and get up and move about more?

PE: 我们可不可以不要总是(21468)坐在工作场所(51)，多起来(14)动一动(105)？

HT: 我们可以不要长时间(2368)的坐在工作间里(2)，经常站起来(0)并且四处走走(55)么？

Notes:

	Pink	Yellow	Blue	Green
PE vs. HT	> 9.07	> 25.5	> ratio n/a PE=HT+14	> 1.91

Example 8.3

Sentence No. 35 (BLEU score = 0.2195):

SogouT

ENG: But exercise is a very specific subset of physical activity.

PE: 但锻炼是身体活动的一个非常特别(171987)的子集(53)。

HT: 但是锻炼是非常特定(11299)的某些(25039)身体活动。

Notes:

	Pink	Yellow
PE vs. HT	> 15.22	< 472.43

OpenSubtitles

ENG: But exercise is a very specific subset of physical activity.

PE: 但锻炼是身体活动的一个非常特别(14829)的子集(3)。

HT: 但是锻炼是非常特定(11299)的某些(3626)身体活动。

Notes:

	Pink	Yellow
PE vs. HT	> 1.31	< 1208.67

Example 8.4

Sentence No. 61 (BLEU score = 0.2435):

SogouT

ENG: Well, a third of the people in the world are not sufficiently active for good health.

PE: 那么(129870)世界上有三分之一的人没有为了健康(59763)而足够的活跃身体(0)。

HT: 也就是说(11429)世界上三分之一的人对于良好的健康状况(9)不是十分活跃(603)。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 11.36	< 603	> 6640.33

OpenSubtitles

ENG: Well, a third of the people in the world are not sufficiently active for good health.

PE: 那么(77977)世界上有三分之一的人没有为了健康(4334)而足够的活跃身体(0)。

HT: 也就是说(1990)世界上三分之一的人对于良好的健康状况(0)不是十分活跃(4)。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 39.18	< ratio n/a HT=PE+4	> ratio n/a PE=HT+4334

Example 8.5

Sentence No. 48 (BLEU score = 0.2447):

SogouT

ENG: We would like adults to have moderate activity of at least 30 minutes, 5 days a week.

PE: 我们希望(253604)大人(2663)有一个至少 30 分钟，每周 5 天(22)的适度活动(0)。

HT: 我们想要(32933)成年人(3031)一周有 5 天做(0)至少 30 分钟不剧烈的活动(0)。

Notes:

	Pink	Yellow	Blue	Green
PE vs. HT	> 7.70	< 1.14	> 22	= 0

OpenSubtitles

ENG: We would like adults to have moderate activity of at least 30 minutes, 5 days a week.

PE: 我们希望(60310)大人(9776)有一个至少 30 分钟，每周 5 天(1)的适度活动(0)。

HT: 我们想要(49562)成年人(903)一周有 5 天做(0)至少 30 分钟不剧烈的活动(0)。

Notes:

	Pink	Yellow	Blue	Green
PE vs. HT	> 1.22	> 10.83	> 1	= 0

Example 8.6

Sentence No. 66 (BLEU score = 0.2550):

SogouT

ENG: Patterns that are similar in young people, boys tend to be more active than girls, and children get less active as they age.

PE: 年轻人的模式类似(29693)，男孩往往比女孩更活跃，孩子越长大(48)越不活跃(0)。

HT: 在年轻人当中，这种模式是相似的(198)，男孩比女孩更活跃，孩子们会随着年龄增长(176)变得不那么活跃(0)。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 149.96	< 3.67	= 0

OpenSubtitles

ENG: Patterns that are similar in young people, boys tend to be more active than girls, and children get less active as they age.

PE: 年轻人的模式类似(2366)，男孩往往比女孩更活跃，孩子越长大(14)越不活跃(0)。

HT: 在年轻人当中，这种模式是相似的(30)，男孩比女孩更活跃，孩子们会随着年龄增长(5)变得不那么活跃(0)。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 78.87	> 2.8	= 0

Example 8.7

Sentence No. 31 (BLEU score = 0.2554):

SogouT

ENG: We often hear of physical activity and exercise being used interchangeably.

PE: 我们经常听到(29880)身体活动和锻炼(16128)互换(2184)使用(161091)。

HT: 我们经常听说(14823)身体活动和锻炼(16128)这两个词使用的时候(66)是可以互换的(8)。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 2.02	> 2440.77	> 273

OpenSubtitles

ENG: We often hear of physical activity and exercise being used interchangeably.

PE: 我们经常听到(29556)身体活动和锻炼互换(100)使用(9162)。

HT: 我们经常听说(16961)身体活动和锻炼这两个词使用的时候(3)是可以互换的(0)。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 1.74	> 3054	> 100

Example 8.8

Sentence No. 79 (BLEU score = 0.2593):

SogouT

ENG: And some of the ways that we can do this are to think about using alternative forms of transport.

PE: 我们可以做的一些方法(77051)是考虑使用替代(11132)交通工具的方式(0)。

HT: 其中的一些办法(57871)就是我们可以考虑使用多样的(614)交通方式(230)。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 1.33	> 18.13	< 230

OpenSubtitles

ENG: And some of the ways that we can do this are to think about using alternative forms of transport.

PE: 我们可以做的一些方法(10448)是考虑使用替代(551)交通工具的方式(0)。

HT: 其中的一些办法(57871)就是我们可以考虑使用多样的(18)交通方式(8)。

Notes:

	Pink	Yellow	Blue
PE vs. HT	< 5.54	> 30.61	< 8

Example 8.9

Sentence No. 41 (BLEU score = 0.2727):

SogouT

ENG: And these are all exercise-related activities, because they're structured and we're largely pursuing them for fitness and health benefit.

PE: 这些(281175)都是和锻炼(16128)相关的活动，因为它们是结构化(534)的，我们锻炼在很大程度上是为了追求健身(8331)和健康的好处。

HT: 以上这些(358)都是与锻炼(16128)有关的活动，因为他们是有结构(37)，并且主要追求健康和完美的体型(2)。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 785.41	> 14.43	> 4165.5

OpenSubtitles

ENG: And these are all exercise-related activities, because they're structured and we're largely pursuing them for fitness and health benefit.

PE: 这些(80484)都是和锻炼相关的活动，因为它们是结构化(3)的，我们锻炼在很大程度上是为了追求健身(572)和健康的好处。

HT: 以上这些(11)都是与锻炼有关的活动，因为他们是有结构(11)，并且主要追求健康和完美的体型(0)。

Notes:

	Pink	Yellow	Blue
PE vs. HT	> 7316.73	< 3.67	> 572

Example 8.10

Sentence No. 72 (BLEU score = 0.2727):

SogouT

ENG: So for example, if you have a desk-based job, then it's likely that you spend quite a lot of your day sitting or sedentary.

PE: 所以例如，如果你是伏案工作(50)，那么很可能(25629)你一天里有很多的时间(13)都坐着或久坐不动(9)。

HT: 例如，如果你拥有一份案头工作(0)，那么你极有可能(1278)一天大量时间(423)坐着或者不怎么活动(3)。

Notes:

	Pink	Yellow	Blue	Green
PE vs. HT	> 50	> 20.05	< 32.54	> 3

OpenSubtitles

ENG: So for example, if you have a desk-based job, then it's likely that you spend quite a lot of your day sitting or sedentary.

PE: 所以例如，如果你是伏案工作(1)，那么很可能(2247)你一天里有很多的时间(15)都坐着或久坐不动(0)。

HT: 例如，如果你拥有一份案头工作(0)，那么你极有可能(38)一天大量时间(33)坐着或者不怎么活动(0)。

Notes:

	Pink	Yellow	Blue	Green
PE vs. HT	> 1	> 59.13	< 2.2	= 0

There are 32 strings in the ten sentences in total. The green string in Sentence 48 and the blue string in Sentence 66 return zero occurrences in both corpora. When using SogouT corpus as a reference, there are 24 strings where the PE output scores higher in the number of occurrences, which accounts for 75%. This is not true only for six strings, which accounts for 25%. As for OpenSubtitles corpus, there are 23 strings where the PE

output scores higher in the number of occurrences, which accounts for 71.88%. This is not true only for six strings, which accounts for 18.75%. We can see that the result of the relationship between the corresponding outputs for the two conditions (HT vs PE) for Sentences 31, 35, and 61 are consistent with that of SogouT, while other sentences all have some slight differences. In a word, using both OpenSubtitles and SogouT as 'reference corpora', PE subtitles appear to use more frequent vocabulary (individual words and multi-word units) than HT subtitles. This means PE subtitles might look more familiar to the participants and thus were easier to process than HT subtitles (see Section 4.2.4). Because only ten sentences have been analysed, this observation has provided more insight into the study but cannot be conclusive.

Along with the result of TQA, which suggests there is no significant difference in the quality between PE and HT subtitles, the two supplementary methods have provided some insights into the cause of the unexpected results of eye-tracking and questionnaires, which is, the quality of HT subtitles is not higher than that of PE subtitles. As mentioned in Section 7.3.1, the premise of all the hypotheses in this research is that the quality of raw MT subtitles was lower than PE subtitles, which was lower than HT subtitles. If the premise is not totally correct, it is reasonable to have unexpected results that HT subtitles were sometimes less comprehensible and more cognitively demanding.

Chapter 9 Conclusions

9.1 Research aims

This research was motivated by the language barrier faced by the rapidly developing MOOCs sector in China and around the world. To remove this barrier, the more and more widely used MT seems to provide an alternative or complementary solution, as it can generate translated subtitles in a very short time. Nevertheless, very little attention has been paid to the use and utility of MT for MOOC content. MT and post-editing have been extensively studied by scholars from both the translation field and computing field as discussed in Chapter 2. Also, considerable research has been done in the field of MOOCs and that of subtitling. However, there are still gaps to be filled. While there is a lot of research on each of these topics, there is very little research linking them together, and this is what this PhD thesis does.

The main purpose of this study was to test the impact of machine translated subtitles on Chinese viewers' reception of MOOC content. The researcher was interested in whether the viewers' reception of raw machine translated subtitles is different compared to fully post-edited machine translated subtitles and human translated subtitles. Meanwhile, a comparison between post-edited translation and human translation was also conducted although it was assumed that there would be fewer quality differences between these two conditions.

In short, the aim of this research was to answer the overarching question:

Is there a difference in reception between participants who are offered raw machine translated subtitles and those who are offered full post-edited machine translated subtitles and human translated subtitles?

The concept "reception" is operationalized in this study using Gambier's (2009) model, which is based on the three R's: response, reaction and repercussion. Response refers to the initial physical response of a viewer to the subtitle and the rest of the MOOC image. It was measured using glance count, glance duration, and fixation count. Reaction involves the cognitive follow-on from initial response, and is concerned with how much effort is involved in processing the initial stimulus and what is understood by the viewer. It was measured partly through average fixation duration and partly through testing viewers' comprehension of the MOOC content. Repercussion refers to attitudinal and sociocultural dimensions of AVT consumption, and was captured using an attitude survey.

The researcher firstly used two core methods, questionnaires and eye-tracking, to study viewers' reception of translated subtitles. The results of the first part of the post-task questionnaire, comprehension testing, demonstrated that the participants who were shown the post-edited subtitles (Group PE) scored the best, and the participants who were shown the raw MT subtitles (Group RAW) scored the worst. The results of the second part of the questionnaire, the attitude survey, revealed that Group PE had the best attitude towards their subtitles, while Group HT had a surprising result that they were the least satisfied with the subtitles provided.

The results for cognitive effort measured through the eye-tracking method

demonstrated that the three groups returned no statistically significant differences in most eye-tracking metrics. More specifically, the results for Group PE and Group HT were opposite to what had been hypothesized. Because all the hypotheses related to the two groups were built upon the premise that the quality of full PE subtitles was lower than HT subtitles, to further study if this premise was correct, the quality of MT output and HT output needed to be compared. Hence, the researcher carried out two supplementary methods, TQA and frequency analysis, for PE subtitles and HT subtitles.

For TQA, the researcher used both human evaluation and automatic evaluation, the former was conducted by ten Chinese translation novices, and the latter was implemented by using the BLEU score. When doing human evaluation, the novices were working based on the metrics adapted from the well-known MQM model. The TQA shows that the two types of subtitles have no significant difference in quality, while the frequency analysis indicated that compared with both in-domain and out-of-domain corpora, PE subtitles appear to use more frequent vocabulary than HT subtitles, indicating that viewers probably were more familiar with PE subtitles and spent less time processing them, which could lead to a positive effect on their reception metrics. The results of both supplementary methods suggest that the quality of HT subtitles is not higher than that of PE subtitles, which means the premise for the hypotheses was not totally correct (the part of the premise that the quality of PE subtitles was higher than that of raw MT subtitles was correct though).

In summary, this research has shown a difference in reception between participants who are offered raw machine translated subtitles and those who are offered full post-edited machine translated subtitles and human translated subtitles. Participants who

were offered full PE subtitles scored better overall on the selected reception metrics than those who were offered raw MT subtitles. HT subtitles, on the other hand, did not necessarily lead to better reception as expected; in contrast, the participants who were offered HT subtitles performed the worst in some of the selected reception metrics. Notwithstanding this, most participants had a positive attitude towards the subtitles regardless of their type, which means MT could potentially be useful for translating English subtitles for MOOCs into Chinese, and may contribute to the long-term development of MOOCs in China.

9.2 Contributions

This study has been successful in answering the main research question by applying an adapted reception model and adopting a mixed-methods approach that combines questionnaires, eye-tracking, TQA and frequency analysis. It has bridged the gap between reception studies and MT, and has explored the possibility of applying MT to MOOCs.

9.2.1 Contributions to MOOCs

As discussed in Chapter 2, there is a growing demand for translation of subtitles for MOOCs among users not only in China, but also other countries, especially given that MOOCs have been developing quickly in recent years. Language is a big obstacle for Chinese people who wish to learn via MOOCs, due to their low level of English. However,

little attention has been given to Chinese language in the research on subtitling MOOCs. The present research has advanced the field by recruiting native Chinese speakers to empirically study English-to-Chinese subtitles for MOOCs. The MOOC platform, Coursera, and the MT engine used for this research, Google Translate, are both accessible and have been localized in China, which means this research is relevant to Chinese people's study and life, and could have real impact.

In addition, the literature shows that fansubbing and crowdsourcing are the most popular ways to translate subtitles for MOOCs in China now, thus this research has opened a new door in this field. Results show that the quality of machine translated subtitles and human translated subtitles for the MOOC used in this study did not differ much, and the reception by Chinese viewers of machine translated subtitles is theoretically not necessarily lower than for human translated subtitles. Therefore, this study not only adds to research on MOOCs, but also hopefully will bring inspiration to the industry. It has paved the way for the wider application of MT in subtitling MOOCs, making it possible to shorten the time for production of subtitled MOOCs and save on labour costs.

9.2.2 Contributions to reception studies

Section 2.4 shows that reception studies of translated text are still limited in quantity. Further to this, it can be seen that: 1) very few reception studies are about machine translated subtitles; 2) very few reception studies are about the subtitles of MOOCs; and 3) very few are about Chinese viewers or the English-Chinese language pair. This

research has addressed these three issues directly by utilizing Gambier's reception model to study Chinese viewers' reception of machine translated subtitles (from English to Chinese) for MOOCs. While Gambier's model has been successfully used by a few scholars for reception studies of translated text, this research has provided further evidence of the usefulness of this approach. This research has combined Gambier's model with two well-established methods in translation studies, eye-tracking and questionnaires. Results have consolidated the fact that a combination of the two methods is effective in reception studies. Apart from that, in the post-task questionnaire, the technology acceptance model (TAM) was applied for designing questions. To the best of my knowledge, it is the first time that TAM, a popular model in the field of information systems, has been used for reception studies of translated text.

There has been little research on MT in reception studies, which is undoubtedly unbalanced compared with the speed of its development and the increasing attention paid to it. People's attitudes towards MT and its reception are unclear and deserve more investigation. This research has contributed to filling these gaps by dividing viewers into three groups, observing and comparing their reception of human translated and machine translated (both raw and post-edited) subtitles. Results show that Chinese viewers' reception of machine translated subtitles was positive, and, for some measures, even better than for human translated subtitles.

9.2.3 Contributions to the mixed-methods approach

Chapter 3 discussed the significance of triangulation when doing research. This PhD research responds to this approach by adopting mixed methods, including eye-tracking, questionnaires, TQA and frequency analysis. There are previous studies that use each of these methods or a combination of two of them, but none has employed four of them together. Using such a mixed-methods approach not only helped to assure the validity of this research, but also made it innovative and new.

Both TQA and frequency analysis have complemented the methods of eye-tracking and questionnaires, and the results of the former two have explained the unexpected results of the latter two. In a word, this research has provided a new and effective mixed-methods approach to reception studies of translated text, and has presented an example of a combination of reception studies and TQA.

9.2.4 Contributions to MT

This research expands MT studies from the following six viewpoints:

1) it provides a reception study of MT;
2) it focuses on the MT from English to Chinese;
3) it presents a quality assessment of MT output, including human evaluation and automatic evaluation;
4) it compares MT output (raw and post-edited) with HT output;

5) for the first time, it adopts a mixed-methods approach of eye-tracking, questionnaires, TQA and frequency analysis, to investigate MT;
6) it explores the possibility of applying MT to AVT, especially subtitling MOOCs.

The existing studies on MT are mostly about improving, testing, or training the MT engine from the perspective of computer science, few studies have explored MT from the user's perspective, and even fewer have focused on users' reception of MT. This research stands on the user side, investigating Chinese viewers' reception of two types of machine translated subtitles, raw and full post-edited. The results demonstrate that the quality of MT is improving significantly and indicate there is a great prospect for the wider application of MT in the translation of various texts or spoken languages.

