國立臺灣大學文學院翻譯碩士學位學程
碩士論文

Graduate Program in Translation and Interpretation
College of Liberal Arts
National Taiwan University
Master thesis

英中同步口譯之句末時間差與譯文精確性之關聯

A Research on the Relationship between Tail-to-tail span and Quality in English to Chinese Simultaneous Interpreting

陳品秀
Ping-hsiu (Kimberly) Chen

指導教授：范家銘 博士
Dr. Chiaming (Damien) Fan

中華民國 107 年 7 月
July 2018

doi:10.6342/NTU202000181

摘要

同步口譯中之「同步性」(synchronicity) 係指原文與譯文發生之時間十分接近，近乎完全同步。但同步性可能因口譯員或情境而異，口譯員可能習慣於緊跟講者或刻意拉長距離，但也可能於特定情境下另做選擇。此一差異長久以來為口譯研究中的熱門主題，談討同步性或其與品質、語言方向、口譯員資歷等面向之關聯。同步性的具體探討乃透過原文與譯文間的「時間差」(time lag) 切入。過往研究多採取聽譯時間差 (ear-voice span) 為衡量時間差的方式，即量測原文開端與譯文開端間的時間差。2003 年時，韓國學者 Tae-hyung Lee 提出另一種衡量時間差的方式：句末時間差 (tail-to-tail span) 。句末時間差以原文結尾與譯文結尾為時間差的量測點，乍看與聽譯時間差似乎十分雷同，然而兩者間卻有儘管細微但仍值得細究之差異， Lee 認為應在研究上予以區別和重視。兩種時間差最關鍵之區別為，口譯員對於聽譯時間差較有自行掌控之可能，而句末時間差則相對受限。其次，譯者對於聽譯時間差長短可能各有偏好，或偏好較短的時間差以求降低記憶負擔，或選擇較長時間差以求理解更完整；相較之下，少有口譯員會選擇拉長句末時間差，因為句末時間差過長可能會犧牲原可用於理解和處理下一句原文的心力，在無法預知下一句原文難易輕重時，維持較短的句末時間差是較佳之策。本研究承襲 Lee 對句末時間差的研究，探討句末時間差與口譯品質間的關係，分析兩百餘組句子中句末時間差與下一句之口譯品質間的關係。本研究的核心假設為：某句句末時間差長短與下一句之口譯品質呈現反向關係，即某句譯文之句末時間差愈長，下一句之品質愈差；句末時間差愈短，下一句之品質則愈好。研究結果發現，整體而言，句末時間差短的譯文平均獲得較高分的口譯品質分數，且平均句末時間差較短的受試者獲得之平均口譯分數也較高。在部分受試者身上也發現同一位受試者不同句的口譯品質也會受到句末時間差長短的影響。

關鍵字：同步口譯、句末時間差、聽譯時間差、口譯品質

doi:10.6342/NTU202000181

Abstract

Synchronicity is a key factor in simultaneous interpreting (SI). While it is a basic element in SI, the level of synchronicity still varies by interpreting practitioner and by context sometimes. Synchronicity has long been the topic of much interest in the field of interpreting research. Past studies have looked at synchronicity itself and its relationship with such aspects as quality, directionality and experience. Most studies measure synchronicity through “ear-voice span”, the time that elapses between the onset of the source speech and the onset of the corresponding target speech. In 2003, researcher Tae-hyung Lee proposed a new parameter for measuring synchronicity – tail-to-tail span, the time that elapses between the end of a source speech sentence and the end of the corresponding target speech sentence. Though tail-to-tail span appears to be highly similar with ear-voice span, crucial differences exist. The main difference involves the manipulation of span length. While ear-voice span can be extended or shortened through adjusting the length of interpreter's pause between sentences and for such purposes as better comprehension or lighter memory load, shortening tail-to-tail span requires more effort, and there is little rationale for extending one's tail-to-tail span, as processing of the upcoming sentence could be at stake. This research follows Lee's footstep in exploring the relationship between tail-to-tail span and interpreting quality, and conducts an experiment to examine the relationship between a sentence's tail-to-tail span and the next sentence's interpreting quality. The research hypothesizes that tail-to-tail span length and interpreting quality are negatively correlated. Results based on over 200 sentences show that the hypothesis holds true not just for the experiment data as a whole, but also for inter-

doi:10.6342/NTU202000181

participant comparison (i.e. participants with shorter TTS tend to have better quality), and in certain cases, for intra-participant comparison (i.e. the same participant produces renditions of better quality when TTS is shorter).

Keywords: simultaneous interpreting, tail-to-tail span, ear-voice span, quality

doi:10.6342/NTU202000181

Acknowledgment

I am truly blessed to have so many encouraging and supportive people to lend generous support and accompany me to the end of this years-long journey.

I would like to especially thank my advisor, Professor Damien Fan, for continuously giving me guidance and support under all kinds of circumstances. I would also like to thank Professor Michelle Wu and Emma Ju for their valuable suggestions. My deep gratitude also goes to all my fellow classmates and peers from NTU, NTNU and FJU who have participated in the experiment.

I would also like to thank the GPTI office and Vicky for being the best and kindest facilitator. And the R04 interpreting class, thank you for being the strongest support.

Finally, I would like to thank my family and my boyfriend, who have been there for me the longest time.

doi:10.6342/NTU202000181

Table of Contents

doi:10.6342/NTU202000181

doi:10.6342/NTU202000181

List of figures

doi:10.6342/NTU202000181

List of Tables

doi:10.6342/NTU202000181

doi:10.6342/NTU202000181

I. Introduction

Simultaneous interpreting is a mode of interpreting where the interpreter utters the interpretation at the same time as the speaker utters his/her speech. The simultaneous nature of simultaneous interpreting is what makes simultaneous interpreting highly useful in today's world of time scarcity. With the aid of technology, interpreting is no longer limited to the time-consuming consecutive mode, where the speaker and interpreter take turns to speak and thus consumes more time. However, the synchronous nature of simultaneous interpreting also makes it a highly challenging task. Listening to one stream of speech while uttering another stream of speech is not an ability that comes without hard practice. Simultaneous interpreting remains challenging even after interpreters have acquired and familiarized themselves with the skill. When interpreting simultaneously, interpreters are under constant struggle to pay sufficient attention to comprehend the speaker's speech, reformulate the speaker's messages into the target language, and strive to prevent losing track of messages temporarily stored in the interpreter's memory to be translated into the target language. At times, certain challenges in simultaneous interpreting could slow interpreters down. For example, they may encounter a term that does not have an equivalent in the target language, so they need to search for a near-equivalent instead; or the speaker's speech may contain cultural specific items, which require interpreters to elaborate the cultural connotations to the listeners. Sometimes a stumble caused by the challenge snowballs into a long silent pause or an error in comprehension or speech production. These situations along with many other challenging situations may cause interpreters to fall further behind the speaker. Slowing down and falling behind affect the "synchronicity" of simultaneus interpreting, which refers to the state of source speech and target speech being uttered and unfolded at the same time.

doi:10.6342/NTU202000181

If interpreters fall behind and do not manage to catch up as soon as possible, there could be negative consequences for both the listeners and interpreters themselves.

Synchronicity is a very important factor of simultaneous interpreting. Simultaneous interpreting is a highly goal-oriented task, which means that achieving its goal – facilitating cross-cultural communication to the extent possible – is quintessential. Simultaneous interpreting timely delivered to the ears of the listeners facilitates communication. It helps listeners coordinate the information they hear with the other real-time events around them such as laughter provoked by the speaker's joke, information shown on the speaker's slides, questions that the speaker throws out to the audience, or an opportunity for the listeners to voice their opinion. When the information arrives in the ears of the listeners too late, the listeners miss the optimal chance to engage in the discourse.