In this research, MT was applied to the translation of subtitles, which is a field that emphasizes timeliness and accuracy. The abovementioned positive results can reinforce the connection between MT, AVT, and the film and TV industry, thus potentially reducing the delay caused by translating subtitles in the future. Furthermore, the subtitles in this research were for MOOCs, an important way of educating with a wide range of users all over the world. The researcher has bridged the gap between MOOCs and MT, and shown that the language barrier of learning via MOOCs is likely to be removed by MT, so that more people can enjoy this method of education and benefit from it. In short, this research will hopefully promote the wider spread of education.

As MT keeps developing quickly, it has caused disputes in society. People are both curious and sceptical about it. Therefore, it is believed that this study can attract the

attention of people from, but not limited to, both translation studies and computing sciences, both academia and industry, and also contribute to the application of MT in the long run.

9.3 Limitations

Despite the contributions, some limitations should be acknowledged. First of all, although 61 participants were recruited for the main experiment, the eye-tracker failed to track 22 of them, thus only 39 participants completed the full experiment. The eye-tracking problem is particularly associated in the literature with Asian participants, and needs to be taken into account when designing eye-tracking research. When analysing eye-tracking data, data for only a dozen people in each group was valid (Group PE: 12; Group RAW: 14; Group HT: 13). It is known that finding participants for empirical research is usually difficult, hence it is a shame that a significant portion of data was invalid for this research. Ten valid eye-tracking data sets for each group only met the minimum standards the researcher expected, more participants are needed for more robust results. In retrospect, there are several shortcomings in experimental setup that need to be improved: 1) lighting was not controlled. The experiment was conducted in an independent office with a few big windows, the change of natural light may have had some influence on the brightness of the screen and the eye-tracking process, although the process was only seven minutes long; 2) the movement of the participants was not fully controlled. There was no chin rest for the participants, so although they were told not to, they could still move unconsciously during the seven minutes, which would certainly affect the eye-tracking accuracy; 3) the position of the subtitles was too low.

Because the software used for adding subtitles to the MOOC video did not have the feature of adjusting the position of subtitles, the default position was too close to the bottom of the screen, so that some fixations that drifted off the bottom of the subtitled area may not have been captured even though they were valid fixations.

Secondly, in regard to the human translated subtitles, the human translator was a bilingual (an English teacher in a Chinese school) rather than a professional translator. Although a large number of subtitles are translated by fansubbers and crowd workers, who are mostly non-professional volunteer translators, there are also a large number of subtitles being translated by professional translators. This research has only studied non-professional human translation; certain results would possibly be different if the subtitles were translated by a professional translator.

Thirdly, because the frequency analysis is a complementary method to TQA, which did not suggest any difference between HT and PE subtitles, only ten sentences were investigated. Results show that the frequency of the vocabulary in HT subtitles was lower than that of PE subtitles. While this frequency analysis is limited in scope, it is still indicative of a line of research worth pursuing. In addition, one fear with the increasing use of MT is that translated texts will become more homogenized, as more frequent formulations tend to get re-used in MT (Kenny, 2020). This research, however, points to a possible advantage of such homogenization – frequently occurring vocabulary or combinations of vocabulary units might actually aid comprehension.

Fourthly, the ten evaluators for TQA were either Chinese undergraduates or postgraduates in translation studies. They were more like translation novices rather than professional translators. If conditions allow, it may be better to recruit professional translators for TQA. However, it has to be noted that even if professionals are recruited, it does not necessarily lead to a high level of agreement, an example is Guerberof (2012), who used three professional revisers and their level of agreement was also low.

9.4 Future work

The scope of this research is necessarily limited and there is much scope for future research in the area. Firstly, much subtitling work is done by professional translators, hence it would be beneficial to expand this research to include professional human translated subtitles and compare them with MT and PE subtitles. Results could then be compared to the results of this research, so that we can see if there is any difference between the two settings, and find out how professional and non-professional subtitles affect the outcome. Secondly, future work may include employing professional translators as evaluators for TQA. Thirdly, in the frequency analysis of this research, only ten sentences were analysed, which only provided an indicative result. Therefore, this could also be taken a step further by analysing all the subtitles. The frequency of the vocabulary in the three types of subtitles could be calculated and compared.

A further step would be the implementation for different types of MOOCs, including some that are very technical. This research has investigated a MOOC that is relatively generic; however, many MOOCs are not, such as those in finance, computer sciences,

physics etc. It would be interesting to see how MT performs in translating these types of content, and how viewers react to these MT subtitles. Also, this research only focuses on Chinese, as MOOCs can be accessed all over the world, a different set of languages would be worth researching, from the most to the least challenging for MT.

References

Agudo, R.R. (2019) *The Language of MOOCs*. Available at: https://www.insidehighered.com/digital-learning/views/2019/01/09/moocs-overwhelming-dependence-english-limits-their-impact-opinion (Accessed: 15 December 2019).

Agrawal, R. and Gupta, N. (2018) *Extracting Knowledge From Opinion Mining*. Pennsylvania: IGI Global.

Alves, F., Pagano, A. and da Silva, I. (2009) 'A new window on translators' cognitive activity: methodological issues in the combined use of eye tracking, key logging and retrospective protocols', in Mees, I., Alves, F. and Göpferich, S. (eds.) *Methodology, Technology and Innovation in Translation Process Research*. Frederiksberg: Samfundslitteratur, pp. 267-291.

Alves, F., Gonçalves, J.L. and Szpak, K. (2012) 'Identifying instances of processing effort in translation through heat maps: An eye-tracking study using multiple input sources', *Proceedings of 24th International Conference on Computational Linguistics*. IIT Bombay, Mumbai, India, 8-15 December 2012, p. 5.

Ambati, V., Vogel, S. and Carbonell, J. G. (2010) *Active Learning and Crowd-Sourcing for Machine Translation*. Available at: https://www.cs.cmu.edu/~jgc/publication/PublicationPDF/Active_Learning_And_Crowd-Sourcing_For_Machine_Translation.pdf (Accessed: 11 October 2019)

Antonini, R. (2005) 'The perception of subtitled humor in Italy', *Humor*, *18*(2), pp. 209-225.

Antonini, R. (2007) 'SAT, BLT, Spirit Biscuits, and the Third Amendment: What Italians make of cultural references in dubbed texts', *Doubts and Directions in Translation Studies: Selected Contributions from the EST Congress*. University of Lisbon, Lisbon, Portugal, 26-29 September 2004. John Benjamins, pp. 153-167.

Aranberri, N., Labaka, G., Diaz de Ilarraza, A., & Sarasola, K. (2014) 'Comparison of post-editing productivity between professional translators and lay users', *Proceeding of AMTA Third Workshop on Post-editing Technology and Practice (WPTP-3).* Vancouver, Canada, 26 October, pp. 20-33.

Armstrong, S., Caffrey, C. and Flanagan, M. (2006) 'Translating DVD subtitles from English-German and English-Japanese using Example-Based Machine Translation', *EU-High-Level Scientific Conference Series MuTra 2006 – Audiovisual Translation Scenarios: Conference Proceedings*. Copenhagen, Denmark, 1-5 May 2006, pp. 1-12.

Arnáiz-Uzquiza, V. (2012) *Subtitling for the deaf and the hard-of-hearing some parameters and their evaluation*. PhD thesis. Universitat Autònoma de Barcelona.

Attnäs, M., Senellart, P. and Senellart, J. (2005) 'Integration of SYSTRAN MT systems in an open workflow', *Proceedings of MT Summit X.* Phuket, Thailand, 12-16 September, pp. 211-218.

Aziz, W., Castilho, S. and Specia, L. (2012) 'PET: a Tool for Post-editing and Assessing Machine Translation', *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey, 21-27 May 2012. European Language Resources Association, pp. 3982–3987.

Aziz, W., Koponen, M., and Specia, L. (2014) 'Sub-sentence level analysis of machine translation post-editing effort', in O'Brien, S., Balling, L. W., Carl, M., Simard, M., and Specia, L. (eds) *Post-editing of Machine Translation: Processes and Applications.* Newcastle: Cambridge Scholars Publishing.

Banerjee, P., Naskar, S.K., Roturier, J., Way, A. and van Genabith, J. (2012) 'Translation quality-based supplementary data selection by incremental update of translation models', *Proceedings of COLING 2012*. IIT Bombay, Mumbai, India, 8-15 December 2012, pp. 149-166.

Banerjee, S. and Lavie, A. (2005) 'METEOR: An automatic metric for MT evaluation with improved correlation with human judgments', *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. University of Michigan, Michigan, USA, 29 June. Association for Computational Linguistics, pp. 65–72.

Beaven, T., Comas-Quinn, A., Hauck, M., de los Arcos, B. and Lewis, T. (2013) 'The Open Translation MOOC: Creating Online Communities to Transcend Linguistic Barriers', *Journal of Interactive Media in Education*, *2013*(3), pp. 1-14. doi: http://doi.org/10.5334/2013-18.

Behnke, M., Miceli Barone, A.V., Sennrich, R., Sosoni, V., Naskos, T., Takoulidou, E., Stasimioti, M., Menno, V.Z., Castilho, S., Gaspari, F. and Georgakopoulou, P. (2018) 'Improving machine translation of educational content via crowdsourcing', *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC2018).* Miyazaki, Japan, 7-12 May 2018. European Language Resources Association, pp. 3343-3347.

Bentivogli, L., Bisazza, A., Cettolo, M. and Federico, M. (2016) 'Neural versus phrase-based machine translation quality: a case study', *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, USA, 1-5 November 2016. Association for Computational Linguistics, pp. 257-267.

Blain, F., Senellart, J., Schwenk, H., Plitt, M. and Roturier, J. (2011) 'Qualitative analysis of post-editing for high quality machine translation', *MT Summit XIII: the Thirteenth Machine Translation Summit.* Xiamen, China, 19-23 September 2011. Asia-Pacific Association for Machine Translation, pp.164-171.

Blignaut, P., and Wium, D. (2014) 'Eye-tracking data quality as affected by ethnicity and experimental design', *Behavior Research Methods*, *46*(1), pp. 67–80.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H. and Soricut, R. (2014) 'Findings of the 2014 workshop on statistical machine translation', *Proceedings of the ninth workshop on statistical machine translation*. Baltimore, Maryland, USA, 26 June 2014. Association for Computational Linguistics, pp. 12-58.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Logacheva, V., Monz, C. and Negri, M. (2016) 'Findings of the 2016 conference on machine translation', *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany, 11-12 Aug 2016. Association for Computational Linguistics, pp. 131-198.

Brace, I. (2013) *Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research*. London: Kogan Page Publishers.

Briggs, N. (2018) 'Neural machine translation tools in the language learning classroom: Students' use, perceptions, and analyses', *The JALT CALL Journal*, *14*(1), pp. 3-24.

Brown, P., Cocks, J., Pietra, S.D., Pietra, V.D., Jelinek, F., Mercer, R. and Roossin, P. (1988) 'A statistical approach to French/English translation', *Proceeding RIAO '88 User-Oriented Content-Based Text and Image Handling*. Grenoble, France, 21-24 March 1988, pp. 810-828.

Bryman, A. (2016) *Social Research Methods*. Oxford: Oxford University Press.

Bucaria, C. (2005) 'The Perception of the Non-Verbal: How Italians React to What They see but isn't Explained', *EU High Level Conference Series – Challenges of Multidimensional Translation: Conference Proceedings.* Saarbrücken, Germany, 2-6 May 2005.

Bucaria, C. and Chiaro, D. (2007) 'End-user perception of screen translation: The case of Italian dubbing', *Tradterm*, *13*, pp. 91-118.

Cadwell, P., Castilho, S., O'Brien, S. and Mitchell, L. (2016) 'Human factors in machine translation and post-editing among institutional translators', *Translation Spaces*, *5*(2), pp. 222-243.

Caffrey, C. (2008a) 'Using pupillometric, fixation-based and subjective measures to measure the processing effort experienced when viewing subtitled TV anime with pop-up gloss', in Gopferich, S. and Rasmussen, A. (eds.) *Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing*. Copenhagen: Samfundslitteratur, pp. 125-144.

Caffrey, C. (2008b) 'Viewer perception of visual nonverbal cues in subtitled TV Anime 1', *European Journal of English Studies*, *12*(2), pp. 163–178.

Caffrey, C. (2009) *Relevant abuse? Investigating the effects of an abusive subtitling procedure on the perception of TV anime using eye tracker and questionnaire*. PhD thesis. Dublin City University. Available at: http://doras.dcu.ie/14835/ (Accessed: 11 October 2019)

Cai, X. (2015) 'Fansubbing humor: A mainland China case study', *Journalism*, *5*(9), pp. 435-453.

Carl, M., Dragsted, B., Elming, J., Hardt, D. and Jakobsen, A.L. (2011) 'The process of post-editing: a pilot study', *Copenhagen Studies in Language*, *41*, pp. 131-142.

Castilho, S, Aziz, W. and Specia, L. (2011) 'Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles', *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Hissar, Bulgaria, 12-14 September 2011. Association for Computational Linguistics, pp. 97-103.

Castilho, S., O'Brien, S., Alves, F. and O'Brien, M. (2014) 'Does post-editing increase usability? A study with Brazilian Portuguese as target language', *Proceedings of EAMT*. Dubrovnik, Croatia, 17-19 June 2014. European Association for Machine Translation, pp. 183-190.

Castilho, S. (2016) *Measuring acceptability of machine translated enterprise content*. PhD thesis. Dublin City University. Available at: http://doras.dcu.ie/21342/ (Accessed: 11 October 2019)

Castilho, S. and O'Brien, S. (2016) 'Evaluating the impact of light post-editing on usability', *Proceedings of 10th International Conference on Language Resources and Evaluation* (LREC 2016). Portorož, Slovenia, 23-28 May 2016. European Language Resources Association, pp. 310-316.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Way, A. and Georgakopoulou, P. (2018) 'Evaluating MT for massive open online courses', *Machine Translation*, *32*(3), pp.255-278.

Catley, A. (1999) *Methods on the move: a review of veterinary uses of participatory approaches and methods focussing on experiences in dryland Africa*. Available at: http://agris.fao.org/agris-search/search.do?recordID=GB2013203544 (Accessed: 11 September 2019).

CBN Weekly (2019) *2019 Xin Yixian Chengshi Guanfang Mingdan Chulu: Ni De Chengshi Pai Di Ji? [Official list of 2019's new first-tier cities: where does your city*

rank?]. Available at: https://www.yicai.com/news/100200192.html (Accessed: 2 December 2019)

Chan, Y.S. and Ng, H.T. (2008) 'MAXSIM: A maximum similarity metric for machine translation evaluation', *Proceedings of ACL-08: HLT*. Ohio, Columbus, USA, 15-20 June 2008. Association for Computational Linguistics, pp. 55-62.

Chatterjee, R., Weller, M., Negri, M. and Turchi, M. (2015) 'Exploring the planet of the apes: a comparative study of state-of-the-art methods for mt automatic post-editing', *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, 27-31 July 2015. Association for Computational Linguistics, pp. 156-161.

Chesterman, A. (2007) 'Bridge concepts in translation sociology', in Wolf, M. and Fukari, A. (eds.) *Constructing a Sociology of Translation*. Amsterdam: Benjamins Translation Library, p. 171.

Chiaro, D. (2004) 'Investigating the perception of translated Verbally Expressed Humour on Italian TV', *ESP Across Cultures*, *1*(1), pp. 35–52.

Chiaro, D. (2007) 'The effect of translation on humour response', *Doubts and directions in translation studies*, *72*, pp. 137–152.

China Education Center. (2019) *Project 211 and 985*. Available at: https://www.chinaeducenter.com/en/cedu/ceduproject211.php (Accessed: 2 December 2019)

Coldewey, D. (2018) *Judge says "literal but nonsensical" Google translation isn't consent for police search*.
Available at: http://social.techcrunch.com/2018/06/15/judge-says-literal-but-nonsensical-google-translation-isnt-consent-for-police-search/ (Accessed: 10 September 2019).

Cormier, D. (2008) *The CCK08 MOOC – Connectivism course, 1/4 way – Dave's Educational Blog*. Available at: http://davecormier.com/edblog/2008/10/02/the-cck08-mooc-connectivism-course-14-way/ (Accessed: 5 September 2019).

Costa, Â., Ling, W., Luís, T., Correia, R. and Coheur, L. (2015) 'A linguistically motivated taxonomy for Machine Translation error analysis', *Machine Translation*, *29*(2), pp.127-161.

Costa-Jussà, M.R., Formiga, L., Petit, J. and Fonollosa, J.A. (2014) 'Detailed Description of the Development of a MOOC in the Topic of Statistical Machine Translation', in Gelbukh A., Espinoza F.C., and Galicia-Haro S.N. (eds.) *Human-Inspired Computing and Its Applications*. Tuxtla Gutiérrez, Mexico, 16-22 November 2014. Springer, pp. 92-98. doi: https://doi.org/10.1007/978-3-319-13647-9_10.

Costa-Jussà, M.R., Formiga, L., Torrillas, O., Petit, J. and Fonollosa, J.A.R. (2015) 'A MOOC on approaches to machine translation', *The International Review of Research in Open and Distributed Learning*, *16*(6), pp. 174-205.

Creswell, J.W. and Clark, V.L.P. (2017) *Designing and conducting mixed methods research*. 3rd edn. Newbury Park: Sage Publications.

Cui, Q. (2014) 'Lun jiqi fanyi de yi hou bianji [Post-editing of machine translation]', *Zhongguo Fanyi [Chinese Translation Journal]*, *6*, pp. 68–73.

d'Ydewalle, G., Muylle, P. and van Rensbergen, J. (1985) 'Attention shifts in partially redundant information situations', in Oregan, J.K. and Levy-Schoen, A. (eds.) *Eye movements and human information processing: Selected/Edited Proceedings of the Third European Conference on Eye Movements*. Dourdan, France, September 1985. Elsevier, pp.375-384.

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', *Eye Movements from Physiology to Cognition*, pp. 313-321. doi: https://doi.org/10.1016/B978-0-444-70113-8.50047-3.

Dale, E. and Chall, J. (1948) 'A Formula for Predicting Readability', *Educational Research Bulletin*, *27*, pp.11-20.

Davis, F.D. (1985) *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. PhD thesis. Massachusetts Institute of Technology).

Davis, F.D., Bagozzi, R.P. and Warshaw, P.R. (1989) 'User acceptance of computer technology: a comparison of two theoretical models', *Management Science*, *35*(8), pp. 982–1003.

Day, M.Y. and Lee, C.C. (2016) 'Deep learning for financial sentiment analysis on finance news providers', *Proceedings of 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. San Francisco, CA, USA, 18-21 August 2016. IEEE, pp. 1127-1134.

de Almeida, G. and O'Brien, S. (2010) 'Analysing post-editing performance: correlations with years of translation experience', *Proceedings of the 14th annual conference of the European association for machine translation.* St. Raphaël, France, 27-28 May 2010. EAMT, pp. 27–28.

de Andrade Stupiello, É.N. (2007) 'Ethical implications of translation technologies. *Translation Journal, 12*(1), pp. 1-6.

Denkowski, M. and Lavie, A. (2010) 'Extending the METEOR machine translation evaluation metric to the phrase level', *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, CA, USA, 1-6 June 2010. Association for Computational Linguistics, pp. 250-253.

Densmer, L. (2014) *Light and Full MT Post-Editing Explained*. Available at http://info.moravia.com/blog/bid/353532/Light-and-Full-MT-Post-Editing-Explained (Accessed: 8 September 2019).

DePalma, D. (2013) *Post-editing in practice*. Available at: http://www.tcworld.info/e-magazine/translation-and-localization/article/post-editing-in-practice/ (Accessed: 11 September 2019).

Depraetere, I. (2010) 'What counts as useful advice in a university post-editing training context? Report on a case study', *EAMT 2010: proceedings of the 14th annual conference of the European association for machine translation.* St. Raphaël, France, 27-28 May 2010. EAMT, pp. 1-7.

Devore, J.L. (2011) *Probability and Statistics for Engineering and the Sciences*. Boston: Cengage learning.

Díaz-Cintas, J. and Muñoz Sánchez, P. (2006) 'Fansubs: Audiovisual translation in an amateur environment', *The Journal of Specialised Translation*, *6*(1), pp. 37–52.

Dillinger, M. (2016) *MT escaped from the Lab!* Available at: https://amtaweb.org/wp-content/uploads/2016/10/Dillinger_AMTA2016Keynote_dist.pdf (Accessed: 15 December 2019)

do Carmo, F.E.M. (2017) *Post-editing: a theoretical and practical challenge for translation studies and machine learning.* PhD thesis. Universidade do Porto.

Doddington, G. (2012) 'The role of score calibration in speaker recognition', *Proceedings of Thirteenth Annual Conference of the International Speech Communication Association*. Portland, Oregon, USA, 9-13 September 2012. International Speech Communication, pp 1103-1106.

Doherty, S. (2012) *Investigating the effects of controlled language on the reading and comprehension of machine translated texts: A mixed-methods approach.* PhD thesis. Dublin City University. Available at: http://doras.dcu.ie/16805/ (Accessed: 11 October 2019)

Doherty, S. and O'Brien, S. (2014) 'Assessing the usability of raw machine translation output: A user-centered study using eye tracking', *International Journal of Human Computer Interaction*, *30*(1), pp. 40–51.

Doherty, S. and Kruger, J.L. (2018) 'Assessing Quality in Human-and Machine-Generated Subtitles and Captions', in Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (eds.) *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer, pp. 179-197.

Doherty, S., Kruger, J.L., Dwyer, T., Perkins, C., Redmond, S. and Sita, J. (2018) 'The development of eye tracking in empirical research on subtitling and captioning', *Seeing into Screens: Eye Tracking and the Moving Image,* London: Bloomsbury Publishing, pp. 46-64.

Drugan, J. and Babych, B. (2010) 'Shared resources, shared values? Ethical implications of sharing translation resources', *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User*. Denver, CO., USA, 4 November, 2010. CNGL, pp. 3-9.

Drugan, J., Strandvik, I. and Vuorinen, E. (2018) 'Translation Quality, Quality Management and Agency: Principles and Practice in the European Union', in Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (eds.) *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer, p. 39.

Duchowsky, A.T. (2017) *Eye Tracking Methodology: Theory and Practice (Third Edition).* London: Springer.

Dugast, L., Senellart, J. and Koehn, P. (2007) 'Statistical Post-Editing on SYSTRAN's Rule-Based Translation System', *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, 23 June 2007. ACL, pp. 220-223.

Ellis, N.C. (2002) 'Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition', *Studies in second language acquisition, 24*(2), pp. 143-188.

Embassy of the People's Republic of China in Ireland (2019) *Education in China: a survey.* Available at: http://ie.china-embassy.org/eng/Education/EducationDevelopmentinChina/t112967.htm (Accessed: 2 December 2019)

Etchegoyhen, T., Bywood, L., Fishel, M., Georgakopoulou, P., Jiang, J., Van Loenhout, G., Del Pozo, A., Maucec, M.S., Turner, A. and Volk, M. (2014) 'Machine Translation for Subtitling: A Large-Scale Evaluation', *Proceedings of LREC*. Reykjavik, Iceland, 26-31 May 2014. European Language Resources Association, pp. 46-53.

Eyckmans, J., Anckaert, P. and Segers, W. (2009) 'The perks of norm-referenced translation evaluation', *Testing and assessment in translation and interpreting*, pp.73-93.

Farrús, M., Costa-Jussà, M.R., Mariño, J.B. and Fonollosa, J.A.R. (2010) 'Linguistic-based evaluation criteria to identify statistical machine translation errors',

Proceedings of 14th Annual Conference of the European Association for Machine Translation. St. Raphaël, France, 27-28 May 2010. EAMT, pp. 167-173.

Federico, M., Negri, M., Bentivogli, L. and Turchi, M. (2014) 'Assessing the impact of translation errors on machine translation quality with mixed-effects models', *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Dohar, Qatar, 25-29 October 2014. Association for Computational Linguistics, pp. 1643-1653.

Fernández Costales, A. (2013) 'Crowdsourcing and collaborative translation: mass phenomena or silent threat to Translation Studies?' *Hermēneus*, *15*, pp. 85–110.

Fernández-Torné, A. and Matamala, A. (2016) 'Machine translation in audio description? Comparing creation, translation and post-editing efforts', *SKASE Journal of Translation and Interpretation*, *9*(1), pp. 64–87.

Filizzola, T. (2017) *Italians' Perception and Reception of British Stand-Up Comedy Humour with Interlingual Subtitles: A Qualitative and Quantitative Study on Eddie Izzard's Shows.* PhD thesis. University College London.

Fishel, M., Georgakopoulou, Y., Penkale, S., Petukhova, V., Rojc, M., Volk, M. and Way, A. (2012) 'From subtitles to parallel corpora', *Proceedings of the 16th EAMT Conference*. Trento, Italy, 28-30 May. EAMT, pp. 3-6.

Flanagan, M. (1994) 'Error classification for MT evaluation', *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*. Columbia, Maryland, USA, 5-8 October 1994. AMTA, pp. 65-72.

Flanagan, M. (2009) 'Using example-based machine translation to translate DVD Subtitles', *Proceedings of the 3rd international workshop on example-based machine translation.* Dublin, Ireland, 12-13 November 2009. CNGL, pp. 85–92.

Flanagan, M. and Christensen, T.P. (2014) 'Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes', *The Interpreter and Translator Trainer*, *8*(2), pp. 257–275.

Flesch, R. (1948) 'A new readability yardstick', *Journal of applied psychology*, *32*(3), p. 221.

French, H.W. (2006) 'Chinese tech buffs slake thirst for US TV shows', *The New York Times*, 9 August, p. 2006.

Frey, L.R., Botan, C.H., and Kreps, G.L. (1991) *Investigating Communication: An Introduction to Research Methods*. Upper Saddle River: Prentice Hall.

Gambier, Y. (2003) 'Introduction: Screen transadaptation: Perception and reception', *The Translator*, *9*(2), pp. 171-189.

Gambier, Y. (2007) 'Challenges in research on audiovisual translation', *Translation research projects*, *2*, pp. 17-25. Available at: http://www.intercultural.urv.cat/media/upload/domain_317/arxius/TP2/gambier.pdf (Accessed: 15 December 2019)

Gambier, Y. (2009) 'Perception and reception of audiovisual translation: Implications and challenges', in Omar, H.C., Haroon, H., and Ghani, A.A. (eds.) *The Sustainability of the Translation Field*. Kuala Lumpur: Malaysian Translators Association, pp. 40–57.

Gambier, Y. (2018) 'Translation studies, audiovisual translation and reception', *Reception Studies and Audiovisual Translation, 141*, p. 43.

Gao, D. (2014a) '"Muke": hexin linian, shijian fansi yu wenhua anquan ['MOOC': core concept, practical reflection and cultural security]', *Dongbei Shida Xuebao: Zhexue Shehui Kexue Ban [Journal of Northeast Normal University - Philosophy and Social Sciences], 5*, pp. 178–186.

Gao, D. (2014b) 'MOOC Re de leng sikao – guoji shang dui MOOCs kecheng jiaoxue liu da wenti de shensi [Cold thinking of MOOC heat - International review of six major problems in MOOCs]', *Yuancheng Jiaoyu Zazhi [Journal of Distance Education]*, *32*(2), pp. 39–47.

Garant, M. (2009) 'A case for holistic translation assessment', *AFinLA-e: Soveltavan kielitieteen tutkimuksia*, (1), pp. 5-17.

Garg, S., Saini, A. and Khanna, N. (2016) 'Is sentiment analysis an art or a science? Impact of lexical richness in training corpus on machine learning', *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Jaipur, India, 21-24 September 2016. IEEE, pp. 2729-2735.

Georgakopoulou, P. (2012) 'Challenges for the audiovisual industry in the digital age: the ever-changing needs of subtitle production', *The Journal of Specialised Translation*, *17*, pp. 78–103.

Gerber-Morón, O., Szarkowska, A. and Woll, B. (2018). The impact of text segmentation on subtitle reading. *Journal of Eye Movement Research*, *6*(5), pp.1-12.

Gerber-Morón, O. (2019) *Subtitle segmentation quality across screens*. PhD thesis. Universitat Autònoma de Barcelona.

Giménez, J. and Màrquez, L. (2008) 'A smorgasbord of features for automatic MT evaluation', *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio, USA, 19 June. ACL, pp. 195-198.

Goldberg, J.H. and Kotval, X.P. (1999) 'Computer interface evaluation using eye movements: methods and constructs', *International Journal of Industrial Ergonomics*, *24*(6), pp. 631-645.

Goldberg, J.H. and Wichansky, A.M. (2003) 'Eye tracking in usability evaluation: A practitioner's guide', *The Mind's Eye*, pp. 493-516.

Goldman, M. (2008) *Statistics for Bioinformatics*.
Available at: https://www.stat.berkeley.edu/~mgoldman/Section0402.pdf (Accessed: 14 December 2019)

González, L.P. (2007) 'Fansubbing anime: insights into the "butterfly effect" of globalisation on audiovisual translation', *Perspectives*, *14*(4), pp. 260–277.

Graham, Y., Baldwin, T., Moffat, A. and Zobel, J. (2017) 'Can machine translation systems be evaluated by the crowd alone', *Natural Language Engineering*, *23*(1), pp. 3-30.

Gratton, C. and Jones, I. (2010) *Research Methods for Sports Studies*. London: Taylor & Francis.

Green, S., Heer, J. and Manning, C.D. (2013) 'The efficacy of human post-editing for language translation', *Proceedings of the SIGCHI conference on human factors in computing systems*. Paris, France, 27 April – 2 May 2013. ACM SIGCHI, pp. 439-448.

Gries, S.T. (2010) 'Useful statistics for corpus linguistics', *A mosaic of corpus linguistics: Selected approaches*, *66*, pp. 269-291.

Grůber, M. and Matoušek, J. (2010) 'Listening-test-based annotation of communicative functions for expressive speech synthesis', *International Conference on Text, Speech and Dialogue*. Brno, Czech Republic, 6-10 September 2010. Springer, pp. 283-290.

Guerberof, A. (2009) 'Productivity and quality in MT post-editing', *Proceedings of MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT*. Ottawa, Ontario, Canada, 26-30 Aug 2009. AMTA, pp. 1-9.

Guerberof, A. (2012) *Productivity and quality in the post-editing of outputs from translation memories and machine translation.* PhD thesis. Universitat Rovira i Virgili.

Guo, C., Shim, J. and Otondo, R. (2010) 'Social network services in China: An integrated model of centrality, trust, and technology acceptance', *Journal of Global Information Technology Management*, *13*(2), pp. 76–99.

Guokr MOOC Academy (2014) *2014 Nian Muke Xuexizhe Diaocha Baogao [2014 MOOC Learners Survey]*. Available at: https://mooc.guokr.com/post/610674/ (Accessed: 5 September 2019).

Guokr MOOC Academy (2015) *2015 Zaixian Xuexi Da Diaocha: Zaixian Zhiye Xuexi Jueqi, Fufei "10 Bei Su" Zengzhang [Online Learning Survey 2015: Online career learning is on the rise, with a 10-fold increase in fee paying].* Available at: https://www.jiemodui.com/N/41550.html%7D (Accessed: 8 September 2019).

Guokr MOOC Academy (2017) *2016 Zhishi Qingnian Baogao: Zaixian Jiaoyu Baijia Zhengming, Zhishi Fufei Shidai Lailin [2016 Educated Youth Report: Online education is booming, the era of paying for knowledge is coming].* Available at: https://36kr.com/p/5061911 (Accessed: 5 September 2019).

Haggard, S., Brown, S., Mills, R., Tait, A., Warburton, S., Lawton, W. and Angulo, T. (2013) *The maturing of the MOOC: Literature review of massive open online courses and other forms of online distance learning.* Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/240193/13-1173-maturing-of-the-mooc.pdf (Accessed: 11 October 2019)

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M. and Liu, S. (2018) 'Achieving human parity on automatic chinese to english news translation', *arXiv preprint arXiv:1803.05567*. Available at: https://arxiv.org/abs/1803.05567 (Accessed: 11 October 2019)

HCR (2016) *Shendu: 2006 Nian Zhongguo Muke Hangye Yanjiu Baipishu [The White Paper on China's MOOC Industry 2016].*
Available at: http://www.sohu.com/a/115837986_400678 (Accessed: 11 September 2019).

He, T. (2014) 'Fansubs and Market Access of Foreign Audiovisual Products in China: The Copyright Predicament and the Use of No Action Policy', *Oregon Review of International Law*, *16*(2), pp. 307-346.

Hew, K.F. and Cheung, W.S. (2014) 'Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges', *Educational Research Review*, *12*, pp. 45–58.

Heyn, M. (1996). 'Integrating machine translation into translation memory systems', *Proceedings of the EAMT Machine Translation Workshop (TKE'96)*. Vienna, Austria, 29-30 August 1996. EAMT, pp. 113-126.

Hofstadter, D. (2018) 'The Shallowness of Google Translate', *The Atlantic*. Available at: https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/ (Accessed: 10 September 2019).

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: Oxford University Press.

House, J. (1977) *A Model for Translation Quality Assessment*. Tübingen: Gunter Narr Verlag.

House, J. (1997) *Translation Quality Assessment: A Model Revisited*. Tübingen: Gunter Narr Verlag.

House, J. (2001) 'Translation quality assessment: Linguistic description versus social evaluation', *Meta: journal des traducteurs/Meta: translators' Journal*, *46*(2), pp. 243-257.

House, J. (2014) *Translation Quality Assessment: Past and Present*. Abingdon : Routledge.

Hsiao, C. (2014) 'The Moralities of Intellectual Property: Subtitle Groups as Cultural Brokers in China', *The Asia Pacific Journal of Anthropology*, *15*(3), pp. 218–241.

Hu, H. (2013) 'MOOC migration', *Diverse Issues in Higher Education*, *30*(4), p. 10.

Hu, K. and Cadwell, P. (2016) 'A comparative study of post-editing guidelines', *Baltic Journal of Modern Computing*, *4*(2), pp. 346-353.

Hu, K. and O'Brien, S. (2016) *Applying TAM (Technology Acceptance Model) to testing MT acceptance.* Available at: https://ec.europa.eu/info/sites/info/files/tef2016_kehu_en.pdf (Accessed: 8 September 2019).

Hu, K. (2018) 'MT use in China', *MultiLingual*, *29*(6), pp. 32-38.

Hu, Q. (2009) 'Zhongguo zimuzu yu xin ziyou zhuyi gongzuo lunli [Working ethics of Chinese fansubbers and neo-liberalism]', *Xinwenxue Yanjiu [Journalism Studies], 101*, pp. 177-214.

Iarossi, G. (2006) *The power of survey design: A user's guide for managing surveys, interpreting results, and influencing respondents*. Washington: The World Bank.

iiMedia (2018) *2018 Zhongguo Zaixian Jiaoyu Hangye Baipishu [2018 China Online Education Industry White Paper].*
Available at: https://www.iimedia.cn/c400/63080.html (Accessed: 5 September 2019).