For interpreters, keeping up with the speaker has its inherent benefits. Interpreters need to retain in memory the information that is already heard but are waiting to be vocalized in the target language. Therefore, the further interpreters lag behind, the more items they would need to retain in memory. It is generally believed that human beings have a limited memory span, and the more items one needs to keep in mind, the greater the chance that some items would be lost. In addition, having to comprehend the speaker and translate the message while keeping many things in memory is a very challenging task. Though multi-tasking is inevitable for performing simultaneous interpreting, if interpreters employ strategies to effectively cope with or relieve the heavy memory load, it would most certainly make the task less strenuous and could also allow interpreters to pay more attention to comprehending the speaker's message and reformulating it into the target language.

doi:10.6342/NTU202000181

On the other hand, while keeping up with the speaker benefits both the listeners and interpreters, keeping too close to the speaker is not always necessarily the better alternative. In the majority of cases, interpreters do not really interpret "simultaneously" with the speaker. The interpretation follows the speaker's speech at a small but necessary distance behind. This distance allows interpreters, who rarely can predict exactly what the speaker is going to say, to grasp a minimal sense of the speaker's message which they can use to start their sentence in the target language. If interpreters follow too closely behind the speaker, they may risk a "false start", where the way the sentence is started makes following the speaker's message and logic challenging. Interpreters may need to correct themselves or patch up missing or unfitting information afterwards, which would often require them to spend more time than they would have had the false start not occured. The minimal sense of the speaker's message interpreters have grasped when they consciously fall behind the speaker allows interpreters to determine how best to start their sentence in the target language. The distance at which interpreters fall behind the speaker is generally referred to as "décalag", or "lag". Given the importance of décalage to simultaneous interpreting, it has attracted many pedagogical and academic attention. For conference interpreting trainees learning simultaneous interpreting, the question of "when to start" is one of the challenges they encounter very early on in their training. Many studies also looked at the relationship between décalage and many other dimensions of simultaneous interpreting.

One of the commonly studied topics is the "ear-voice-span" (EVS), the duration of time that elapses between the onset of a speaker's source language speech and the onset of the interpreter's target language production. Interpreters' EVS length reveals how much information they have received before uttering their translation, and also provides a hint to the load of information interpreters may be retaining in their memory. As the interpreted discourse unfolds, interpreters manage to overcome different challenges on the way, and their EVS would thus vary from sentence to sentence, or even from clause to clause or word to word.

Despite all the attention on EVS, its opposite - how far interpreters lag behind the speaker at the end of a sentence – has been much neglected. Simultaneous interpreting is a very dynamic activity. The lag interpreters start with varies as they run into different situations throughout the sentence, and thus end with a shorter or longer lag. Although looking at the EVS of the next sentence could provide hints to the lag at the end of the sentence, it is inevitably distorted. The EVS of the next sentence embodies the speaker's pause length and the interpreters' pause length between the two sentences, because pausing between sentences is a natural act in language production. While it may be meaningful to incorporate the speaker and interpreters' pause lengths into interpreting studies, the study of the speaker's pause length puts interpreters in a passive position. Rarely are interpreters able to dictate the speaker's pause length. On the other hand, interpreters can play a relatively more active role in adjusting the actual lag time at the end of the sentence and interpreters' pause length between two sentences by adopting various strategies to control the length of their lags and pauses.

The lag time at the end of sentences in simultaneous interpreting was finally picked up by English-Korean conference interpreting researcher Tae-hyung Lee in 2003, when he proposed a new parameter for measuring the lag at the end of sentence in a study. The parameter was named tail-to-tail span (TTS). Lee (2003) defines TTS as "the time lag between the tail of a speaker's sentence and the tail of an interpreter's translated sentence."

doi:10.6342/NTU202000181

The results of his 2003 research revealed that TTS does exert an influence on the output quality of simultaneous interpreting. However, since then TTS has remained rather unexplored. The majority of simultaneous interpreting lag time research focused on EVS. But there are good reasons to establish TTS as a sub-branch of lag time study separated from the studies of EVS on top of the reasons already discussed above. While some of the variables, such as language combination, information density, discourse type (Yagi, 2000), that the length of EVS is subject to also apply to TTS, one crucial difference is idiosyncratic preferences. Empirical evidence and experiential anecdotes both described that interpreters consciously choose to finish producing one's target speech as quickly as possible, hence shortening TTS on purpose, so as to allow for better comprehension for the coming segments (Yagi, 2000; Chang & Schallert, 2007). However, some interpreters prefer to lengthen EVS for better information preservation, whereas others opt for the opposite (Kraviarová, 2013). Given the consistent preference for reducing the length of TTS, it is worth further exploring what the effects of shorter TTS on SI quality are.

doi:10.6342/NTU202000181

II. Literature Review

2.1 Simultaneous interpretation and synchronicity

The term simultaneous interpretation (SI) is a commonly used term to refer to a mode of interpreting service. As the name of the mode itself suggests, simultaneous interpretation is of a simultaneous nature in that, as the speaker speaks, the interpreter interprets at the same time. This is but a very rough definition of simultaneous interpretation. Since the documented debut of SI in the 1920s, practitioners of the interpreting profession as well as researchers from the fields of conference interpreting and psychology have come up with many ways to define simultaneous interpretation (Seeber, 2015). The global association of professional conference interpreters, the International Association of Conference Interpreters (AIIC), for example, defines SI as a mode of interpreting that includes the interpreter "sit[ting] in a booth" and "immediately speaks their interpretation" upon hearing the speaker's message through the headphones (AIIC, 2011). AIIC's definition includes the simultaneity of SI, specifying the immediacy of the target speech production. In addition, it also specifies the working environment and apparatus, painting a clear picture of interpreters sitting in a booth, listening to the speaker through headphones and re-speaking into the microphone. Yet, there are other modes of interpreting that requires the interpreter to listen to the speaker and explicitly voice the translation immediately, but does not require the interpreter to sit in a booth or use equipment. Chuchotage, or whispering, for example, does not require the interpreter to be in a booth, but the interpreter still has to produce the rendition while listening to the speaker at the same time. For the sake of this research, the definition that draws a line between SI and the other major interpreting mode - consecutive interpreting (CI) - will be used. SI is defined as: a mode of interpreting where "the interpreter produces his/her speech while the interpreted

doi:10.6342/NTU202000181

speaker is speaking/signing – though with a lag of up to a few seconds" (Gile, 2018). The other mode consecutive interpreting (CI) is a mode where the speaker pauses deliberately for the interpreter to vocally produce the translation, after which the speaker will resume speaking.

An obvious difference between the two modes is the temporal relationship between the source speech and the target speech, as illustrated in Figure 1 and Figure 2 below.

Figure 1 Relationship between speaker's speech and interpreter's translation in CI

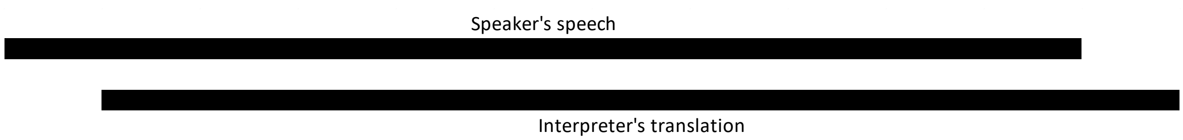


Figure 2 Relationship between speaker's speech and interpreter's translation in SI

The diagram clearly shows the distinction between SI & CI. In consecutive mode, the speaker's speech and the interpreter's vocalized rendition of the speech do not overlap one another, where as in simultaneous mode, the source speech and target speech largely overlap each other. It is worth adding that in CI, part of the translation process usually begins already while the speaker is still speaking. This is evident through observing interpreter's notes, where symbols, substitutes, paraphrases in both the source language and target language can

doi:10.6342/NTU202000181

be found, suggesting that the translation process – source text comprehension, conversion and target text production – takes place even before the notes are actually jotted down, only that the entire translation process is completed at a later stage when the speaker pauses for the interpreter to vocalize his/her translation. For SI, on the other hand, the entire translation process, from listening and understanding to vocalizing the target speech, occurs in synchrony with the speaker's source speech.