Iriarte, M.M. (2017) *The Reception of subtitling for the deaf and hard of hearing: viewers' hearing and communication profile & subtitling speed of exposure*. PhD thesis. Universitat Autònoma de Barcelona.

Ivarsson, J. and Carroll, M. (1998) 'Code of good subtitling practice', *Subtitling*. Simrishamn: TransEdit, pp. 1-2. Available at: https://www.esist.org/wp-

content/uploads/2016/06/Code-of-Good-Subtitling-Practice.PDF.pdf (Accessed: 15 December 2019)

Ive, J., Max, A. and Yvon, F. (2018) 'Reassessing the proper place of man and machine in translation: a pre-translation scenario', *Machine Translation*, *32*(4), pp. 279-308.

Jacobs, A. (2013) 'Two cheers for Web U', *New York Times*, 20 April, pp. 1–7.

Jacoby, J. (2014) 'The disruptive potential of the Massive Open Online Course: A literature review', *Journal of Open, Flexible, and Distance Learning*, *18*(1), pp. 73-85.

Jensen, G. M. and Mostrom, E. (2012) *Handbook of Teaching for Physical Therapists - E-Book*. Amsterdam: Elsevier Health Sciences.

Jia, M., Gong, D., Luo, J., Zhao, J., Zheng, J. and Li, K. (2019a) 'Who can benefit more from massive open online courses? A prospective cohort study', *Nurse education today*, *76*, pp. 96-102.

Jia, Y., Carl, M. and Wang, X. (2019b) 'How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study', *Journal of Specialised Translation*, *31*, pp. 60-86.

Jiménez-Crespo, M.A. (2009) 'The evaluation of pragmatic and functionalist aspects in localization: towards a holistic approach to Quality Assurance', *The Journal of Internationalization and Localization*, *1*(1), pp. 60-93.

Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G. and Hughes, M. (2017) 'Google's multilingual neural machine translation system: Enabling zero-shot translation', *Transactions of the Association for Computational Linguistics*, *5*, pp. 339-351.

Joshi, A., Kale, S., Chandel, S. and Pal, D.K. (2015) 'Likert scale: Explored and explained', *British Journal of Applied Science & Technology*, *7*(4), p.396.

Junczys-Dowmunt, M., Dwojak, T. and Hoang, H. (2016) 'Is neural machine translation ready for deployment? A case study on 30 translation directions', *arXiv preprint arXiv:1610.01108.* Available at: https://arxiv.org/abs/1610.01108 (Accessed: 10 September 2019)

Just, M.A. and Carpenter, P.A. (1980) 'A theory of reading: From eye fixations to comprehension', *Psychological review*, *87*(4), p. 329.

Kenny, D. (2020) 'Machine Translation', in Baker, M. and Saldanha, G. (eds) *The Routledge Encyclopedia of Translation Studies,* 3rd Edition. London and New York: Routledge, pp. 305-310.

Khalilov, M. (2018) *The good, the bad and the ugly of machine translation for customer service*. Available at: https://unbabel.com/blog/machine-translation-customer-service/ (Accessed: 8 October 2019).

Kilgarriff, A. (1996) 'Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved LOB-Brown comparison', *Proceedings of ALLC-ACH Conference*. Bergen, Norway, 25-29 June 1996. Association for Computers and the Humanities, pp. 169-172.

Kim, T.K. (2017) 'Understanding one-way ANOVA using conceptual figures', *Korean Journal of Anesthesiology*, *70*(1), p. 22.

Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L. and Chissom, B.S. (1975) 'Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel', *Institute for Simulation and Training*, *56* Available at: https://stars.library.ucf.edu/istlibrary/56 (Accessed: 5 September 2019).

Kirchhoff, K., Capurro, D. and Turner, A. (2012) 'Evaluating user preferences in machine translation using conjoint analysis', *Proceedings of European Association for Machine Translation.* Trento, Italy, 28-30 May. EAMT, pp. 119-126.

Klubička, F., Toral, A. and Sánchez-Cartagena, V.M. (2017) 'Fine-grained human evaluation of neural versus phrase-based machine translation', *The Prague Bulletin of Mathematical Linguistics*, *108*(1), pp. 121-132.

Koehn, P. (2010) *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn, P. (2016) *The State of Neural Machine Translation (NMT).* Available at: https://omniscien.com/state-neural-machine-translation-nmt/.

Koehn, P. and Knowles, R. (2017) 'Six challenges for neural machine translation', *arXiv preprint arXiv:1706.03872.* Available at: https://arxiv.org/abs/1706.03872 (Accessed: 10 September 2019)

Kolowich, S. (2013) 'The professors who make the MOOCs', *The Chronicle of Higher Education*, 18 March. Available at*:* https://www.chronicle.com/article/The-Professors-Behind-the-MOOC/137905 (Accessed: 11 October 2019)

Koponen, M. (2012) 'Comparing human perceptions of post-editing effort with post-editing operations', *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montreal, Canada, 7-8 June 2012. Association for Computational Linguistics, pp. 181–190.

Koponen, M., Aziz, W., Ramos, L. and Specia, L. (2012) 'Post-editing time as a measure of cognitive effort', *Proceedings of Workshop of Post-editing Technology and Practice (WPTP-2012)*. San Diego, CA, USA, 28 October – 1 November 2012. AMTA, pp. 11-20.

Koponen, M. (2016) 'Is machine translation post-editing worth the effort? A survey of research into post-editing and effort', *The Journal of Specialised Translation*, *25*, pp. 131-148.

Kordoni, V., van den Bosch, A.P.J., Kermanidis, K.L., Sosoni, V., Cholakov, K., Hendrickx, I.H.E., & Huck, M. (2016). 'Enhancing access to online education: Quality machine translation of MOOC content', *Proceedings of the International Conference on Language Resources and Evaluation (LREC) 2016*. Portorož, Slovenia, 23-28 May 2016. European Language Resources Association, pp. 16-22.

Kovačič, I. (1995) 'Reception of subtitles: The non-existent ideal viewer', Translatio (Nouvelles de la FIT/FIT Newsletter), *14* (3-4), pp. 376–383.

Krosnick, J.A. and Fabrigar, L.R. (1997) 'Designing rating scales for effective measurement in surveys', in Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (eds.) *Survey Measurement and Process Quality*. New York: John Wiley, pp. 141-164.

Kundu, C., Das, R.K. and Sengupta, K. (2015) 'A High Performance Semi-Supervised Learning technique for Non-Standard Word Classification in Bengali News Corpus', *Journal of Intelligent Computing*, *6*(3), p. 75.

Künzli, A. and Ehrensberger-Dow, M. (2011) 'Innovative subtitling: A reception study', in Alvstad, C., Hild, A., and Tiselius, E. (eds.) *Methods and Strategies of Process Research*. Amsterdam: Benjamins Translation Library, pp. 187–200.

Kuo, S.Y. (2014) *Quality in subtitling: theory and professional reality*. PhD thesis. Imperial College London.

Łabendowicz, O. (2018) *The Impact of Audiovisual Translation Modality on the Reception and Perception of Culture-Specific References*. PhD thesis. University of Lodz Repository.

Lacruz, I., Shreve, G.M. and Angelone, E. (2012) 'Average pause ratio as an indicator of cognitive effort in post-editing: A case study', *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation of the Americas*. San Diego, CA, USA, 28 October – 1 November 2012. AMTA, pp. 29–38.

Lacruz, I., Denkowski, M. and Lavie, A. (2014) 'Cognitive Demand and Cognitive Effort in Post-Editing', *Proceedings of AMTA 2014*. Vancouver, Canada, 22-26 October 2014. AMTA, pp. 73-84.

Lancaster, G. (2007) *Research Methods in Management*. Abingdon: Routledge.

Landis, J.R. and Koch, G.G. (1977) 'An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers', *Biometrics*, *33*(2), pp. 363-374.

Larose, R. (1998) 'Méthodologie de l'évaluation des traductions', *Meta: journal des traducteurs/Meta: Translators' Journal*, *43*(2), pp. 163-186.

Läubli, S., Sennrich, R. and Volk, M. (2018) 'Has machine translation achieved human parity? A case for document-level evaluation', *arXiv preprint arXiv:1808.07048*. Available at: https://arxiv.org/abs/1808.07048 (Accessed: 11 October 2019)

Lee, M.M., Lin, M.F.G. and Bonk, C.J. (2007) 'OOPS, Turning MIT Opencourseware into Chinese: An analysis of a community of practice of global translators', *The International Review of Research in Open and Distributed Learning*, *8*(3), pp. 1-21.

Lee, T. E. (2014) 'An Exploratory Study of Fan-based Subtitling Culture in China: With a Reference to YYeTs', *Fanyi Xue Yanjiu Qikan [Studies of Interpretation and Translation],* (18), pp. 101–120.

Lee, Y., Kozar, K. A. and Larsen, K. R. (2003) 'The technology acceptance model: Past, present, and future', *Communications of the Association for information systems*, *12*(1), p. 50.

Legris, P., Ingham, J. and Collerette, P. (2003) 'Why do people use information technology? A critical review of the technology acceptance model', *Information & Management*, *40*(3), pp. 191–204.

Leonard, S. (2005) 'Progress against the law Anime and fandom, with the key to the globalization of culture', *International journal of cultural studies*, *8*(3), pp. 281–305.

Lepre, O. (2015) 'Cultural References in Fansubs: When Translating is a Job for Amateurs', in Perego, E. and Bruti, S. (eds.) *Subtitling Today: Shapes and Their Meanings*. Newcastle: Cambridge Scholars Publishing, pp. 77-98.

Levenshtein, V.I. (1966) 'Binary codes capable of correcting deletions, insertions, and reversals', *Soviet physics doklady*, *10*(8), pp. 707-710.

Li, M. and Zhu, X. (2013) 'Yinghan jiyi cuowu fenlei ji shuju tongji fenxi [Error patterns and statistical analyses of an English-Chinese machine translation corpus]', *Shanghai Ligong Daxue Xuebao [Journal of University of Shanghai for Science and Technology],* *35*(3), pp. 201–207.

Lin, C.Y. and Och, F.J. (2004) 'Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics', *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Barcelona, Spain, 21-26 July. ACL, pp. 605-613.

Lison, P. and Tiedemann, J. (2016) 'Opensubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles', *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, 23-28 May. European Language Resources Association, pp. 923-929. Available at: https://www.duo.uio.no/handle/10852/50459 (Accessed: 13 September 2019).

Liu, S., Liang, L., and Michelson, E. (2014). 'Migration and Social Structure: The Spatial Mobility of Chinese Lawyers', *Law & Policy*. *36* (2), pp. 165–194.

Lommel, A., Burchardt, A. and Uszkoreit, H. (2014) 'Multidimensional Quality Metrics (MQM)', *Tradumàtica*, *12*, pp. 0455–0463.

Lommel, A. (2018) 'Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies', in Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (eds.) *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer, pp. 109-127.

Lv, Y. and Li, M. (2015) 'On the New Features of Online Film and TV Subtitle Translation in China', *International Journal of English Linguistics*, *5*(6), p. 122.

Madnani, N. (2007) 'Getting started on natural language processing with Python', *ACM Crossroads*, *13*(4), p. 5.

Mangiron, C. (2016) 'Reception of game subtitles: an empirical study', *The Translator*, *22*(1), pp. 72-93.

Mao, E., Srite, M., Bennett Thatcher, J. and Yaprak, O. (2005) 'A research model for mobile phone service behaviors: empirical validation in the US and Turkey', *Journal of Global Information Technology Management*, *8*(4), pp. 7-28.

Matamala, A. (2015) 'The ALST project: technologies for audiovisual translation', *Proceedings of the 37th edition of Translating and the Computer Conference (TC37)*. London, UK, 26-27 November. AsLing, pp. 1-24. Available at: https://ddd.uab.cat/pub/presentacions/2015/145198/L2-11.55-Matamala-ALST.pdf (Accessed: 11 December 2019)

Matthews, B. and Ross, L. (2010) *Research methods: A Practical Guide for the Social Sciences*. Edinburgh: Pearson Education Ltd.

Mayer, R.E. (2017) 'Using multimedia for e-learning', *Journal of Computer Assisted Learning*, *33*(5), pp. 403-423.

McCoy, S., Everard, A. and Jones, B. M. (2005) 'An examination of the technology acceptance model in Uruguay and the US: a focus on culture', *Journal of Global Information Technology Management*, *8*(2), pp. 27–45.

McEnery, T. and Hardie, A. (2011) *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Melby, A. (2002) 'The translator workstation', in Newton, J. (ed.) *Computers in Translation.* London: Routledge, pp. 167-185.

Mesa-Lao, B. (2012) 'The next generation translator's workbench: post-editing in CASMACAT v. 1.0', *Proceedings of the 34th Translating and the Computer Conference, ASLIB*. London, UK, 29-30 November. AsLing, pp. 1-11.

Miró, J.D.V., Silvestre-Cerdà, J.A., Civera, J., Turró, C. and Juan, A. (2015) 'Efficient generation of high-quality multilingual subtitles for video lecture repositories', in Conole, G., Klobučar, T., Rensing, C., Konert, J., and Lavoué, E. (eds.) *Design for Teaching and Learning in a Networked World.* Berlin: Springer, pp. 485-490.

Miró, J.D.V., Baquero-Arnal, P., Civera, J., Turró, C. and Juan, A. (2018) 'Multilingual videos for moocs and oer', *Journal of Educational Technology & Society*, *21*(2), pp. 1-12.

Mitchell, L. (2015) 'The potential and limits of lay post-editing in an online community', *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT2015).* Antalya, Turkey, 11-13 May. EAMT, pp. 67-74.

MOE (2018) *Ministry of Education launches 490 national selected online open courses - Ministry of Education of the People's Republic of China*. Available at: http://en.moe.gov.cn/news/press_releases/201801/t20180119_325124.html (Accessed: 5 September 2019).

Moorkens, J. and O'Brien, S. (2013) 'User attitudes to the post-editing interface', *Proceedings of Machine Translation Summit XIV: Second Workshop on Post-editing Technology and Practice.* Nice, France, 2-6 September 2013. IAMT, pp. 19–25.

Moorkens, J., O'Brien, S., da Silva, I.A., de Lima Fonseca, N.B. and Alves, F. (2015) 'Correlations of perceived post-editing effort with measurements of actual effort', *Machine Translation*, *29*(3-4), pp. 267-284.

Morien, D. (2007) *Business Statistics*. South Melbourne: Cengage Learning Australia.

Morse, J.M. (2016) *Mixed method design: Principles and procedures*. Abingdon: Routledge.

Müller, M. and Volk, M. (2013) 'Statistical machine translation of subtitles: From OpenSubtitles to TED', in Gurevych, I., Biemann, C. and Zesch, T. (eds.) *Language processing and knowledge in the Web.* Berlin: Springer, pp. 132-138.

MQM (2015) *Multidimensional Quality Metrics Definition*. Available at: http://www.qt21.eu/mqm-definition/definition-2015-12-30.html (Accessed: 9 September 2019).

Murray, S. (2019) Moocs struggle to lift rock-bottom completion rates, *Financial Times*, 4 March. Available at: https://www.ft.com/content/60e90be2-1a77-11e9-b191-175523b59d1d (Accessed: 5 September 2019).

Nagao, M. (1984) A framework of a mechanical translation between Japanese and English by analogy principle, in Elithorn, A. and Banerji, R. (eds.) *Artificial and Human Intelligence*. Amsterdam: Elsevier Science Publishers. Available at: http://www.mt-archive.info/Nagao-1984.pdf (Accessed: 3 December 2019)

Nida, E.A. (1964) *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Leiden: Brill Archive.

Niu, H., Li, J. and Zhao, Y. (2019) 'SmartBullets: A Cloud-Assisted Bullet Screen Filter based on Deep Learning', *arXiv preprint arXiv:1905.05925*. Available at: https://arxiv.org/abs/1905.05925 (Accessed: 11 October 2019)

Nobre, A., Mallmann, E.M., Nobre, V. and Mazzardo, M.D. (2018) 'MOOC and OER: identity management', in Queirós, R. (ed.) *Emerging Trends, Techniques, and Tools for Massive Open Online Course (MOOC) Management.* Hershey: IGI Global, pp. 1-23.

Oakes, M.P. and Farrow, M. (2006) 'Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries', *Literary and Linguistic Computing*, *22*(1), pp.85-99.

O'Brien, S. (2006) 'Pauses as indicators of cognitive effort in post-editing machine translation output', *Across Languages and Cultures*, *7*(1), pp. 1–21.

O'Brien, S. (2009) 'Eye tracking in translation process research: methodological challenges and solutions', *Methodology, Technology and Innovation in Translation Process Research*, *38*, pp. 251–266.

O'Brien, S. (2010) 'Introduction to post-editing: Who, what, how and where to next', *Presentado en The Ninth Conference of the Association for Machine Translation in the Americas.* Denver, Colorado, 31 October – 5 November 2010. Available at: https://amta2010.amtaweb.org/AMTA/papers/6-01-ObrienPostEdit.pdf (Accessed: 13 October 2019)

O'Brien, S. (2011) 'Towards predicting post-editing productivity', *Machine translation*, *25*(3), pp. 197–215.

O'Hagan, M. (2011) 'Community translation: Translation as a social activity and its possible consequences in the advent of Web 2.0 and beyond', *Linguistica Antverpiensia*, *10*, pp. 11–23.

Ong, C.S., Lai, J.Y. and Wang, Y.S. (2004) 'Factors affecting engineers' acceptance of asynchronous e-learning systems in high-tech companies', *Information & management*, *41*(6), pp. 795-804.

Onwuegbuzie, A.J. and Teddlie, C. (2003) 'A framework for analyzing data in mixed methods research', *Handbook of mixed methods in social and behavioral research*, *2*, pp. 397–430.