Synchronicity is a highly important feature of simultaneous interpretation. However, as shown in Figure 2 and also indicated in Gile's definition of SI, SI is not perfectly simultaneous. There is in fact a short lag between the speaker's source speech and the interpreter's target speech. Paradoxically, this lag is crucial to the synchronicity of SI, for a number of reasons. From a realistic perspective, a slight delay at the onset of SI can almost certainly be anticipated, because it is impossible that the interpreter would know beforehand how the speaker would start exactly and commence the target speech interpreting at the same time with the speaker (Davidson, 1992). The interpreter needs to first listen for a very short while to understand a minimal portion of the source speech before they can start producing the target speech (Grübl, 2013). In addition, after interpreters have grasped a minimally workable sense of the source speech, they may still require some time to reformulate the sense into the target speech.

Since the existence of lag time is universal in natural SI settings, the next questions is whether this lag would gradually impair the synchronicity of SI. In fact, while a short delay at the start is necessary, most interpreters will attempt to carry on with the interpreting process at a balanced and steady rate (Jones, 2014). In doing so, interpreters maintain a steady and

doi:10.6342/NTU202000181

managable lag time. An appropriate length of lag serves multiple purposes. For interpretation service users, they expect continuous and synchronous interpreting output (Jones, 1998). A study surveying speakers and listeners of events that use conference interpretation service revealed that synchronicity is amongst the main expectations of service users (Kurz, 2001). When interpreters fall behind the speaker too much, audience may feel irritated and start to doubt if the interpretation is having problems (Hamon et al., 2009). For interpreters themselves, managing the distance at which they fall behind the speaker is also a quintessential task. In SI, interpreters are constantly faced with the difficult choice between listening to more and not forgetting too much. When interpreters increase their lag, they can comprehend the source speech better. But they also take on risk for doing so. For instance, they would have to keep large chunks of information in mind as a result. Interpreters would have to strike a balance between comprehending the source speech, formulating the target speech, and retaining the yet-to-be-vocalized segments in mind in order to pull off the challenging task of SI. With so many sub-tasks to execute and so much information to process in compressed time, interpreters' working memory then plays a crucial role.

2.2 Time lag and working memory

Working memory (WM) is loosely defined as the cognitive ability to hold information in mind and work with the information mentally (Diamond, 2013). While there are many models attempting to account for the detailed mechanisms of working memory as well as models for how SI works, it is generally agreed that SI relies heavily on the interpreter's WM, because SI involves holding information in mind and working on it within a very short time. Therefore, the role of WM is much emphasized in SI research.

doi:10.6342/NTU202000181

WM is central to cognitive tasks such as language comprehension, analytical reasoning, planning and memory storage, all of which are crucial to performing SI (Cowan, 2000/01). A general agreement on the definition and components of WM is still lacking in the research field, but prominent theories include that of Baddeley's multi-component model (2000) and Cowan's activated long-term memory model (2000/01). In Baddeley's multi-component model, WM consists of four components - two storage subsystems, an episodic buffer and a central executive. The two storage subsystems are the phonological loop and the visual-spatial sketchpad, holding phonological and visual information respectively. The episodic buffer activates pieces of information in long-term memory that are relevant to the incoming new information. The final component, the central executive, is in charge of controlling all the mental processes and 'slave systems'. Cowan's activated long-term memory model, on the other hand, posits that working memory is the subset of items in long-term memory that is activated. In Cowan's model, items in long-term memory can be tiered into the inactivated subset not available for cognitive processing unless they receive enough stimulation, the activated subset that is not under the focus of attention but available for cognitive processing, and the subset that is being processed, or in other words, under the focus of attention (Cowan, 2000/01). Cowan also proposed two types of limitations to working memory. The first is a capacity limitation of the focus of attention, which has the capacity of holding around 4 items a time. The second is a time limitation of activation. Each activation is estimated to last for 10 to 30 seconds, after which activated items gradually fade out of working memory. An inactive item can be re-activated by directing the focus of attention to it. Both systems of WM theories have been applied to SI research.

doi:10.6342/NTU202000181

To better examine how WM and SI are linked, SI is further broken down into several "efforts" after Gile's effort model. Gile's effort model breaks down the complexity of SI into four efforts, with each effort being a cognitive task demanding mental resource. The four efforts are: (1) the listening and analysis effort that makes sense of the incoming sound wave of the source speech and decodes the message it carries; (2) the memory effort that retains relevant information in short-term memory for future processing; (3) the production effort that encodes the information retained in short-term memory into the target language; and (4) the coordination effort that controls each of the efforts and allocates mental resource to each of the efforts (Gile, 1995).

Using either Cowan's or Baddeley's model would reveal that SI is highly taxing for WM. By Baddeley's multi-component model, interpreters' are heavily burdened when they perform SI. One of the burdens arise from the necessity for WM to process two streams of speech at the same time – to comprehend the incoming source speech and utter the target speech. Baddeley has identified an effect on WM that is present when one attempts to process two streams of speech a time, or more specifically, to vocally say something while attempting to memorize something else. This effect is known as the articulatory suppression effect. According to Baddeley, when one is memorizing a set of items, one undergoes a subvocal rehearsal process, where one silently pronounces the words to refresh their memory traces of the items to be remembered. However, when one is forced to articulate an irrelevant stream of speech, the subvocal rehearsal process is disabled, and the memory traces of the items to be remembered cannot be refreshed and face a greater chance of being forgotten. In a more general sense, articulatory suppression refers to the impairment to auditorily-received information caused by irrelevant vocal language production. Articulatory suppression is

doi:10.6342/NTU202000181

exactly what happens during SI – interpreters are listening and decoding one stream of language while they articulate another stream of language at the same time. Past research has found that articulatory suppression has detrimental effects on short-term memory, demonstrated by poor recall performance in experimental settings (Murray, 1965). In SI, while the interpreter struggles with the articulatory suppression effect, the central executive also has to manage to retain the items queuing up in the episodic buffer. Within the framework of Cowan's activated long-term memory model, SI is also highly onerous owing to the fact that items heard and being decoded, items retained in memory and items in the process of target language encoding all compete for attention. In both models, there seem to be the dilemma of attention competition between processing and memory. In Baddeley's model, when the phonological loop processes incoming information and produces the target speech, information piles up in the episodic buffer. In Cowan's model, when attention is being given to encoding and decoding, the time limit and capacity for the activated subsets that await processing gradually run out. Therefore, an appropriate length of lag time would be a period of time that does not exceed the limits of the interpreter's memory store but allows the interpreter to give incoming information sufficient processing. The working memory models provide support from a theoretical perspective for the argument that maintaining an appropriate lag is crucial for SI.

2.2 EVS and factors influencing EVS length

Since more than half a century ago, researchers have been exploring various aspects of lag time, especially on the concept of an "appropriate lag time". Yet, researchers have yet to settle on a definitive conclusion. Studies have approached the length of lag time from various

doi:10.6342/NTU202000181

angles. A central concept to lag time research is ear-voice-span (EVS). EVS refers to the time that elapse between the onset of a source speech segment and the onset of the corresponding target speech segment. Another concept mirroring the EVS concept is tail-to-tail span (TTS), a concept proposed by Lee (2003) which describes the lag between the end of a source speech segment and the end of the corresponding target speech segment. The majority of research on lag time since more than half a century ago have focused mostly on EVS and yielded a wealth of results.

The debate over whether an appropriate or universal range of EVS remains an unsettled issue, for a number of reasons. Some studies have concluded that EVS generally falls within the range of 2 to 10 seconds (Oléron & Nanpon, 2002), or interpreters follow a "pause-delimited" rule, which posits that interpreters would generally start speaking before an input segment ends (Frieda Goldman-Eisler, 1972). However, these findings are not exactly clear indications of how long an appropriate lag truly is. In fact, lag time could vary due to a number of reasons. There are contextual factors such as language pair, or more precisely, syntactic differences between the source and target text; other contextual factors include directionality and source speech features such as the speaker's speech rate or information density of the source text (Anderson, 1994).