Orero, P. (2018) *Deliverable IO4 MOOC | ACT / Accessible Culture & Training*. Available at: http://pagines.uab.cat/act/content/io4-mooc (Accessed: 8 September 2019).

Orero, P., Doherty, S., Kruger, J.L., Matamala, A., Pedersen, J., Perego, E., Romero-Fresco, P., Rovira-Esteva, S., Soler-Vilageliu, O. and Szarkowska, A. (2018) 'Conducting experimental research in audiovisual translation (AVT): A position paper', *JosTrans: The Journal of Specialised Translation*, (30), pp.105-126. Available at: https://www.jostrans.org/issue30/art_orero_et_al.php (Accessed: 12 December 2019)

Orlič, D., Cestnik, B. and Urbančič, T. (2014) 'What can videolectures. net and translectures do for opening higher education to the multicultural world', *Proceedings of the 15th International Conference on Computer Systems and Technologies*. Ruse, Bulgaria, 27-28 June 2014. ACM, pp. 409–416.

Orrego-Carmona, D. (2015) *The reception of (non) professional subtitling.* PhD thesis. Universitat Rovira i Virgili.

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, doi: http://dx.doi.org/10.1556/084.2016.17.2.2.

Ortiz-Boix, C. and Matamala, A. (2015) 'Quality assessment of post-edited versus translated wildlife documentary films', *Proceedings of 4th Workshop on Post-editing Technology and Practice (WPTP4)*. Miami, USA, 3 November 2015. AMTA, pp. 16-30.

Ortiz-Boix, C. and Matamala, A. (2016) 'Post-editing wildlife documentary films: A new possible scenario?' *Jostrans: The Journal of Specialized Translation*, (26), pp. 187-210. Available at: http://www.jostrans.org/issue26/art_ortiz.php (Accessed: 4 December 2019)

Owczarzak, K., van Genabith, J. and Way, A. (2007) 'Dependency-based automatic evaluation for machine translation', *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*. New York, USA, 26 April 2007. Association for Computational Linguistics, pp. 80-87.

Padó, S., Galley, M., Jurafsky, D. and Manning, C. (2009) 'Robust machine translation evaluation with entailment features', *Proceedings of the Joint Conference of the 47th*

Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore*,* 2-7 August 2009. Association for Computational Linguistics, pp. 297-305.

Panchenko, L. (2013) *Massive open online course as an alternative way of advanced training for higher educational establishment professors*. Available at: http://pedagogicaljournal.luguniv.edu.ua/archive/2013/N1/articles/3/Panchenko_eng.pdf (Accessed: 11 October 2019)

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2002) 'BLEU: a method for automatic evaluation of machine translation', *Proceedings of the 40th annual meeting on association for computational linguistics.* Philadelphia, Pennsylvania, USA, 7 – 12 July 2002. Association for Computational Linguistics, pp. 311-318.

Park, N., Roman, R., Lee, S. and Chung, J.E. (2009) 'User acceptance of a digital library system in developing countries: An application of the Technology Acceptance Model', *International journal of information management*, *29*(3), pp. 196-209.

Perego, E. and Del Missier, F. (2008) 'Is a Reading Situation Better than Another for Subtitled Film Viewers', in Lavaur, J. and Serban, A. (eds.) *Handbook of Abstracts, Audiovisual Translation: Multidisciplinary Approaches (La traduction audiovisuelle: Approches pluridisciplinaires).* Montpellier: De Boeck, p. 34.

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

Perego, E., Laskowska, M., Matamala, A., Remael, A., Robert, I.S., Szarkowska, A., Vilaró, A. and Bottiroli, S. (2016) 'Is subtitling equally effective everywhere? A first cross-national study on the reception of interlingually subtitled messages', *Across Languages and Cultures*, *17*(2), pp. 205-229.

Phelan, M. (2017) 'Analytical assessment of legal translation: a case study using the American Translators Association framework', *The Journal of Specialized Translation*, *27*, pp. 189-210.

Pierce, J.R., Carroll, J.B., Hamp, E.P., Hays, D.G., Hockett, C.F., Oettinger, A.G. and Perlis, A. (1966) *Languages and machines: computers in translation and linguistics. Research report, Automatic Language Processing Advisory Committee (ALPAC)*. Available at: http://www.mt-archive.info/ALPAC-1966.pdf (Accessed: 13 October 2019)

Pituch, K.A. and Lee,Y. (2006) 'The influence of system characteristics on e-learning use', *Computers & Education*, *47*(2), pp. 222–244.

Plitt, M. and Masselot, F. (2010) 'A productivity test of statistical machine translation post-editing in a typical localisation context', *The Prague bulletin of mathematical linguistics*, *93*, pp.7-16.

Poole, A. and Ball, L.J. (2006) 'Eye tracking in HCI and usability research', in Ghaoui, C. (ed.) *Encyclopedia of human computer interaction*. Denmark: Idea Group Reference, pp. 211-219.

Popović, M. and Ney, H. (2009) 'Syntax-oriented evaluation measures for machine translation output', *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, 30-31 March 2009. Association for Computational Linguistics, pp. 29-32.

Popović, M. (2011) 'Hjerson: An open source tool for automatic error classification of machine translation output', *The Prague Bulletin of Mathematical Linguistics*, *96*, pp. 59–67.

Popović, M. (2015) 'chrF: character n-gram F-score for automatic MT evaluation', *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, 17-18 September 2015. ACL, pp. 392-395.

Popović, M. (2018) 'Error classification and analysis for machine translation quality assessment', in Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (eds.) *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer, pp. 129-158.

Przybocki, M., Sanders, G. and Le, A. (2006) 'Edit distance: a metric for Machine Translation evaluation', *LREC-2006: Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy, 22-28 May 2006. European Language Resources Association, pp. 2038–2043.

Qi, B. (2019) 'Analysis of the Combination of Information Technology and Government Functions – Take China MOOC as an Example', *MATEC Web of Conferences*, *267*(02015), pp. 1-4. doi: 10.1051/matecconf/201926702015.

Qi, J., Li, L., Li, Y. and Shu, H. (2009) 'An extension of technology acceptance model: Analysis of the adoption of mobile data services in China', *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research*, *26*(3), pp. 391-407.

Qian, Y. (2014) *Zhongguo De Yizhipian 'Si' Le Ma? Shang Shiji De Huihuang Yijing Bu Zai [Are the Chinese Dubbed Films 'Dead'? The Glory of the Last Century is Gone].* Available at: http://hlj.ifeng.com/culture/art/detail_2014_12/29/3350755_0.shtml (Accessed: 11 September 2019).

Qin, Y. and Specia, L. (2015) 'Truly exploring multiple references for machine translation evaluation', *Proceedings of the 18th Annual Conference of the European*

Association for Machine Translation. Antalya, Turkey, 11-13 May 2015. EAMT, pp. 113-120.

Rayson, P. and Garside, R. (2000) 'Comparing corpora using frequency profiling', *Proceedings of the workshop on Comparing corpora*. Hong Kong, China, 7 October 2000. ACL, pp. 1-6.

Rayner, K. and McConkie, G.W. (1976) 'What guides a reader's eye movements?', *Vision research*, *16*(8), pp. 829-837.

Rembert-Lang, L.D. (2010) 'Reinforcing the tower of babel: The impact of copyright law on fansubbing', *Intell. Prop. Brief*, *2*, p.21.

Romero-Fresco, P. (ed.) (2015) *The Reception of Subtitles for the Deaf and Hard of Hearing in Europe*. Bern: Peter Lang.

Ross, S.M. (2014) *Introduction to Probability and Statistics for Engineers and Scientists*. Cambridge: Academic Press.

Roturier, J., Mitchell, L. and Silva, D. (2013) 'The ACCEPT post-editing environment: A flexible and customisable online tool to perform and analyse machine translation post-editing', *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*. Nice, France, 2 September. IAMT, pp. 119–128.

Roturier, J. (2006) *An investigation into the impact of controlled English rules on the comprehensibility, usefulness and acceptability of machine-translated technical documentation for French and German users.* PhD thesis. Dublin City University. Available at: http://doras.dcu.ie/18190/ (Accessed: 13 October 2019)

Sager, J.C. (1989) 'Quality and standards: The evaluation of translations', *The translator's handbook*, *2*, pp.91-102.

Saldanha, G. and O'Brien, S. (2013) *Research methodologies in translation studies*. Abingdon: Routledge.

Sanchis-Trilles, G., Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., Germann, U., González-Rubio, J., Hill, R.L., Koehn, P. and Leiva, L.A. (2014) 'Interactive translation prediction versus conventional post-editing in practice: a study with the CasMaCat workbench', *Machine Translation*, *28*(3-4), pp. 217-235.

Schauffler, S.F. (2012) *Investigating Subtitling Strategies for the Translation of Wordplay in Wallace and Gromit-An Audience Reception Study.* PhD thesis. University of Sheffield.

Schepers, J., Wetzels, M. and de Ruyter, K. (2005) 'Leadership styles in technology acceptance: do followers practice what leaders preach?', *Managing Service Quality: An International Journal*, *15*(6), pp. 496-508.

Schepers, J. and Wetzels, M. (2007) 'A meta-analysis of the technology acceptance model: Investigating subjective norm and moderation effects', *Information & management*, *44*(1), pp. 90-103.

Schmitt, A. and Minker, W. (2012) *Towards Adaptive Spoken Dialog Systems*. Berlin: Springer Science & Business Media.

Schrøder, K.C. (1999) 'The best of both worlds?: media audience research between rival paradigms', in Alasuurari, P. (ed.) *Rethinking the media audience*. London: Sage Publications, pp. 38-68.

Secară, A. (2005) 'Translation evaluation: A state of the art survey', *Proceedings of the eCoLoRe/MeLLANGE workshop.* Leeds, UK, 21-23 March 2005. University of Leeds, pp. 39-44.

Secară, A. (2015) *RU Ready 2 Explor?: Creative Spellings in Subtitling*. PhD thesis. University of Leeds.

Sennrich, R., Birch, A. and Junczys-Dowmunt, M. (2016) *Advances in Neural Machine Translation*. Available at: http://homepages.inf.ed.ac.uk/rsennric/amta2016-tutorial.pdf (Accessed: 13 October 2016)

Shah, D. (2018) *By The Numbers: MOOCs in 2018 — Class Central*. Available at: https://www.classcentral.com/report/mooc-stats-2018/ (Accessed: 5 September 2019).

Shreve, G.M. and Angelone, E. (2010) *Translation and Cognition*. Amsterdam : John Benjamins Publishing.

SMI. (2015). *Begaze manual version 3.5*. Teltow: SMI.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006) 'A study of translation edit rate with targeted human annotation', *Proceedings of association for machine translation in the Americas*. Cambridge, Massachusetts, USA, 8-12 August 2006. AMTA, pp. 1-9.

Snover, M., Madnani, N., Dorr, B.J. and Schwartz, R. (2009) 'Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric', *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, 30-31 March 2009. ACL, pp. 259-268.

Snover, M., Madnani, N., Dorr, B. and Schwartz, R. (2009) 'TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate', *Machine Translation*, *23*(2-3), pp. 117-127.

Specia, L. and Farzindar, A. (2010) 'Estimating machine translation post-editing effort with HTER', *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*. Denver, Colorado, USA, 31 October – 4 November 2010. CNGL, pp. 33-41.

Sriram, S. (2019) *Certificates boost MOOC completion rates*, *University of Michigan News*. Available at: https://news.umich.edu/certificates-boost-mooc-completion-rates/ (Accessed: 5 September 2019).

Starmass (2019) *China tiered cities*. Available at: https://www.starmass.com/china-tiered-cities/ (Accessed: 2 December 2019)

Straub, D., Keil, M. and Brenner, W. (1997) 'Testing the technology acceptance model across cultures: A three country study', *Information & management*, *33*(1), pp. 1–11.

Stymne, S. and Ahrenberg, L. (2012) 'On the practice of error analysis for machine translation evaluation', *Proceedings of LREC 2012.* Istanbul, Turkey, 21-27 May 2012. European Language Resources Association, pp. 1785-1790.

Sullivan, L.E. (2009) *The SAGE Glossary of the Social and Behavioral Sciences*. London: SAGE Publications.

Suojanen, T., Koskinen, K. and Tuominen, T. (2015) 'Usability as a focus of multiprofessional collaboration: A teaching case study on user-centered translation', *Connexions: International Professional Communication Journal, 3*(2), pp. 147-166

Szarkowska, A., Krejtz, I., Pilipczuk, O., Dutka, Ł. and Kruger, J.L. (2016) 'The effects of text editing and subtitle presentation rate on the comprehension and reading patterns of interlingual and intralingual subtitles among deaf, hard of hearing and hearing viewers', *Across Languages and Cultures*, *17*(2), pp. 183-204.

Tahmassebi, A., Gandomi, A.H. and Meyer-Baese, A. (2018) 'An Evolutionary Online Framework for MOOC Performance Using EEG Data', *Proceedings of 2018 IEEE Congress on Evolutionary Computation (CEC)*. Rio de Janeiro, Brazil, 8-13 July 2018. IEEE, pp. 1-8.

TAUS (2010) *Post-editing in Practice*. Available at: https://www.taus.net/think-tank/reports/postedit-reports/postediting-in-practice (Accessed: 9 September 2019).

TAUS (2015) *DQF and MQM Harmonized to Create an Industry-Wide Quality Standard – TAUS*. Available at: https://www.taus.net/academy/news/press-release/dqf-and-mqm-harmonized-to-create-an-industry-wide-quality-standard (Accessed: 9 September 2019).

TAUS (2016a) *TAUS Post-Editing Guidelines*. Available at: https://www.taus.net/think-tank/articles/postedit-articles/taus-post-editing-guidelines (Accessed: 20 January 2016).

TAUS (2016b) *DQF Tools Updated with DQF-MQM Error Types – TAUS*. Available at: https://taus.net/academy/news/press-release/dqf-tools-updated-with-dqf-mqm-error-types (Accessed: 9 September 2019).

TAUS (2017) *Quality Management White Paper – TAUS*. Available at: https://www.taus.net/academy/reports/translate-reports/taus-quality-management-white-paper (Accessed: 9 September 2019).

Tavakoli, M. and Gerami, E. (2013) 'The effect of keyword and pictorial methods on EFL learner's vocabulary learning and retention', *Porta Linguarum*, *19*, pp. 299-316.

Teddlie, C. and Tashakkori, A. (2003) 'Major issues and controveries inthe use of mixed methods in the social and behvioral sciences', in Tashakkori, A. and Teddlie, C. (eds.) *Handbook of mixed methods in social & behavioral research*. London: SAGE, pp. 3–50.

Teixeira, C.S. and O'Brien, S. (2018) 'Overcoming methodological challenges of eye tracking in the translation workplace', *Eye Tracking and Multidisciplinary Studies on Translation*, *143*, p. 33.

Thelen, M. (2008) 'Translation Quality Assessment or Quality Management & Quality Control of Translation?', *Translation and Meaning, Part 8,* pp. 411-424.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A. and Sawaf, H. (1997) 'Accelerated DP based search for statistical translation', *Proceedings of Fifth European Conference on Speech Communication and Technology*. Rhodes, Greece, 22-25 September 1997. European Speech Communication Association, pp. 1-4.

Toral, A. and Way, A. (2018) 'What level of quality can Neural Machine Translation attain on literary text?', in Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (eds.) *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer, pp. 263-287.

Toral, A., Castilho, S., Hu, K. and Way, A. (2018) Attaining the unattainable? Reassessing claims of human parity in neural machine translation, *Proceedings of the Third Conference on Machine Translation*, *Volume 1: Research Papers*. Brussels, Belgium, 31 October – 1 November 2018. Association for Computational Linguistics, pp. 113–123. Available at: https://arxiv.org/abs/1808.10432 (Accessed: 13 October 2019)

Torres-Hostench, O., Moorkens, J., O'Brien, S. and Vreeke, J. (2017) 'Testing interaction with a Mobile MT post-editing app', *Translation & Interpreting*, *9*(2), pp. 138-150.

Toury, G. (1995) *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins Publishing.

Turovsky, B. (2016) *Ten years of Google Translate*. Available at: https://www.blog.google/products/translate/ten-years-of-google-translate/ (Accessed: 3 December 2019)

van Raaij, E.M. and Schepers, J.J. (2008) 'The acceptance and use of a virtual learning environment in China', *Computers & Education*, *50*(3), pp. 838–852.

van Slype, G. (1979) *Critical study of methods for evaluating the quality of machine translation*. Available at: http://aei.pitt.edu/39751/1/A4102.pdf (Accessed: 13 October 2019)

Vasey, M.W. and Thayer, J.F. (1987) 'The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution', *Psychophysiology*, *24*(4), pp. 479-486.

Venkatesh, V. and Davis, F.D. (1996) 'A model of the antecedents of perceived ease of use: Development and test', *Decision sciences*, *27*(3), pp. 451–481.

Venkatesh, V. (2000) 'Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model', *Information systems research*, *11*(4), pp. 342–365.

Venkatesh, V. and Davis, F.D. (2000) 'A theoretical extension of the technology acceptance model: Four longitudinal field studies', *Management science*, *46*(2), pp. 186–204.

Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D. (2003) 'User acceptance of information technology: Toward a unified view', *MIS quarterly*, *27*(3), pp. 425-478.

Venkatesh, V. and Bala, H. (2008) 'Technology acceptance model 3 and a research agenda on interventions', *Decision sciences*, *39*(2), pp. 273–315.

Vermeer, H.J. (1978) 'Ein Rahmen für eine allgemeine Translationstheorie', *Lebende Sprachen*, *23*(3), pp. 99-102.