2.2.1 Language pair and syntactic differences

Various researchers and practitioners have pointed out the importance for interpreters to start interpreting when they have heard enough information from the speaker. This relates to the concept in SI referred to as "chunking". Chunking is a technique interpreters use to cope with the unending influx of source speech that often comes in long and complicated sentences.

doi:10.6342/NTU202000181

Interpreters would segment the long source speech into smaller but still meaningful chunks to process, and connect the chunks as appropriate. Interpreters' lag often marks such a chunk for the interpreter, signifying a point where they feel that they are confident to produce meaningful rendition of the source speech based on what has been heard so far. Goldman-Eisler (1972) observed that interpreters frequently wait until they hear the verb phrase (predicate) of the sentence to begin interpreting. Lee (2002) and Ono et al. (2008) found that the "noun + verb" combination is a frequent initial chunk for interpreters.

While a large set of languages begin sentences with a "noun + verb" structure, a number of other languages permit different syntactic orders. Korean for example, is a subject-object-verb (SOV) language. German is assumed to be an "underlying" SVO language as well as a "free-word-order language" (from Haider 1991:12, cited in Setton, 1999) which not only permits the SVO sentence structure but also allows various permutations of the subject, verb and object order, albeit with case markers aiding language users to determine a word's case regardless of its position in the sentence. Chinese, a highly context-dependent language, as opposed to the form-reliant English language, also permits both SOV and SVO structures. Korean is another SOV language that is discourse-based, allowing comparatively free word order as well (Park 1991, cited in Kim, 2005).

If the verb phrase is required for the interpreter to start interpreting, then it should come quite naturally that interpreters would encounter some challenges when interpreting from an SOV language into an SVO language. Yet studies have yielded mixed results. Kim's study compared the EVS of both directions of the Korean-Japanese language pair and the Korean-Chinese language pair (Kim, 2005). Korean and Japanese are syntactically more similar, both

doi:10.6342/NTU202000181

with a predominantly SOV structure, while Chinese, in comparison is more skewed towards an SVO structure. Surprisingly, the study did not find significant difference in EVS lengths between the Korean-Japanese language pair and the Korean-Chinese language pair. In further analyses, Kim inferred that the interpreter's adoption of the anticipation strategy could explain why the expected EVS difference was not found. On the other hand, a number of researchers have cited performing SI from an SOV language into an SVO language posed problems for interpreters (Bevilacqua, 2009; Lee, 2002). Goldman (1972) concluded specifically in her research comparing English to French, French to English and German to English SI that the verb-final structure of German caused German to English SI to have a longer EVS than English to French and French to English SI. Ono et al. (2008) also found that SI from Japanese into English had a longer time lag compared to SI from English to Japanese due to the Japanese word order.

2.2.2 Speech rate

During the EVS vacancy, two activities the interpreter engage silently in inside their minds that ultimately determine the length of the EVS are their listening and processing. Interpreters need to hear enough information and at the same time process and reformulate the information into the target language so that they can start vocalizing the rendition. Considering these two aspects, the variance in speech rate could affect EVS in two different ways. On the one hand, fast speech rate means a large amount of information for the interpreter to process in a given time, and could require longer time for the interpreter to process all of the information. Barik (1973) and Gerver (1969) both found that fast speech rate caused interpreters to lag further behind in SI. On the other hand, when speech rate is faster, information comes in faster, the interpreter is able to accumulate understanding of the

doi:10.6342/NTU202000181

source speech faster, and may thus be able to have a shorter lag. Lee's (2002) research on English-Korean SI found that when the source speech was delivered at a higher rate, the interpreters had shorter EVS.

2.2.3 Idiosyncratic preferences and other factors

The interpreter's subjective judgment or habit could also influence the length of lag (Yagi, 2000). For example, an interpreter may prefer information storage over anticipation or vice versa (Pöchhacker, 2016). Grübl (2013) cited classroom discussions amongst interpreting students at the Karl-Franzens-Universität Graz on their choice over longer or shorter EVS, where the student cohort that had above-average EVS repeatedly mentioned the importance of gaining a clear understanding of the source speech, and the below-average EVS cohort stressed their concern about the negative effects that come with overburdened memory.

In addition to language pair, speech rate and individual preferences, researchers have also observed or speculated a number of other factors that could influence the length of lag time. Jian's (2009) research looked at whether the use of gestures by the speaker caused any difference in EVS length but found no conclusive relationship. Camille & Defrancq (2018) looked at whether gender could be used as a predictor of EVS length, testing the assumption of whether women had a shorter EVS than men did. They collected more than 10,000 samples, but found that the analysis results did not support the assumption that women had a shorter EVS than men did and also found no significant correlation between gender and EVS length.

Finally, a much neglected aspect in lag time is its spill-over effect. Spill-over effect in SI refers to the phenomenon where the interpreter has exerted too much effort in one or more upstream segment, leading to the depletion of mental resources for processing the coming

doi:10.6342/NTU202000181

segments (2000). This phenomenon has also been described by Gile (2008) as "imported cognitive load". Gile proposed to analyze interpreter's cognitive load from a "local level analysis", where the cognitive load of the interpreter is analyzed at "grammatical sentence level". In other words, sentences or independent clauses that could independently express an idea are treated as a local unit and were analyzed independently. During the event of a source speech unit, the interpreter's cognitive processing includes two parts. One part is the residual information of the previous source speech unit, and the other part is information from the current source speech unit. The residual part of the previous unit is "imported" into the current unit, while part of the current unit that can't be processed completely before the current source speech unit's end is "exported" to the next unit. Time lag in interpreting works in a similar way. If an interpreter starts with a 3-second EVS, but encounters problem triggers half-way through the unit and pauses for 1 second to overcome the problem triggers, the extra time spent would spillover to the next unit. The next unit's EVS would not necessarily increase by one second, as the actual EVS would depend on other factors such as how long the speaker paused between the current unit and the next unit and the complexity of the starting segment of the next unit. However, it is highly likely that "imported" lag time from the previous unit could affect the processing of the next unit and the length of its EVS (Seeber, 2011). If the previous unit had an extremely long lag, the interpreter would have to allocate a considerable portion of his/her effort for the current unit to process the "imported" part of the previous unit, risking not hearing or forgetting what is being said in the local unit.

Imported lag exists throughout the entire SI discourse. In fact, apart from the very first unit of a discourse, the EVS of each of the units within the discourse would more or less be influenced by the previous unit's lag. Therefore, when looking at lag time, EVS and

doi:10.6342/NTU202000181

synchronicity of SI, the “import” from the previous unit is also a key consideration. However, there has not been much research into how the lag of the previous unit affects the current unit, apart from the study on TTS by Lee (2003).

2.3 Tail-to-tail span

The term “tail-to-tail span” (TTS) first appeared in 2003 (Lee, 2003). The term was coined by researcher Lee Tae-Hyung to describe “the time lag between the end of a sentence by a speaker and the beginning of the translated sentence by an interpreter”. To calculate this span, Lee uses the formula below:

$$TTS = SBP^{*} + EVS - IBP^{**}$$

*SBP (Speaker’s between-pauses): the length of pause between speaker’s sentences
**IBP (Interpreter’s between-pauses): the length of pause between interpreter’s sentences

Figure 3 below illustrates the formula for calculating tail-to-tail span:

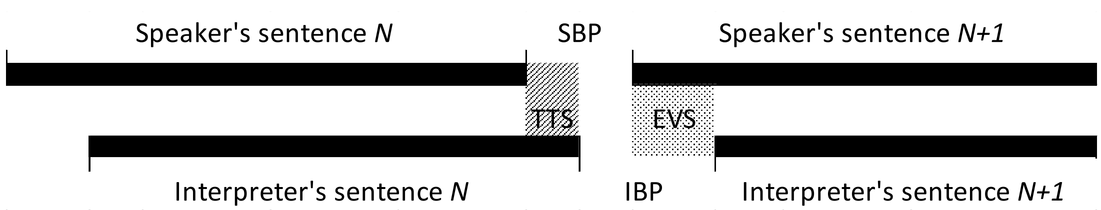


Figure 3 Formula for calculating TTS

doi:10.6342/NTU202000181

Although TTS itself may be measured by time-coding the recordings of source speech and its rendition by using sound editing or sound wave analysis software, formulizing tail-to-tail span has the benefit of taking into consideration the other elements that will affect SI quality.