Vieira, L.N. and Specia, L. (2011) 'A Review of Translation Tools from a Post-Editing Perspective', *Proceedings of the Third Joint EM+/CNGL Workshop Bringing MT to the Users: Research Meets Translators" (JEC '11)*. Luxemburg, 14 July 2011. CNGL, pp. 33–42.

Vilar, D., Xu, J., Luis Fernando, D.H. and Ney, H. (2006) 'Error Analysis of Statistical Machine Translation Output', *Proceedings of LREC2006*. Genoa, Italy, 22-28 May 2006. European Language Resources Association, pp. 697-702.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2015) 'Show and tell: A neural image caption generator', *Proceedings of the IEEE conference on computer vision and pattern recognition.* Boston, MA, USA, 7-12 June 2015. IEEE, pp. 3156-3164.

Volk, M. (2008) *The Automatic Translation of Film Subtitles. A Machine Translation Success*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.490.7635&rep=rep1&type=pdf (Accessed: 13 October 2019)

Volk, M., Sennrich, R., Hardmeier, C. and Tidström, F. (2010) 'Machine translation of TV subtitles for large scale production', *Proceedings of JEC 2010.* Denver, Colorado, USA, 31 October – 4 November 2010. CNGL, pp. 53-62.

W. French, H. (2006) *Chinese Tech Buffs Slake Thirst for U.S. TV Shows - The New York Times*. Available at: https://www.nytimes.com/2006/08/09/world/asia/09china.html (Accessed: 8 September 2019).

Wang, D. (2015) 'Subtitling – From a Chinese Perspective', in Bruti, S. and Perego, E. (eds.) *Subtitling Today: Shapes and Their Meanings*. Cambridge: Cambridge Scholars Publishing, p. 203.

Wang, L., Zhang, X., Tu, Z., Way, A. and Liu, Q. (2016) 'Automatic construction of discourse corpora for dialogue translation', *arXiv preprint arXiv:1605.06770*. Available at: https://arxiv.org/abs/1703.05330 (Accessed: 13 October 2019)

Wang, W. (2013) *Yizhipian Meikuangyuxia, Lao Yanyuan Cheng Reng You Cunzai De Biyao [The Situation of Dubbed Films has Gone From Bad to Worse; the Old Actor Said that There is Still a Need for Them to Exist].* Available at: http://www.chinanews.com/cul/2013/06-24/4962866.shtml (Accessed: 11 September 2019).

Way, A. (2018) 'Quality expectations of machine translation', in Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (eds.) *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer, pp. 159-178.

Wen, W. (2017) *Meijumi Qun: Meijie Xiaofei Yu Rentong Jiangou [American Drama Fans: Media Consumption and Identity Construction].* Linden: Beijing Book Co. Inc.

Wenger, E. (1999) *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.

Williams, L.J. and Abdi, H. (2010) 'Fisher's least significant difference (LSD) test', *Encyclopedia of research design*, *218*, pp. 840-853.

Williams, M. (2013) 'A holistic-componential model for assessing translation student performance and competency', *Mutatis Mutandis: Revista Latinoamericana de Traducción*, *6*(2), pp. 419-443.

Wołk, K. and Marasek, K. (2015) 'Polish-English speech statistical machine translation systems for the IWSLT 2013', *arXiv preprint arXiv:1509.09097.* Available at: https://arxiv.org/abs/1306.2485 (Accessed: 13 October 2019)

Wongseree, T. (2019) 'Understanding Thai fansubbing practices in the digital era: a network of fans and online technologies in fansubbing communities', *Perspectives*, pp. 1-15. doi: https://doi.org/10.1080/0907676X.2019.1639779.

Wu, B. and Chen, X. (2017) 'Continuance intention to use MOOCs: Integrating the technology acceptance model (TAM) and task technology fit (TTF) model', *Computers in Human Behavior*, *67*, pp. 221-232.

Wu, J.H., Wang, S.C. and Lin, L.M. (2007) 'Mobile computing acceptance factors in the healthcare industry: A structural equation model', *International journal of medical informatics*, *76*(1), pp. 66-77.

Wu, W. and Bai, Q. (2018) 'Why Do the MOOC Learners Drop Out of the School? – Based on the Investigation of MOOC Learners on Some Chinese MOOC Platforms', *Proceedings of 2018 1st International Cognitive Cities Conference (IC3)*. Okinawa, Japan, 7-9 August 2018. IEEE, pp. 299-304.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. and Klingner, J. (2016) 'Google's neural machine translation system: Bridging the gap between human and machine translation'*, arXiv preprint arXiv:1609.08144.* Available at: https://arxiv.org/abs/1609.08144 (Accessed: 13 October 2019)

Wray, A. (2002) *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Ye, D., Huang, P., Hong, K., Tang, Z., Xie, W. and Zhou, G. (2015) 'Chinese microblogs sentiment classification using maximum entropy', *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*. Beijing, China, 30-31 July 2015. ACL, pp. 171-179.

Yoon, C. (2009) 'The effects of national culture values on consumer acceptance of e-commerce: Online shoppers in China', *Information & Management*, *46*(5), pp. 294–301.

Yuan, S. and Liu, X. (2014) 'Zhongguo daxue MOOC shijian xianzhuang ji gongyou wenti [The current situation and common problems of MOOC practice in Chinese universities]', *Xiandai Yuancheng Jiaoyu Yanjiu [Modern Distance Education Research]*, p. 20.

Zehnalová, J. (2013) 'Tradition and Trends in Translation Quality Assessment', in Zehnalová, J., Molnár, O. and Kubánek, M. (eds.) *Tradition and Trends in Trans-Language Communication*. Olomouc: Univerzita Palackého, pp. 41-58.

Zeman, D., Fishel, M., Berka, J. and Bojar, O. (2011) 'Addicter: what is wrong with my translations?', *The Prague Bulletin of Mathematical Linguistics*, *96*, pp. 79-88.

Zhang, Q. and Lu, Z. (2014) 'The writing of Chinese characters by CFL learners: Can writing on Facebook and using machine translation help?', *Language Learning in Higher Education*, *4*(2), pp. 441-467.

Zhang, Q., Liu, X. and Fu, J. (2018) 'Neural networks incorporating dictionaries for Chinese word segmentation', *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, Louisiana, USA, 2-7 February 2018. Association for the Advancement of Artificial Intelligence, pp. 5682-5689.

Zhao, S. (2013). *Shengtai Fanyixue Shijiao Xia De Waiguo Gongkaike Zimu Fanyi [Subtitle Translation of Foreign Open Courses from the Perspective of Ecological Translation Studies]*. PhD thesis. Beijing International Studies University.

Zhechev, V. (2012) 'Machine translation infrastructure and post-editing performance at Autodesk', *Proceedings of AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*. San Diego, USA, 28 October – 1 November 2012. AMTA, pp. 87–96.

Zheng, W., Wang, W., Liu, D., Zhang, C., Zeng, Q., Deng, Y., Yang, W., He, P. and Xie, T. (2019) 'Testing untestable neural machine translation: an industrial case', *Proceedings of the 41st International Conference on Software Engineering: Companion.* Montreal, QC, Canada, 25-31 May 2019. IEEE, pp. 314-315.

Zhou, S. (2018) *Has AI surpassed humans at translation? Not even close!* Available at: https://www.skynettoday.com/editorials/state_of_nmt (Accessed: 8 October 2019).

Appendices

Contents

Appendix A1: Informed Consent Form (English)

Research study title: End Users' Acceptance of Machine Translation in a MOOC Environment (EN-ZH)

Institution: School of Applied Language and Intercultural Studies, Dublin City University

Principal investigators: Ke Hu, Dr Sharon O'Brien and Prof Dorothy Kenny

Purpose of the research: To investigate to what extent the end users accept the machine translated subtitles for MOOCs

- I have read the Plain Language Statement and I have understood the information provided in it.
- I have been given the opportunity to ask questions to the researchers by email, and my questions and concerns have been answered by the researchers.
- I am aware that I will be asked to complete a pre-recruitment questionnaire and an online English test
- I am aware that I will probably be asked to watch a MOOC video – and that data will be recorded during this time using an eye-tracker – and a post-experiment questionnaire, and that my participation will last up to a maximum of 45 minutes.
- I confirm that my involvement in the study is voluntary and I am aware that I may withdraw from this study at any point.
- I am aware that my answers are confidential, and I understand that confidentiality of the information provided is subject to legal limitations.

☐ By ticking this box, you confirm that you have read and understood the information in this section and you consent to take part in this research project.

Appendix A2: Informed Consent Form (Chinese)

研究题目：慕课环境下终端用户对机器翻译的接受度（英-中）

学校：都柏林城市大学应用语言与跨文化研究学院

研究人员：胡克，Sharon O'Brien 博士，Dorothy Kenny 教授

研究目的：调查终端用户对慕课的机器翻译字幕的接受程度

- 我已阅读该声明，我了解其包含的信息。
- 我有机会通过电子邮件向研究人员提出问题，研究人员已经回答了我的问题和疑虑。
- 我知道我将要完成一份实验前测问卷和一份在线英文测试。
- 我知道我很可能会观看一个慕课视频，过程中的数据将会被眼动仪记录，我还可能要完成一份实验后问卷，我的参与过程将不超过 45 分钟。
- 我确认我自愿参与该研究，我知道我随时可以退出该研究。
- 我知道我的答案是保密的，我明白我所提供的信息的机密性处于法律限制范围内。

☐ 勾选这里表示您确认您已经阅读并了解此部分的信息，您同意参与本研究项目。

Appendix B1: Plain Language Statement (English)

This study is part of a research project called “Human Factors in Machine Translation”. This study is being funded by the Science Foundation Ireland through the ADAPT project at Dublin City University and is being carried out at Dublin City University (or the named Chinese University) by Ke Hu (ke.hu2@mail.dcu.ie) under the supervision of Dr Sharon O’Brien and Prof Dorothy Kenny.

To be eligible to participate in this experiment, you must be a native speaker of Chinese, you must be aged 18 or over, and you must be a registered undergraduate. As a participant, you will be firstly asked to complete a pre-recruitment questionnaire and an online English test. Participants who meet the requirements will be asked to watch a MOOC video using an eye-tracker, and complete a post-experiment questionnaire. Your entire participation will take no longer than 45 minutes of your time.

We are required by DCU’s Research Ethics Committee to provide you with the following additional information concerning your participation in the study:

We anticipate no potential risks to you from involvement in this research study and we will make all the necessary arrangements to protect the anonymity and confidentiality of the data. Each participant will be assigned a number before we start to process the data, so that your identity will never be visible during the analysis and dissemination of results. As we are not assessing your abilities or competencies, but rather your use and acceptance of machine translated subtitles, we anticipate that the collected data cannot be damaging to you in any way. Nonetheless, you are advised that the confidentiality of the information provided cannot always be guaranteed by researchers and can only be protected within the limitations of the law – i.e., it is possible for data to be subject to subpoena, freedom of information claim or mandated reporting by some professions.

We anticipate that you might benefit indirectly from this study as our aim is to research how well the machine translated subtitles can work for people. As a participant, your

participation will help the assessment and improvement of machine translation. In turn, it may bring you more opportunities to watch more foreign MOOCs in the future and be beneficial to your self-improvement.

During this study, the data will be handled exclusively by the three researchers named in this invitation to participate. The study is scheduled to be completed by August 2019, and you will be informed of its outcome by means of an infographic by then. You will also have the option to have a more detailed, plain language report on direct request to the researchers. Afterwards, the data will be disposed in a way that protects the security and confidentiality of the data.

Your involvement in this research study is voluntary and you may withdraw from the study at any point without repercussion. If you have further questions, please do not hesitate to contact the researcher by sending an email to the address provided above.

If you have any concerns about this study and wish to contact an independent person, please contact:

The Secretary, Dublin City University Research Ethics Committee, c/o Research and Innovation Support, Dublin City University, Dublin 9. Tel +353 1 7008000

Thank you in advance,

Ke Hu
Sharon O'Brien
Dorothy Kenny

Appendix B2: Plain Language Statement (Chinese)

声明

该研究是《机器翻译的人为因素》研究项目的一部分，由爱尔兰科学基金会通过在都柏林城市大学的 ADAPT 项目资助。该项目由胡克（ke.hu2@mail.dcu.ie）在 Sharon O'Brien 博士和 Dorothy Kenny 教授的指导下于都柏林城市大学开展。

参加本实验的要求如下：您必须以中文为母语，您必须年满 18 岁，您必须是在校本科生。作为参与者，您将完成一份实验前测问卷和一份在线英语测试。符合要求的参与者将在安装有眼动仪的电脑上观看一个慕课视频，并完成一份实验后问卷。您的全部参与时间将不超过 45 分钟。

在 DCU 研究伦理委员的要求下，以下是我们提供的关于您参与本研究的其他信息：

我们预计您参与本研究不存在任何潜在风险，我们将采取一切必要措施来保护数据的匿名性和保密性。在我们开始处理数据之前，每个参与者将被分配一个号码，以便在分析和宣传结果期间，您的身份永不可见。由于我们的目的并非评测您的能力，而是评测您对机器翻译字幕的使用和接受度，我们预计所收集的数据不会对您造成任何伤害。尽管如此，研究人员无法一直保证您所提供的信息的机密性，您的信息只会在法律范围内受到保护。

我们预计您可能会间接受益于本研究，因为我们的目标是研究机器翻译字幕的效果。作为参与者，您的参与将有助于机器翻译的评测和改进。作为回报，它可能会让您在未来有更多机会观看更多的外国慕课，有利于您的自我完善。

在本研究中，数据将由本邀请函中的三名研究人员专门处理。该研究预计将于 2019 年 8 月前完成，届时研究成果将会以信息图的方式告知您。您还可以直接要求研究人员提供更详细的报告。之后，数据将会在安全、保密的情况下被处理掉。

您参与本研究是自愿的，您可以随时退出研究，不会造成任何不良后果。如果您还有其他问题，请随时联系研究人员，发送电子邮件至上文地址。

如果您对本研究有任何疑问并希望联系有关人员，联系方式如下：

The Secretary, Dublin City University Research Ethics Committee, c/o Research and Innovation Support, Dublin City University, Dublin 9.
电话： +353 1 7008000

谢谢！

胡克
Sharon O’Brien
Dorothy Kenny

Appendix C1: Pre-task Questionnaire (English)

1) Are you a native speaker of Chinese?
   - Yes
   - No

2) Age:

3) Sex:
   - Male
   - Female

4) What year are you in at university?

   _____

5) What is your major?

   _____

6) Do you have experience in using machine translation applications (Google Translate/ Baidu Translate etc.)?
   - Yes
   - No

   (If you replied 'yes', please go to question 7. If you replied 'no', please go to question 10.)

7) When did you start to use machine translation?

   _____

8) How frequently do you use it?

   _____

9) What do you think of the quality of machine translation?

   _____

10) Please indicate the reason why you don't use machine translation.

    _____

11) Do you have experience in watching movies or other videos with machine translated subtitles?
    - Yes
    - No
    - Uncertain

12) Do you have experience in learning via MOOCs?
    - Yes
    - No

    (If you replied 'yes', please go to question 13. If you replied 'no', you can click on the "Submit" and complete the questionnaire.)

13) Do you have experience in learning via MOOCs with machine translated subtitles?
    - Yes
    - No
    - Uncertain

14) Have you studied the MOOC entitled "Sit Less, Get Active" before?
    - Yes
    - No

15) Do you believe the machine translated subtitles will adequately transfer the meaning of the source text?
    - Yes
    - No

Appendix C2: Pre-task Questionnaire (Chinese)

1) 你的母语是中文吗?
   - 是
   - 否

2) 您的年龄?

3) 您的性别是?
   - 男
   - 女

4) 您现在是大学几年级?

   _____

5) 您的专业是什么?

   _____

6) 您有没有使用过机器翻译应用（例如谷歌翻译、百度翻译等）?
   - 有
   - 没有

   （如果您的回答是“有”，请移步第 7 题；如果您的回答是“没有”，请移步第 10 题。）

7) 您从什么开始使用机器翻译的?

   _____

8) 您使用机器翻译的频率是?

   _____

9) 您认为机器翻译的质量如何?

   _____

10) 请列出您不使用机器翻译的原因。

   _____

11) 您有没有看过配有机器翻译字幕的电影或其他视频?

- 有
- 没有
- 不确定

12) 您有没有学习过慕课?

- 有
- 没有

（如果您的回答是“有”，请继续答题；如果您的回答是“没有”，您可以点击“提交”完成本问卷。）

13) 您有没有学习过配有机器翻译字幕的慕课?

- 有
- 没有
- 不确定

14) 您有没有学习过“少坐，多活动”这门慕课?

- 有
- 没有

15) 您认为机器翻译字幕会充分传达源语言的内容吗?

- 会
- 不会

Appendix D1: Post-task Questionnaire for Pilot Study (English)

Part I: Comprehension Testing

1) What counts as physical activity?
   A. Dancing
   B. Walking the dog
   C. Climbing the stairs
   D. Housework or gardening
   E. All of the above

2) Which example below counts as a muscle strengthening activity?
   A. Running
   B. Yoga
   C. Lifting weights at the gym
   D. Gardening

3) Which of the following would NOT be considered as a sedentary behaviour?
   A. Driving in your car
   B. Sitting at a desk working
   C. Standing on the train to work
   D. Reading a book on the sofa

4) Which of the following is not active transport?
   A. Walking
   B. Cycling
   C. Driving a car
   D. None of the above

5) Which of the following is not caused by sedentary behaviour?
   A. Obesity
   B. Type 2 diabetes
   C. Eye disease
   D. Some cancers
   E. Premature mortality

6) Which of the following is not exercise?
   A. Training to take part in a marathon
   B. Walking along the pavement

C. Going to gym for strength training
D. Going to a yoga class

7) According to physical activity guidelines, children should do 60 minutes of activity every day, that's ____
   A. Running
   B. Jumping
   C. Both of the above

8) According to physical activity guidelines, adults should ____
   A. have moderate activity at least 150 minutes every week
   B. have strength building activities
   C. minimize the time they spend sitting
   D. All of the above

9) True or false: Doing moderate activity is enough to meet the physical activity guidelines.
   A. True
   B. False

10) True or false: If you want to do physical activity only 10 minutes a day, 5 minutes a day, that's not going to be helpful.
   A. True
   B. False

11) True or false: Children get more active as they age.
   A. True
   B. False

12) True or False: Physical activity is a very specific subset of exercise.
   A. True
   B. False

Part II: Attitude Survey

13) The subtitles allow me to fully understand the contents of the MOOC.

☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

14) The subtitles are useful to me.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

15) The subtitles are easy to understand.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

16) Interacting with the subtitles does not require a lot of my mental effort.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

17) I would find it easy to get the information I need from subtitles.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

18) The subtitles are clear and understandable.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

19) I enjoy reading the subtitles.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

20) I am satisfied with the subtitles.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

21) If I have a chance, I would use machine translation to translate English subtitles in the future, because I know it will do a good job.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

22) I would recommend machine translation to my friends if they need to translate subtitles.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

I could comprehend the subtitles…

23) If there was no one around to tell me what to do as I go.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

24) If I could call someone for help if I got stuck.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

25) If I had a lot of time.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

26) If I had just the built-in help facility (e.g.: online dictionary) for assistance.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

Appendix D2: Post-task Questionnaire for Pilot Study (Chinese)

第一部分： 理解测试

1) 下列哪一项属于身体活动?
   A. 跳舞
   B. 遛狗
   C. 爬山
   D. 做家务或园艺
   E. 以上全部

2) 下列哪一项属于肌肉增强活动?
   A. 跑步
   B. 瑜伽
   C. 举重
   D. 园艺

3) 下列哪一项不属于久坐?
   A. 开车
   B. 坐在桌前办公
   C. 站在火车上
   D. 坐在沙发上看书

4) 下列哪一项不属于活跃的交通方式?
   A. 走路
   B. 骑自行车
   C. 自驾车
   D. 以上全部

5) 下列哪一项不会由久坐导致?
   A. 肥胖
   B. 2 型糖尿病
   C. 眼科疾病
   D. 某些癌症
   E. 过早死亡

6) 下列哪一项不属于锻炼？
   A. 为了参加马拉松开始跑步
   B. 沿着人行道走
   C. 去健身房进行肌肉增强活动
   D. 上瑜伽课

7) 根据身体活动指南，孩子每天都要进行 60 分钟的活动，即 ____
   A. 跑步
   B. 跳跃
   C. 以上全部

8) 根据身体活动指南，成年人应该 ____
   A. 每周进行至少 150 分钟的适度活动
   B. 进行力量训练活动
   C. 最少化坐着的时间
   D. 以上全部

9) 真假：做适度活动就足以达到身体活动指南的要求
   A. 真
   B. 假

10) 真假：如果你每天只做十分钟或五分钟的身体活动，那并没什么帮助
   A. 真
   B. 假

11) 真假：孩子越长大活动量越大
   A. 真
   B. 假

12) 真假：身体活动是锻炼的一个特别的子集
   A. 真
   B. 假

第二部分: 态度调查

13) 字幕让我完全掌握了慕课的内容

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

14) 字幕对我很有用
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

15) 字幕很容易理解
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

16) 阅读字幕不需要花费我太多脑力
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

17) 从字幕中获取有用信息对我来说不难
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

18) 字幕明白易懂
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

19) 我享受阅读字幕的过程
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

20) 我对字幕很满意
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

21) 如果将来有机会，我会使用机器翻译来翻译英文字幕，因为我知道它会翻译的不错
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

22) 如果我朋友需要翻译字幕，我会推荐他们使用机器翻译
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

23) 即使身边没有人指导，我也能理解字幕
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

24) 如果遇到困难时能找到人帮忙，我就能理解字幕
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

25) 如果有充足的时间，我就能理解字幕
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

26) 如果电脑有内置的求助软件（例如在线词典等），我就能理解字幕
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

Appendix E1: Post-task Questionnaire for Main Experiment (English)

Part I: Comprehension Testing

1) What counts as physical activity?
   A. Dancing
   B. Walking the dog
   C. Climbing the stairs
   D. Housework or gardening
   E. All of the above

2) Which of the following benefit of physical activity is not mentioned in the video?
   A. It can help reduce our risk of multiple diseases.
   B. It can help us to maintain healthy weight.
   C. It can help us to improve the quality of our life.
   D. None of the above.

3) Which of the following statement about physical activity is not right?
   A. It is any movement that uses energy.
   B. It is not structured.
   C. It is different from exercise.
   D. It is pursued for fitness benefits.

4) Which example below counts as a muscle strengthening activity?
   A. Running
   B. Yoga
   C. Lifting weights at the gym
   D. Gardening

5) Which of the following is not active transport?
   A. Walking
   B. Cycling
   C. Driving a car
   D. None of the above

6) Which of the following is not caused by sedentary behaviour?
   A. Obesity

B. Type 2 diabetes
C. Eye disease
D. Some cancers
E. Premature mortality

7) Which of the following is not exercise?
A. Training to take part in a marathon
B. Walking along the pavement
C. Going to gym for strength training
D. Going to a yoga class

8) According to physical activity guidelines, children should do 60 minutes of activity every day, that's ____
A. Running
B. Jumping
C. Both of the above

9) According to physical activity guidelines, adults should ____
A. have moderate activity at least 150 minutes every week
B. have strength building activities
C. minimize the time they spend sitting
D. All of the above

10) True or false: Doing moderate activity is enough to meet the physical activity guidelines.
A. True
B. False

11) True or false: If you want to do physical activity only 10 minutes a day, 5 minutes a day, that's not going to be helpful.
A. True
B. False

12) True or false: Children get more active as they age.
A. True
B. False

13) True or False: Physical activity is a very specific subset of exercise.
A. True
B. False

Part II: Attitude Survey

14) The subtitles allow me to fully understand the contents of the MOOC.

☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

15) The subtitles are useful to me.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

16) The subtitles are easy to understand.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

17) Interacting with the subtitles does not require a lot of my mental effort.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

18) I would find it easy to get the information I need from subtitles.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

19) The subtitles are clear and understandable.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

20) I enjoy reading the subtitles.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

21) I am satisfied with the subtitles.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

22) If I have a chance, I would use machine translation to translate English subtitles in the future, because I know it will do a good job.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

23) I would recommend machine translation to my friends if they need to translate subtitles.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

I could comprehend the subtitles…

24) If there was no one around to tell me what to do as I go.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

25) If I could call someone for help if I got stuck.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

26) If I had a lot of time.
☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

27) If I had just the built-in help facility (i.e.: online dictionary) for assistance.

☐Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

Appendix E2: Post-task Questionnaire for Main Experiment (Chinese)

第一部分： 理解测试

1) 下列哪一项属于身体活动?
   A. 跳舞
   B. 遛狗
   C. 爬山
   D. 做家务或园艺
   E. 以上全部

2) 下列哪一项关于身体活动的好处没有在视频中提到?
   A. 有助于减少患病风险
   B. 有助于维持健康体重
   C. 有助于提升生活品质
   D. 以上全部都被提及

3) 下列哪一项关于身体活动的陈述有错?
   A. 消耗能量的任何活动都是身体活动
   B. 它不是结构化的
   C. 它与锻炼不同
   D. 它的目的是为了达到健身的好处

4) 下列哪一项属于肌肉增强活动?
   A. 跑步
   B. 瑜伽
   C. 举重
   D. 园艺

5) 下列哪一项不属于活跃的交通方式?
   A. 走路
   B. 骑自行车
   C. 自驾车
   D. 以上全部

6) 下列哪一项不会由久坐导致?
   A. 肥胖

B. 2 型糖尿病
C. 眼科疾病
D. 某些癌症
E. 过早死亡

7) 下列哪一项不属于锻炼？
A. 为了参加马拉松开始跑步
B. 沿着人行道走
C. 去健身房进行肌肉增强活动
D. 上瑜伽课

8) 根据身体活动指南，孩子每天都要进行 60 分钟的活动，即 ____
A. 跑步
B. 跳跃
C. 以上全部

9) 根据身体活动指南，成年人应该 ____
A. 每周进行至少 150 分钟的适度活动
B. 进行力量训练活动
C. 最少化坐着的时间
D. 以上全部

10) 真假：做适度活动就足以达到身体活动指南的要求
A. 真
B. 假

11) 真假：如果你每天只做十分钟或五分钟的身体活动，那并没什么帮助
A. 真
B. 假

12) 真假：孩子越长大活动量越大
A. 真
B. 假

13) 真假：身体活动是锻炼的一个特别的子集
A. 真
B. 假

第二部分: 态度调查

14) 字幕让我完全掌握了慕课的内容

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

15) 字幕对我很有用

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

16) 字幕很容易理解

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

17) 阅读字幕不需要花费我太多脑力

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

18) 从字幕中获取有用信息对我来说不难

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

19) 字幕明白易懂

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

20) 我享受阅读字幕的过程

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

21) 我对字幕很满意

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

22) 如果将来有机会，我会使用机器翻译来翻译英文字幕，因为我知道它会翻译的不错

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

23) 如果我朋友需要翻译字幕，我会推荐他们使用机器翻译

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

24) 即使身边没有人指导，我也能理解字幕

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

25) 如果遇到困难时能找到人帮忙，我就能理解字幕

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

26) 如果有充足的时间，我就能理解字幕

☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同意

27) 如果电脑有内置的求助软件（例如在线词典等），我就能理解字幕
☐非常同意 ☐ 同意 ☐ 中立 ☐ 不同意 ☐ 非常不同

Appendix F: QA Guidelines (English and Chinese)

类型 (Categories)	解释 (Definition)	例子 (Example)
增译 (Addition)	原文中没有的，译文中却有 The target text includes text not present in the source.	He put his hands in the pockets. 他把手放进外套的口袋。（“外套”即增译）
错译 (Mistranslation)	译文错误翻译了原文 The target content does not accurately represent the source text content.	原文 (source text)：not more than 200 times. 译文 (target text)：超过两百次。
省译 (Omission)	原文中有的，译文中却没有 Content is missing from the translation that is present in the source.	原文：You will be staying in this hotel during your visit in Beijing. 译文：你访问期间就住在这家宾馆。（省译了 Beijing）
语法 (Grammar)	句子结构不通顺，读起来拗口 Issues related to the grammar or syntax of the text, other than spelling and orthography.	原文：It's a cold day, so I just want to stay at home. 译文：很冷，我想只在家。
拼写 (Spelling)	汉字写错 Issues related to spelling of words.	比如“我们”写成了“我门”
文风 (Style)	比如把轻松愉快的广告翻译的严肃沉重，或者正式场合用非正式用语 The text has stylistic problems, for example, the translation of a light-hearted and humorous advertising campaign is in a serious and “heavy” style even though specifications said it should match the style of the source text.	原文：Hey, what's up? 译文：您近来可好？
不合中国习俗 (Locale-convention)	比如在译文中使用英式标点或货币符号等 Issues in locale convention relate to the formal compliance of content with locale-specific conventions, such as use of number formats.	原文：I like painting, dancing and singing. 译文：我喜欢画画，跳舞，唱歌。（此处应该讲逗号转为顿号）
术语 (Terminology)	对一个词语或短语没有使用其通用的或约定俗成的翻译 Terminology issues relate to the use of domain- or organization-specific terminology.	比如把 Hollywood 翻译成“红里屋”而非“好莱坞”
过度直译 (Overly-literal)	因为过度直译而导致译文生硬 The translation is overly literal.	原文：I had a good day 译文：我有个好天。
其它 (Others)	不属于以上任何类型的其它错误 Issues that do not belong to any of the issue types listed above.	请在“错误”栏下方的空格注明您所认为的错误类型。 Please add more details in the “Error” column.

严重度 (Severity)	解释 (Definition)	原文例子 (Example of source text): In his closing comments, Mr O'Neill said Apec would try to ensure "free and open trade" in the region by 2020.
轻微 (Minor)	这类问题不影响内容的可用性或可理解性 (Minor issues are issues that do not impact usability or understandability of the content.)	译文 (target text): 奥尼尔先生在闭幕致辞中表示，亚太经合组织将努力确保到 2020 年在该地区实现“自由奔放的贸易”。
重大 (Major)	这类问题会影响内容的可用性或可理解性，但不会使其无法使用 (Major issues are issues that impact usability or understandability of the content but which do not render it unusable.)	译文 (target text): 奥尼尔先生在关门评论中表示，亚太经合组织将努力确保到 2020 年在该地区实现“自由开放贸易”。
致命 (Critical)	这类问题使句子内容不适合使用。例如，一个改变了句子意思的特别糟糕的语法错误 (Critical issues are issues that render the content unfit for use. For example, a particularly bad grammatical error that changes the meaning of the text would be considered critical.)	译文 (target text): 奥尼尔先生在闭幕致辞中表示，亚太经合组织将努力确保“自由开放贸易”在地区一直到 2020 年。

QA 方法（Steps）：

- 请将翻译错误复制黏贴到“错误”栏里，在“错误类型”和“严重度”的下拉栏里选择对应情况。(Please copy and paste the translation error into the “Error” column and select the corresponding case from the drop-downs of the “Error type” and “Severity” columns)

- 如果错误属于“其它”类型，请在“错误”栏里黏贴的翻译错误末尾注明您所认为的错误类型。(If the error belongs to “Others” type, please add more details at the end of the translation error pasted in the “Error” column.)

- 如果您觉得某个错误属于多种类型，请选择最贴切的那个。(If you feel that an error belongs to more than one error type, please choose the one that is the most appropriate.)

- 请记得在第一列空格里给您较为满意的翻译打勾，如果译文相同则不用打勾。(Please remember to tick the box in the first column if you are satisfied with the translation. If the two translations are the same, then there is no need to tick the box)

Appendix G: Source Text and Target Text for Subtitles

ST[91]	Zumba!
RAW	尊巴！
PE	尊巴！
HT	尊巴！
ST	Squash!
RAW	壁球！
PE	壁球！
HT	壁球！
ST	Tennis.
RAW	网球。
PE	网球。
HT	网球。
ST	Sailing!
RAW	帆船！
PE	航行！
HT	航海！
ST	Basketball!
RAW	篮球！
PE	篮球！
HT	篮球！
ST	Gardening.
RAW	园艺。
PE	园艺。
HT	园艺。
ST	Football.
RAW	足球。
PE	足球。
HT	足球。
ST	Stair walking.
RAW	楼梯走。

[91] ST: Source text
RAW: Raw machine translation
PE: Post-edited machine translation
HT: Human translation

PE 走楼梯。
HT 爬楼梯。

ST Yoga.
RAW 瑜伽。
PE 瑜伽。
HT 瑜伽。

ST Dancing!
RAW 跳舞！
PE 跳舞！
HT 跳舞！

ST Frisbee!
RAW 飞盘！
PE 飞盘！
HT 飞盘！

ST Cycling.
RAW 循环。
PE 骑行。
HT 骑行。

ST Judo.
RAW 柔道。
PE 柔道。
HT 柔道。

ST Basketball.
RAW 篮球。
PE 篮球。
HT 篮球。

ST Surfing!
RAW 冲浪！
PE 冲浪！
HT 冲浪！

ST Running!
RAW 跑！
PE 跑步！
HT 跑步！

ST Hill walking.
RAW 山地行。
PE 徒步。
HT 爬山。

ST Skipping.
RAW 跳绳。
PE 跳绳。
HT 跳绳。

ST Highland dancing.
RAW 高地跳舞。
PE 高地舞。
HT 苏格兰高地舞。

ST But what is physical activity?
RAW 但什么是身体活动？
PE 但什么是身体活动？
HT 但是，身体活动是什么？

ST Physical activity is any movement that uses energy.
RAW 身体活动是使用能量的任何运动。
PE 身体活动是消耗能量的任何运动。
HT 身体活动是任何使用能量的活动。

ST And being physically active is important for people of all ages.
RAW 身体活跃对所有年龄段的人都很重要。
PE 保持身体活跃对所有年龄段的人都很重要。
HT 保持身体的积极状态对所有年龄的人都重要。

ST Physical activity can help reduce our risk for multiple diseases such as coronary heart disease, type 2 diabetes, some types of cancer.
RAW 身体活动可以帮助降低多种疾病的风险，如冠心病，2 型糖尿病，某些类型的癌症。
PE 身体活动可以帮助我们降低多种疾病的风险，如冠心病，2 型糖尿病，某些类型的癌症。
HT 身体活动能帮助减少患多种疾病的风险，比如冠心病，2 型糖尿病，某几种癌症。

ST It can also improve our bone health, help us to maintain healthy weight, and also improve our overall quality of life.
RAW 它也可以改善我们的骨骼健康，帮助我们保持健康的体重，并提高我们的整体生活质量。
PE 它也可以改善我们的骨骼健康，帮助我们保持健康的体重，并提高我们的整体生活质量。
HT 它还能改善我们的骨质健康，帮助我们维持健康的体重，并且还能提升我们整个生活质量。

ST Physical activity is any movement that we do with our bodies.
RAW 身体活动是我们对身体的任何运动。
PE 身体活动是我们通过身体进行的任何运动。
HT 身体活动就是我们身体所做的任何动作。

ST It can be quite light movement, like just walking around a little bit in your office or at home.
RAW 它可以是相当轻的运动，就像在你的办公室或在家里走一点点。
PE 它可以是相当轻的运动，就像在你的办公室或在家里走一点点路。
HT 它可能是相当轻松的活动，比如就在你的办公室或家里走几步。