One very important element that may cause SI quality to deteriorate is the length of speaker's between-sentence pauses (SBP). Many researchers have studied the effect of SBP on SI quality. Speech proportion, "the ratio that speaking time occupies within the total time", for example, is identified as one factor (Lee, 1999). According to Lee's earlier research, high speech proportion causes SI quality to suffer. In other words, when interpreting simultaneously for speakers with shorter or fewer pauses, interpreter's output quality could worsen. Shleshinger (2000) identified a spillover effect in SI that interpreters suffer from when their cognitive resources are "depleted" because of "the effort exerted upstream, in the preceding segment". In other words, if interpreters are not given time to recover from the cognitive overload, SI quality of the new segments will be affected. Gile (2008) pointed out that pauses between "full logical propositions" could reduce interpreter's cognitive load, and long enough pauses could even offset the challenges posed by informationally dense speeches. However, in actual conference settings, speakers and interpreters do not necessarily have the chance to communicate over these issues. Therefore, SBP remains an uncertain factor that could influence SI quality and TTS lengths.

The variables interpreters have control over are IBP or EVS. If interpreters are lagging too far behind, they may try to reduce IBP, or consider not to pause at all through integrating sentence *N+1*'s rendition into sentence *N* when conditions allow so. However, not pausing at all between sentences does not appear to be a strategy that could be frequently adopted.

doi:10.6342/NTU202000181

Previous research on pauses (Wang & Tao, 2015) categorized pauses according to their position in the sentence into syntactic juncture pauses (pauses before sentences and clauses) and unsuitable pauses (pauses that occur at positions outside from the above two). They compared the pause patterns of expert interpreters and trainee interpreters and found that for both groups, around half of the pauses were syntactic juncture pauses, signifying that pausing between sentences or clauses is a common practice.

The length of TTS is a formula of these variables. Different lengths of TTS generate different cognitive load scenarios. Three scenarios may occur:

(1) TTS is a negative number or zero, suggesting that the interpreter finished producing the target language sentence before the speaker finished his/her sentence. This scenario is illustrated in Figure 4 and Figure 5 below.

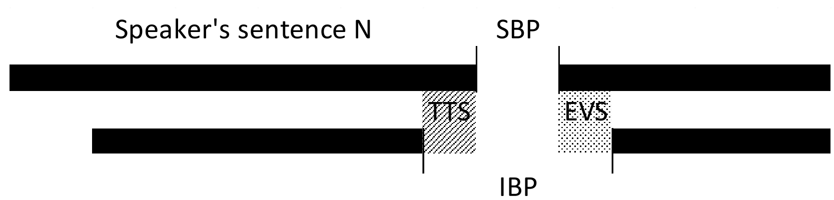


Figure 4 TTS is negative

Figure 5 TTS is zero

doi:10.6342/NTU202000181

The likelihood of these two scenarios happening during real-life SI is not high, considering that interpreters lag behind the speaker from the very beginning of the discourse, and that interpreter's overall speech proportion was found to be lower than that of the speaker, meaning that interpreters start later and move on slower (Lee, 1999). However, there could also be cases where the interpreter successfully anticipates that the end of a sentence is coming, considers it necessary to finish the rendition before or at the same time with the speaker, and manages to do so. This could happen when, for example, the entire discourse is coming to an end and the interpreter manages to end the final sentence together with the speaker; or when the speaker is asking questions to initiate interaction with the audience, and the interpreter speeds up to maintain the dynamics of the interaction.

(2) TTS is shorter than SBP. This suggests that the interpreter finished rendering sentence *N* after the speaker finished sentence *N* but before the speaker began sentence *N+1*. This scenario is illustrated in Figure 6:

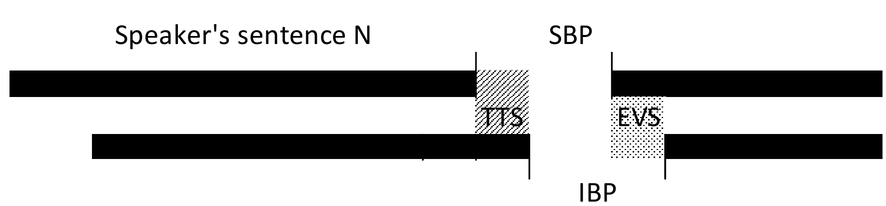


Figure 6 TTS larger than 0 but shorter than SBP

doi:10.6342/NTU202000181

(3) TTS is longer than SBP. This means that the interpreter is still managing to complete the rendition of sentence *N* when the speaker already began with sentence *N+1*. Figure 7 illustrates this scenario:

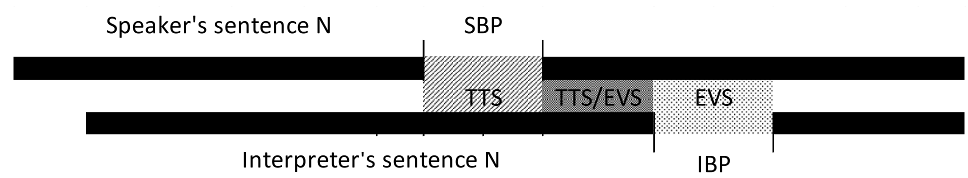


Figure 7 TTS is longer than SBP

In this scenario, the target speech of sentence N overlaps with the source speech of sentence N+1. Therefore, part of the TTS of sentence N will coincide with the EVS of sentence N+1. In other words, sentence N's cognitive load has been imported into sentence N+1. This part can also be called the "imported cognitive load" from sentence N into sentence N+1.

According to Lee, the first two scenarios are ideal for interpreters, because when they finish the sentence just in time, they can enjoy a short period of time in which they either do not have to listen to anything at all or can focus on listening to the incoming new segment. One of the implicit assumption of Gile's effort model is that the total cognitive resource required for any two or three efforts together will be larger than the resource required for a single effort, and this assumption is supported by various anecdotal records (Giles, 2008). Therefore, not having to listen and speak at the same time is certainly a relief for interpreter's cognitive load. However, in real-life, most interpreters do not get to enjoy this kind of scenario often;

doi:10.6342/NTU202000181

rather, the production of the tail of sentence *N* would usually overlap with the head of sentence *N+1*, which is the case of the third scenario. In this scenario, interpreters would have to manage to produce target speech for the segment on hand, while listening to the next sentence the speaker is saying and storing it in working memory; these altogether impose heavier cognitive load for respective tasks. Cognitive load will only start to reduce when the interpreter finishes the production task for sentence *N* and moves to comprehend and produce the target text for sentence *N+1*. In practice, interpreters do try to remain a short TTS to avoid the negative consequences of high cognitive load arising.

A practicing interpreter participating in Chang & Schallert's (2007) interpreting research was quoted on his/her view on consciously manipulating his/her TTS in an interview:

> There seemed to be something important coming, so I finished the sentence as soon as I could. When the next sentence is about something new and specific, I want to finish the previous sentence ASAP. Just to convey the message, even if the style is not good as it could have been. I can't miss to afford the upcoming message (Chang & Schallert, 2007).

Jones also suggests in his book that interpreters should attempt to wrap up their output if they sense that the current sentence they are working on is approaching the end (Jones 1998, as seen in Yagi, 2000). Practitioner's reflections and instructor's views both indicate that controlling the length of TTS has always been a rule interpreters follow implicitly to maintain the quality of SI.

doi:10.6342/NTU202000181

2.3.1 Factors influencing TTS length

In Lee's (2003) study on TTS, he looked into the variables influencing TTS length and how TTS length influenced SI quality. Lee compared two groups of SI renditions, one with TTS longer than 5 seconds and the other with TTS shorter than 3 seconds. He concluded 5 independent variables that influence the length of TTS:

a. The length of the source speech sentence (counted by syllable in a sentence) exerts a "marginally significant correlation" with the length of its TL sentence's.

In other words, the more syllables the source speech sentence has, the longer the TTS of its target speech sentence tends to be. The number of syllables in a sentence is often used to measure sentence readability. In readability indexes, the average length of syllable per word, the average length of syllable per sentence, the percentage of monosyllabic words and the percentage of polysyllabic words are commonly adopted as parameters (Crossley, Allen, & McNamara, 2011; Zamanian & Heydari, 2012). Many of the readability formulae assume that various syllable counts are positively related to sentence difficulty. Therefore, the tendency between syllable count per sentence and its TTS length could suggest a relationship between sentence difficulty and TTS length.

b. The higher the source speech sentence's speech proportion, the shorter the target speech sentence's TTS, which can be approximated as a negative relationship between speech delivery rate and TTS length – the higher the source speech's speech delivery rate, the shorter the target speech's TTS length.

doi:10.6342/NTU202000181

The term speech proportion used by Lee here should be distinguished from the aforementioned speech proportion in relation to SBP, which looks at pauses between sentences and clauses. The speech proportion here refers to the ratio of pauses within a sentence. This result differs from some of the previous researches on speaker's delivery rate. Barik (1973) for example found that English-French interpreters lagged further behind in SI when the delivery became faster; Pio (2003) found that increased source speech delivery rate resulted in more omissions in both expert and student interpreters in the German-Italian language pair. High source speech delivery rate is even considered by Gile to be "one of the most prominent problems for interpreters"(Gile, 1995). Chang also observed that Chinese-English interpreters made omissions, information selections, and generalizations more frequently when speech rate was fast (Chang & Schallert, 2007).

However, Lee's (2002) research on English-Korean SI yields different results; in his study, he found that EVS shortens as speech input rate increases. Lee suggests that a possible explanation for the negative correlation between both types of lag and speech input rate is that, in a given time frame, interpreters are able to comprehend the source speech faster when the source speech is delivered at a higher rate, thus allowing interpreters to begin producing their rendition earlier (Lee, 2003).

c. Interpreter's sentence length has a positive correlation with TTS. In other words, sentences that require interpreters longer time to utter will cause TTS to increase.

d. The higher the interpreter's syllable to speaker's syllable ratio in a sentence, the longer its rendition's TTS. From a temporal perspective, it is understandable that when the interpreter

doi:10.6342/NTU202000181

has more syllables to utter than the speaker does, it will take longer time for the interpreter to finish utteting the sentence.

e. Interpreters syllable per minute and TTS were negatively correlated. In other words, the slower interpreters speak, the longer the TTS. While speaking with a low speech rate would take longer to complete the sentence seems natural from a temporal perspective, it is worth mentioning that Lee also noted in his research that normally interpreters would not lower their speech rate unless they consciously choose to or are pressured to do so; when interpreters slow down, it is likely that they are struggling with processing difficulties such as trying to comprehend the source text or gropling for the correct equivalent in the target language (Lee, 2003).

f. Longer EVS at the beginning results in TL sentences with long TTS. This corresponds with "imported cognitive load" and "imported time lag". However, it should be noted that the length of TTS and EVS would not necessarily be the same. It is hard to infer whether longer EVS suggests anything about interpreter's processing capacity or about characteristics of the source speech, as the EVS length's increase and decrease could be subject to individual preferences or strategies.

To conclude, temporal aspects of the source speech and of the interpreter's rendition play a central role on TTS length, meaning that time management has a role to play in SI accuracy. In addition, based on the deduction that longer sentences (counted by syllables) imply greater difficulty and complexity and the observation that longer source speech sentences tend to lead to slightly longer TTS, it could also be speculated that source material difficulty also influences the length of TTS. Lee's observations on interpreter's speech rate (syllable per

doi:10.6342/NTU202000181

minute) also indicates a possibility that longer TTS is caused by source material difficulty based on the assumption that interpreters do not slow their speech rate unless they encounter processing difficulties, one of which being source material difficulties. This is also in line with previous studies that found that processing difficulties arising from the source material such as numbers, complex sentence structures, and high information density could lead to pauses or speech disfluencies (Cecot, 2001; Ahrens, 2005; Wang & Tao, 2015).

2.3.2 The influence of TTS length on SI quality

Lee also found in his research that while the aforementioned five variables influence the length of TTS, the length of sentence *N*'s TTS has an adverse effect on the accuracy of sentence *N+1*. Lee postulates that longer TTS leads to poorer accuracy in the following sentence because the limited processing capacity for the next sentence would be depleted. More specifically, the listening and analysis effort would be hampered, as large parts of the following sentence would have to be comprehended while the interpreter attempts to finish up target speech production for the previous sentence. The competition for interpreter's limited efforts could lead to saturation and cause a failure sequence.

2.4 EVS and TTS

EVS and TTS are two concepts that are both crucial to the temporal aspects of SI. While they both refer to the distance at which the target speech lags behind the source speech, EVS measures the lag at the onset of a sentence and TTS measures the lag at the end. The length of EVS and that of TTS are inextricably interdependent. A sentence's EVS will usually, to different extents, be exported to the very same sentence's TTS, while this TTS will, to different extents, be exported to the next sentence's EVS and TTS.

doi:10.6342/NTU202000181

Despite the strong association between EVS and TTS, there is one crucial differentiator. One of the many determining factors of the length of EVS is interpreters' idiosyncratic preferences. As discussed in section 2.3, some interpreters prefer a shorter EVS to unburden the memory effort as much as possible, whereas other interpreters prefer a longer EVS for better source speech comprehension. While views split on whether a long EVS or short EVS is better, views on TTS are more consistent. Jones' (2014) publication and Chang & Schallerts' (2007) study both mentioned finishing rendering a sentence as early as possible to spare for the next sentence. Practicing interpreters have repeatedly mentioned finishing rendering a sentence as early as possible to preserve processing capacity for the next sentence. It makes clear sense for interpreters to have a short TTS unless they are unable to do so due to certain challenges such as having to explicate or explain an item in the source speech because the target language lacks an equivalent expression or encountering an unexpected turn at the end of the source speech sentence that requires interpreters to reroute their planned or delivered interpreting output. Therefore, interpreters would benefit from reducing the length of their TTS, while they enjoy more freedom to adjust the length of their EVS. In addition, reducing the length of TTS without hampering SI output quality would require the adoption of strategies such as selective omission, compression, or anticipation, whereas reducing or increasing the length of EVS can be done through strategies such as waiting or stalling.

2.5 Research question and rationale

Lag time is a key dimension and topic of study in SI. Although many studies have explored this topic through examining the lag time at the onset of the speech units through the concept

doi:10.6342/NTU202000181

of EVS, few studies have looked at the lag time at the end of a speech unit, which is referred to as TTS. While a sentence's EVS and TTS frequently overlap each other and affect each other's length, the crucial differences between EVS and TTS discussed in section 2.4 provides ground for establishing TTS as a topic separate from EVS and giving dedicated efforts to exploring how the length of TTS interplays with other aspects of SI.