ST It can be moderate activity, walking the dog, going along the pavement.
RAW 它可以适度的活动，走狗，沿着人行道走。
PE 它可以是适度的活动，遛狗，沿着人行道走。
HT 它有可能是遛狗以及沿路走走之类温和的活动。

ST It can be quite vigorous activity, running or playing sports.
RAW 它可以是相当活跃的运动，跑步或打运动。
PE 它可以是相当强烈的运动，跑步或做运动。
HT 它可能是跑步或者参加体育比赛这些相当充满活力的活动。

ST Any movement of the body is good for our health.
RAW 身体的任何运动都有益于我们的健康。
PE 身体的任何运动都有益于我们的健康。
HT 任何身体活动都有益我们的健康。

ST Our bodies are built to move.
RAW 我们的身体被建造来移动。
PE 我们的身体生来就是为了活动。
HT 我们的身体生来就是要运动。

ST We often hear of physical activity and exercise being used interchangeably.
RAW 我们经常听到身体活动和运动互换使用。
PE 我们经常听到身体活动和锻炼互换使用。
HT 我们经常听说身体活动和锻炼这两个词使用的时候是可以互换的。

ST But do they really mean the same thing?
RAW 但他们真的意思同样吗？
PE 但它们的意思真的一样吗？
HT 但是它们真的一样么？

ST I don't think it's now correct to use them interchangeably.
RAW 我认为现在使用它们是不正确的。
PE 我认为现在交换使用它们是不正确的。

HT 我不觉得现在交换使用它们是正确的

ST Physical activity is the broad umbrella term that covers all bodily movements.
RAW 身体活动是涵盖所有身体运动的广泛术语。
PE 身体活动是涵盖所有身体运动的广泛术语。
HT 身体活动是广义的所有身体活动的统称。

ST But exercise is a very specific subset of physical activity.
RAW 但锻炼是身体活动的一个非常具体的子集。
PE 但锻炼是身体活动的一个非常特别的子集。
HT 但是锻炼是非常特定的某些身体活动。

ST Exercise is pursued for health benefits, for fitness benefits.
RAW 追求健康的健康，健康的益处。
PE 锻炼是为了追求健康，为了健身。
HT 锻炼目的是为了健康或者身材。

ST For example, we might start running to take part in a 10K run.
RAW 例如，我们可能开始运行，参与 10K 的运行。
PE 例如，我们可能为了参加十公里跑步比赛而开始跑步。
HT 例如，为了参加 10 公里长跑我们可能开始跑步。

ST This would improve our aerobic fitness.
RAW 这将提高我们的有氧健身。
PE 这有助于我们的有氧健身。
HT 这将改善我们的心肺功能。

ST We might go to the gym to do strength training, and this will improve our strength fitness.
RAW 我们可以去健身房进行力量训练，这将提高我们的实力。
PE 我们可以去健身房进行力量训练，这将提高我们的体能。
HT 我们可能去健身馆做力量训练，这将会增强我们的力量。

ST We might go to a yoga or Pilates class to improve our balance and coordination, part of our fitness.
RAW 我们可以去瑜伽或普拉提课程来改善我们的平衡和协调，这是我们健身的一部分。
PE 我们可以去瑜伽或普拉提课程来改善我们的平衡和协调，这是我们健身的一部分。
HT 我们可能去上瑜伽或普拉提课程来改善我们的平衡感和协调性，这也是我们健康的一部分。

ST And these are all exercise-related activities, because they're structured and we're largely pursuing them for fitness and health benefit.
RAW 这些都是运动相关的活动，因为它们是结构化的，我们在很大程度上追求健身和健康的好处。
PE 这些都是和锻炼相关的活动，因为它们是结构化的，我们锻炼在很大程度上是为了追求健身和健康的好处。

HT 以上这些都是与锻炼有关的活动，因为他们是有结构并且主要追求健康和完美的体型。

ST Now we know what physical activity is, and now we know the difference between physical activity and exercise.
RAW 现在我们知道身体活动是什么，现在我们知道身体活动和运动之间的区别。
PE 现在我们知道身体活动是什么，现在我们知道身体活动和锻炼之间的区别。
HT 现在我们知道什么是身体活动并且了解身体活动和锻炼之间的区别了。

ST Let's see now how much physical activity we should be doing to gain health benefits.
RAW 现在让我们看看我们应该做多少身体活动来获得健康的好处。
PE 现在来看看我们应该做多少身体活动才能获得健康的好处。
HT 现在让我们看看为了获得健康我们应当做多少身体活动。

ST So let's go to check the physical activity guidelines.
RAW 所以我们去检查体育活动指南。
PE 所以我们去看看身体活动指南。
HT 那么让我们去看看身体活动准则。

ST So what we want is that for children aged 5 to 18, that they do 60 minutes of moderate to vigorous activity every day of the week.
RAW 所以我们想要的是，对于 5 岁到 18 岁的孩子，他们每周每天都做 60 分钟的中度到活跃的活动。
PE 所以我们想要的是，对于 5 岁到 18 岁的孩子，他们每周每天都做 60 分钟的中度到强烈的活动。
HT 对于 5 到 18 岁的孩子来说，我们希望他们每天做 60 分钟的从温和到剧烈的活动。

ST That's running, jumping, all the usual play stuff.
RAW 这是运行，跳跃，所有常见的播放的东西。
PE 就是跑步，跳跃，所有常见的活动。
HT 也就是跑步，跳跃以及所有寻常的体育活动。

ST What we'd also like them to do is some vigorous activity for strengthening their muscles on at least 3 days a week to get the maximum health benefits.
RAW 我们还希望他们做一些有力的活动，每周至少 3 天加强肌肉，以获得最大的健康益处。
PE 我们还希望他们做一些强烈的活动来锻炼肌肉，每周至少 3 天，以获得最大的健康益处。
HT 为了最大限度的获得健康，我们还想要他们一周至少有三天做某种剧烈的活动来增强他们的肌肉力量。

ST We would like adults to have moderate activity of at least 30 minutes, 5 days a week.
RAW 我们希望大人有一个至少 30 分钟，每周 5 天的适度活动。
PE 我们希望大人有一个至少 30 分钟，每周 5 天的适度活动。
HT 我们想要成年人一周有 5 天做至少 30 分钟不剧烈的活动。

ST So that's 150 minutes every week.

RAW 所以这是每周 150 分钟。

PE 所以这是每周 150 分钟。

HT 也就是每周 150 分钟。

ST We would also like them to have strength building activities and to minimize the time they spend sitting.

RAW 我们也希望他们进行实力建设活动，尽量减少他们坐下的时间。

PE 我们也希望他们进行力量训练，尽量减少他们坐着的时间。

HT 我们还想要他们进行力量训练并且让他们坐着不动的时间减到最少。

ST But if you want to do 10 minutes a day, 5 minutes a day, that's better than nothing.

RAW 但是，如果你想每天做 10 分钟，每天 5 分钟，那比一切都好。

PE 但是，如果你想每天只做 10 分钟，每天 5 分钟，那比什么都不做要好。

HT 但是如果你仅仅想要一天锻炼十分钟，或者五分钟也比什么都不做好。

ST We've just learned how much physical activity is needed to get health benefits.

RAW 我们刚刚了解到需要多少身体活动才能获得健康益处。

PE 我们刚刚了解到需要多少身体活动才能获得健康益处。

HT 我们刚刚了解了为了健康要做多少身体活动。

ST But let's see what is physical inactivity, and how it affects our health.

RAW 但是让我们来看看什么是身体不活动，以及它如何影响我们的健康。

PE 但是让我们来看看什么是身体不活动，以及它如何影响我们的健康。

HT 但是让我们看看什么是身体不活动以及它如何影响到我们的健康。

ST Physical inactivity means very low levels of physical activity.

RAW 身体不活动意味着非常低的身体活动水平。

PE 身体不活动意味着非常低的身体活动水平。

HT 身体不活动意味着低水平的身体活动。

ST For example, if we do not do 30 minutes of physical activity in a week, we can define ourselves as being physically inactive.

RAW 例如，如果我们在一周内没有做 30 分钟的身体活动，我们可以将自己定义为身体不活动。

PE 例如，如果我们在一周内没有做 30 分钟的身体活动，我们可以将自己定义为身体不活动。

HT 例如，如果我们一周身体活动不足 30 分钟，我们可以判定我们自己身体不活动。

ST Very low levels of physical activity can increase our risks of getting disease, for example, coronary heart disease, type 2 diabetes, stroke, and cancer.

RAW 非常低的身体活动水平可能增加我们的疾病风险，例如冠心病，2 型糖尿病，中风和癌症。

PE 非常低的身体活动水平可能增加我们的患病风险，例如冠心病，2 型糖尿病，中风和癌症。

HT 非常低水平的身体活动会增加我们患病的风险，例如冠心病，2 型糖尿病，中风以及癌症。

ST Being physically inactive can also affect our bone health, and in time can lead to greater risks of getting osteoporosis.
RAW 身体不活动也会影响我们的骨骼健康，及时可能会导致更大的骨质疏松症风险。
PE 身体不活动也会影响我们的骨骼健康，可能会导致更大的骨质疏松症风险。
HT 身体不活动也会影响我们的骨头健康并且增加得骨质疏松症得风险。

ST But a bit of physical activity is better than none at all.
RAW 但是，有一点体力活动比没有一个更好。
PE 但是，有一点身体活动比什么都没有好。
HT 但是少量的身体活动比什么都不做要好。

ST How active is the world, or how inactive is the world?
RAW 世界有多活跃，还是世界不活跃？
PE 世界有多活跃，世界有多不活跃？
HT 这个世界有多活跃，或者这个世界有多不活跃？

ST Evidence shows that worldwide, 31% of adults are physically inactive.
RAW 有证据显示，全球 31%的成年人身体不活跃。
PE 有证据显示，全球 31%的成年人身体不活跃。
HT 证据表明全世界 31%的成年人身体不活跃。

ST Well, a third of the people in the world are not sufficiently active for good health.
RAW 那么世界上有三分之一的人没有足够的活跃，身体健康。
PE 那么世界上有三分之一的人没有为了健康而足够的活跃身体 。
HT 也就是说世界上三分之一的人对于良好的健康状况不是十分活跃。

ST But not all groups are equally inactive.
RAW 但并不是所有的群体都是同样的活跃。
PE 但并不是所有的群体都一样的不活跃
HT 但是不是所有的人群都一样不活跃。

ST People from high-income nations tend to be less active than those from developing countries.
RAW 来自高收入国家的人们往往不如发展中国家那样活跃。
PE 来自高收入国家的人们往往不如发展中国家那样活跃。
HT 与来自发展中国家的人相比，来自高收入国家的人倾向于不是那么的活跃。

ST Men tend to be more active than women.
RAW 男性往往比女性更加活跃。
PE 男性往往比女性更加活跃。
HT 男性比女性更活跃。

ST And we tend to get less active as we get older.

RAW 而且，随着年龄的增长，我们往往不那么活跃。

PE 而且，随着年龄的增长，我们往往越不活跃。

HT 当我们年龄更大点时，我们倾向于变得不那么活跃。

ST Patterns that are similar in young people, boys tend to be more active than girls, and children get less active as they age.

RAW 年轻人的模式类似，男孩往往比女孩更活跃，孩子年龄越来越小。

PE 年轻人的模式类似，男孩往往比女孩更活跃，孩子越长大越不活跃。

HT 在年轻人当中，这种模式是相似的，男孩比女孩更活跃，孩子们会随着年龄增长变得不那么活跃。

ST That's particularly the case in adolescents.

RAW 青少年尤其如此。

PE 青少年尤其如此。

HT 在青少年中，这种情况更是如此。

ST As few as 20% of 13 to 15-year-old young people are not sufficiently active for good health.

RAW 13 至 15 岁的青少年中只有 20%没有足够的身体健康。

PE 13 至 15 岁的青少年中只有 20%的人缺乏对健康有益的活跃度。

HT 20%的 13 到 15 岁的年轻人对于健康不是非常活跃。

ST Now we know what physical activity is and what physical inactivity is.

RAW 现在我们知道身体活动是什么，身体不活动是什么。

PE 现在我们知道身体活动是什么，身体不活动是什么。

HT 现在我们知道什么是身体活动以及什么是身体不活动。

ST But what does it mean to be sedentary?

RAW 但是久坐的意思是什么？

PE 但是久坐的意思是什么？

HT 但是久坐是什么意思？

ST Being sedentary is when we are sitting or lying down across the day.

RAW 坐着的时候，当我们坐下或躺在这一天。

PE 久坐是指我们整天都坐着或躺着。

HT 久坐指我们整天都坐着或者躺着。

ST So for example, if you have a desk-based job, then it's likely that you spend quite a lot of your day sitting or sedentary.

RAW 所以例如，如果你有一个办公桌的工作，那么很可能你花了很多的时间坐在一起坐下来。

PE 所以例如，如果你是伏案工作，那么很可能你一天里有很多的时间都坐着或久坐不动。

HT 例如，如果你拥有一份案头工作，那么你极有可能一天大量时间坐着或者不怎么活动。

ST And other common examples of sedentary time in everyday life are sitting on the sofa to watch the television and sitting on motorized transport vehicles.

RAW 日常生活中久坐久久的其他常见例子就是坐在沙发上看电视，坐在机动车上。

PE 日常生活中久坐的其他常见例子就是坐在沙发上看电视，坐在机动车上。

HT 在日常生活中，坐在沙发上看电视，在机动车上坐着

ST So for example, driving in your car, taking the train or a bus, or sitting at home on a personal computer or on your tablet.

RAW 所以例如驾驶你的车，乘坐火车或公共汽车，或坐在家里的个人电脑或平板电脑上。

PE 所以例如驾驶你的车，乘坐火车或公共汽车，或坐在家里玩电脑或平板。

HT 例如开车，乘坐火车或巴士或者坐在家里操作电脑或平板都是很普遍的久坐的例子。

ST And we know that people who have higher levels of sitting across their everyday lives are putting their heath at increased risk.

RAW 而且我们知道，在日常生活中拥有较高水平的人会使他们的健康面临更大的风险。

PE 而且我们知道，在日常生活中坐着时间比较久的人们正在增加他们的健康风险。

HT 我们知道那些日常生活中经常坐着的人正在增加他们的健康风险。

ST And this seems to be even in individuals who are achieving physical activity recommendation.

RAW 这似乎甚至在实现体育活动推荐的个人中。

PE 甚至那些实现身体活动的人也好像有这样的情况。

HT 并且甚至那些实现身体活动的人似乎也有这个情况

ST So, and we know that higher levels of sitting are linked with obesity and an increased likelihood of developing type 2 diabetes and cardiovascular disease and some cancers, and even premature mortality.

RAW 所以，我们知道更高层次的坐姿与肥胖有关，增加发生 2 型糖尿病，心血管疾病和某些癌症，甚至过早死亡的可能性。

PE 所以，我们知道坐久了会产生肥胖，增加发生 2 型糖尿病、心血管疾病和某些癌症，甚至过早死亡的可能性。

HT 我们知道经常坐着与肥胖症，2 型糖尿病恶化的可能性，心血管疾病，某些癌症甚至过早死亡有联系。

ST So we need to think of ways to sit less and be less sedentary in our everyday lives.

RAW 所以我们需要考虑在日常生活中少坐少坐的方法。

PE 所以我们需要考虑在日常生活中少坐的方法。

HT 因此，我们需要想出在日常生活中少坐以及多活动的办法来。

ST And some of the ways that we can do this are to think about using alternative forms of transport.

RAW 我们可以做的一些方法是考虑使用替代形式的运输。

PE 我们可以做的一些方法是考虑使用替代交通工具的方式。

HT 其中的一些办法就是我们可以考虑使用多样的交通方式。

ST So can we use active transport like walking or cycling more often instead of relying on motorized transport all the time?

RAW 所以我们可以使用积极的运输，如步行或骑自行车，而不是一直依靠机动交通？

PE	所以我们可以使用活跃的交通方式，如更多的去步行或骑自行车，而不是一直依靠机动交通？
HT	我们可以更加经常的步行或骑自行车而不是一直依赖机动的交通工具么？
ST	Can we break up longer periods of sitting in the workplace and get up and move about more?
RAW	我们可以分开更长一段时间坐在工作场所，起床和走动更多吗？
PE	我们可不可以不要总是坐在工作场所，多起来动一动？
HT	我们可以不要长时间的坐在工作间里，经常站起来并且四处走走么？
ST	Or could we consider spending less time sitting watching the television and try to move about more?
RAW	或者我们可以考虑花更少的时间坐在看电视，并尝试更多的移动吗？
PE	或者我们可不可以考虑花更少的时间坐着看电视，而是更多的动一动？
HT	或者我们可以考虑花更少的时间坐着看电视并且设法多四处走走么？
ST	Hope this video convinced you that regular physical activity is important for our health.
RAW	希望这个视频说服你定期体力活动对我们的健康很重要。
PE	希望这个视频说服你定期身体活动对我们的健康很重要。
HT	希望这段视频能说服你认为有规律的身体活动对我们的健康很重要。
ST	And one of the best ways to be more active is to find activities that you enjoy doing, and make them part of your daily life.
RAW	而更有活力的最佳方法之一就是找到你喜欢做的活动，并将其作为日常生活的一部分。
PE	让自己更有活力的最佳方法之一就是找到你喜欢做的活动，并将其作为日常生活的一部分。
HT	最佳的更加积极的办法之一就是找到你喜欢做的活动并且让它们成为你日常生活的一部分。

“I dream it. I work hard. I grind till I own it.”
Formation – Beyoncé