Lee was the first to formally and systematically study the effects of TTS. While he explored the relationship between TTS and a number of variables in English to Korean SI and provided possible explanations for his results, TTS remains an under-explored topic in comparison with many of other topics in SI research. Other crucial concepts in SI such as EVS, accuracy, fluency, aptitude, and directionality have been studied widely and many of the studies have been replicated in identical settings or tried under different language pairs and directionality, or with different pairs of participants such as trained and self-made interpreters or experts and novice/students, while many of these angles that have been well approached in the study of other SI topics of research are still missing in TTS research. Seeing this gap, this study hopes to replicate Lee's research to enrich and add new dimensions to the field of TTS research.

This research aims to answer how TTS length affects SI quality in SI from English into Chinese. This question has already been answered partly by Lee's research. However, there are a few reasons why this study still sets out to explore the question again.

Firstly, this study explores the relationship between TTS and quality in SI from English into Chinese, as opposed to Lee's study of SI from English into Korean. SI from English into Korean and English into Chinese are both syntactically dissimilar with English; while

doi:10.6342/NTU202000181

English is a strictly SVO and form-dependent language, Korean and Chinese both have an SOV structure. This provides some ground to predict that TTS experiment results in English to Chinese SI may differ from that in English to Korean SI. On the other hand, Chinese is more SVO than Korean in syntactic structure, and Chinese is also considered more "concise and compact" compared to English (Wu, 2001), whereas Lee has mentioned that Korean is verbose compared to English(Lee, 2003). It is thus worth exploring how TTS results of English to Chinese SI would differ from that of English to Korean SI.

Secondly, this study aims to explore the relationship between TTS and interpreting quality holistically, looking at both accuracy and fluency of the SI product. In Lee's research, he found that long TTS had an adverse effect on interpreting accuracy. If long TTS causes poorer accuracy because the interpreter would be more likely to suffer from cognitive resource depletion, then it may be assumed that the target speech fluency would also be affected. Studies have connected types of disfluencies such as silent pauses, self-correction and repetition with processing difficulties during SI. If processing difficulties are reflected by both disfluencies and poorer accuracy, then it is possible that fluency would be affected by TTS length just alike. In addition, fluency is also a quintessential part of interpreting quality. Christodoulides & Lenglet (2014) conducted an experiment where the same piece of SI rendition was played out to listeners under a read fluent mode and a natural SI disfluent mode, and found that listeners rated the fluent mode as more accurate despite the fact that the content of the two versions were identical. It could thus be inferred that an accurate but disfluent speech will still hamper SI quality in the ears of users. For the reasons above, this study will compare the relationship between TTS and holistic SI quality, which includes both accuracy and fluency.

doi:10.6342/NTU202000181

Thirdly, Lee's research used heterogeneous SI live recordings as material, which has strong ecological validity, but the replicability and inter-sample comparability may be weaker due to the heterogeneity of the collected samples. This research, on the other hand, uses experiment material to reduce possible disturbance arising from non-TTS factors and ensure that the relationship between TTS and quality could be replicated in experimental settings to further support Lee's findings.

doi:10.6342/NTU202000181

III. Methods

3.1 General Methods of the Experiment

This experiment aims to collect quantitative data for assessing the impact of TTS on interpreting quality in English-to-Chinese SI by having trainee interpreters perform simultaneous interpretation for a speech adapted from real-life speech. The text was adapted in such a way that certain sentences in the speech could help reflect whether participants' performance was affected by the lengths of their TTS and whether they were able to recover from those effects, if any. The text is manipulated so that the SI performance of the sentences in focus would be affected as little as possible by factors other than TTS. Experiment data will be collected from these designed sentences and processed with SPSS statistics to test the research hypothesis. In addition to the quantitative data gathered from the experiment and statistical processing, qualitative data (i.e. transcripts of the participants' interpretation) will be used to provide possible explanations for the quantitative findings.

3.2 Subjects

Seventeen trainee interpreters were recruited from three interpreter training institutes in Taipei, Taiwan, to participate in the experiment. All seventeen participants are senior students or recent graduates of the three training institutes, and had received at least one full year of training in simultaneous interpreting. It is assumed that completing one year of training in simultaneous interpreting will be adequate for participants to generate SI output with acceptable quality. According to the background survey administered to the participants, the participants had completed two to five SI courses at their respective institutes at the time of the experiment. The mean count of courses taken by the participants is 3.8 courses. All

doi:10.6342/NTU202000181

participants except for one had Chinese as their A language and English as their B language. Three of the participants participated in the pilot study, and the remaining fourteen participants participated in the formal study.

3.3 Material

3.3.1 Overview of the Material

The material for the present experiment is a fifteen minute speech adapted from the original talk notes of Jerry Mintz, a specialist in alternative education, for his TED Talk "Pioneers in Alternative Education". The complete experiment material can be found in the Appendix of this thesis. This piece of text was chosen as the experiment text for several reasons. Firstly, the talk notes are in the form of a speech script, and is thus suitable to be used as the source text for simultaneous interpreting even without any adaptations. The content of the talk is also highly accessible to the general public, because the purpose of the talk is to introduce and promote alternative education to the general public. Finally, the actual speech delivered by the speaker Mintz on the TED stage is substantially different from the talk notes published on Mintz's website. Therefore, it is less likely that the participants would have had listened to or performed simultaneous interpretation with this piece of text before.

The text was further adapted to include a number of critical sentence sets for the collection of TTS and interpretation quality rating, and to level information load across the speech. The adapted text was then read aloud by an English native speaker at a speech rate of 130 wpm. A pilot study with three participants was conducted to evaluate the appropriateness of the speech for simultaneous interpretation, and necessary modifications were made after observing the participants' performance in the pilot study.

doi:10.6342/NTU202000181

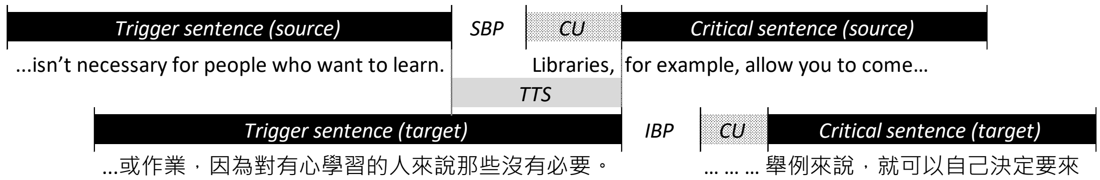
3.3.2 The Design of Critical Sentence sets

The finalized text contains 19 critical sentence sets, which occur at regular intervals throughout the text. The intervals contain similar information loads and have similar delivery speeds. Each critical sentence set contains two sentences: a trigger sentence and a critical sentence. TTS will be collected from the trigger sentence and analyzed for its effects on the performance of the critical sentence.

The beginning of the critical sentence contains a critical unit that is crucial to the comprehension of the critical sentence. Critical units are selected and agreed upon by the researcher and two other fellow interpreter trainees, who did not take part in the experiment as participants. The critical unit serves as one of the parameters for assessing the impact of TTS on interpreting quality. Since the critical unit occurs at the beginning of the sentence, it is likely that its appearance in the source speech overlaps with the target speech of the previous sentence, hence the higher likelihood of participants not hearing or misunderstanding the critical unit. Given the importance of the critical unit to the meaning of the entire sentence, failing to catch the critical unit would be reflected in participants' interpreting performance. It is thus assumed that long TTS will ultimately undermine the interpreter's ability to render the critical unit correctly.

Figure 8 below illustrates the design with one of the actual critical sentences used in the experiment, paired with a fictional translation to illustrate the temporal overlapping between the source speech critical unit and the target speech.

doi:10.6342/NTU202000181



SBP: Speaker's between sentence pause
IBP: Interpreter's between sentence pause
CU: Critical unit

Figure 8 Graphic illustration of the critical unit's position in critical sentence sets

In this example, the critical unit is "libraries". It is assumed that if the participants are still rendering the trigger sentence when the speaker begins the next sentence with "libraries", just as in Figure 8, where TTS and the source speech critical unit overlaps, participants will likely not hear the word "libraries" or misunderstand it, and substitute it with educated guesses such as "these schools" or "the schools' libraries".

3.3.3 Information Density

One of the aims of making adaptations to the original talk notes is to level information load throughout the text. Without so doing, sudden surges or decreases in information density, which has been cited as a cause of variability in cognitive load in simultaneous interpreting in previous studies (Giles, 2008), might affect participants' performance. In particular, abrupt increase in cognitive load has been found to place a heavy burden on interpreters, or even to cause a "knock-on effect" or "failure sequence" in which the cognitive load of a certain segment is carried over to its pursuant segment and distresses the processing (Seeber, 2013; Gumul & Lyda, 2007; Liu & Chiu, 2009). Since the effects of varying information density may in turn influence interpreting performance, the design of the text aims to avert

doi:10.6342/NTU202000181

fluctuations in information load throughout the text to the extent possible, so that interference from factors on interpreting performance other than TTS can be minimized.

To this end, the original text was manipulated to contain equal and symmetric amount of information throughout the text. Information load is measured by "idea units", a unit for analyzing the meaning of a piece of text proposed by Kroll in 1977 (as seen in Crookes, 1990). Kroll's definition of an idea unit was adopted amongst other definitions of idea units because it was designed for analyzing spoken text. Kroll's criteria for chunking idea units are as follow:

(1) a subject and verb counted as one idea unit together with (when present) a (a) direct object, (b) prepositional phrase, (c) adverbial element, or (d) mark of sub-ordination;

(2) full relative clauses counted as one idea unit when the relative pronoun was present;

(3) phrases which occurred in sentence initial position followed by a comma or which were set off from the sentence with commas were counted as separate idea units;

(4) verbs whose structure requires or allows a verbal element as object were counted with both verbal elements as one idea unit;

(5) reduced clauses in which a subordinator was followed by a non-finite verb element were counted as one idea unit;

(6) post-nominal-ing phrases used as modifiers counted as one idea unit;

doi:10.6342/NTU202000181

(7) other types of elements counted as idea units were (a) absolutes, (b) appositives, and (c) verbals.

Following Kroll's criteria, the text was chunked into idea units by the researcher and one of the fellow trainee interpreters who helped select the critical units. The text was then adapted so that each section of the text contained similar numbers of idea units. The sentences in the text are classified into three types. The first and second type are trigger sentences and critical sentences, which together form a critical sentence set. In principle, trigger sentences contain three idea units, with a few exceptions that contain two or four idea units. Critical sentences contain one idea unit, with a few exceptions that have two idea units. The third type of sentence is the sentences that occur before or after the critical sentence sets (i.e. trigger sentence and critical sentence). For the sake of this research, these sentences will be referred to as "non-trigger sentences". Each chunk of non-trigger sentences that occur before, between or after critical sentence sets are referred to as a "non-trigger chunk". Each non-trigger chunk contains eleven idea units, with a few exceptions containing one to three idea units less than the standard number of idea units.

In addition to balancing the number of idea units contained in each part of the text, the duration of pauses are also controlled. Pauses between trigger sentences, between non-trigger sentences and pauses between trigger sentences and critical sentences were manipulated to be of consistent lengths throughout the speech. The pauses between trigger sentences and the pauses between non-trigger sentences are consistently 0.75 seconds, and the pauses between trigger sentences and critical sentences are consistently 2 seconds.

Table 1 below illustrates the structure of the text in terms of sentence types and pause lengths:

doi:10.6342/NTU202000181

Table 1 Structure of the experiment text

Sequence of sentences	Part of text	Type of sentence	Idea unit count
1		Non-trigger sentence #1-a	
		0.75 second pause or natural pause	
2	Non-trigger chunk #1	Non-trigger sentence #1-b	11
		0.75 second pause or natural pause	
3		Non-trigger sentence #1-c	
	Two-second pause		
4		Trigger sentence #1	3
	Critical sentence set #1	0.75 second pause or natural pause	
5		Critical sentence #1	1
	0.75 second pause or natural pause		
6		Non-trigger sentence #2-a	
		0.75 second pause or natural pause	
7	Non-trigger chunk #2	Non-trigger sentence #2-b	11
		0.75 second pause or natural pause	
8		Non-trigger sentence #2-c	

The first rationale for manipulating pauses is to curtail the knock-on effect (or spillover effect) of cognitive load from non-trigger chunks to the critical sentence sets. Although in natural simultaneous interpreting scenarios, the knock-on effect from earlier source speech segments may occur, allowing the cognitive load from non-trigger chunks to be carried over to the critical sentence sets will severely complicate the qualitative analysis. Given the under-

doi:10.6342/NTU202000181

explored state of TTS in interpreting research, the present study aims to contribute to the research of TTS through investigating its impact on interpreting performance and the immediate causes of varying TTS length. If the participant fails to interpret the trigger sentence because of the cognitive load from non-trigger chunks, then the TTS of the trigger sentences will become unavailable. In addition, if the participant's rendition of the trigger sentence is severely impacted by the knock-on effect from the non-trigger chunks, it would be hard to reveal the immediate causes such as idiosyncratic pause patterns or linguistic habits that contribute to the variation of TTS lengths. On the other hand, with the two-second pauses insulating the processing of critical sentence sets from the cognitive load induced by the non-trigger chunks, the qualitative analyses can focus on the attributes of the participants' rendition of the trigger sentence, except for rare cases where a two-second pause is not sufficient for the participant to recover from the carry-over effect of previous segments.

The second rationale for manipulating pauses is to ensure that the joints where two audio files of the source speech are spliced together have equal pause lengths. Since the text is of a considerable length (1,465 words), to prevent having to repeatedly record the entire speech for minor disfluencies, the text was recorded in six batches. Each batch contained around one sixth of the experiment text. When splicing two recording files together, a pause would have to be inserted at the joints to imitate the natural pause between sentences. For the purpose of consistency, a pause length of 0.75 second was used after observing the natural between-sentence pauses of the speaker. As a result, pauses other than those preceding the

doi:10.6342/NTU202000181

3.3.3 Recording of the Material

The text for the pilot study and the formal study were both read out by a North American native speaker of English, who had completed two years of training in conference interpreting as well. The text was provided to the speaker in advance, and the speaker was told to familiarize himself/herself with the text and to read the text at a moderate pace and as naturally as possible.

As described in the previous section, to prevent having to repeatedly record the entire speech over and over again for disfluencies, the speech was recorded in six parts. The speech was recorded with the speaker's iPhone Voice Memo recording function, and spliced and edited into a single audio file containing the complete experiment speech using the audio file editing software Audacity. The volume and pitch of different recording files were also equalized using the software.

After the pilot study, some sentences have been modified and recorded again by the same speaker. These sentences were then inserted into the audio file of the entire speech to replace their previous versions. Where audio files had to be spliced together, pauses were inserted according to the aforementioned pause patterns.

The final material contains 1465 words, and the recording lasted 11 minutes and 18 seconds, with a speech rate of 129.6 wpm.

3.4 Procedures

The participants took part in the experiment one at a time. In total, seventeen experiment sessions were held. Three participants participated in the pilot study. All three pilot

doi:10.6342/NTU202000181

experiments took place within two days and in a quiet classroom in the College of Arts Building at National Taiwan University. Fourteen participants participated in the formal study. The experiments were held across a time span of three weeks, and took place either in the same classroom used for the pilot study or a quiet private study room in the library of National Taiwan University. Each experiment lasted for around thirty minutes.

Upon arrival at the designated location for the experiment, the participants were briefed on the background and procedure of the experiment, and asked to sign a consent form if they agreed to take part in the experiment. They were then asked to fill out a background questionnaire, which collected information including age, institute where they studied, language combination, number of classes in consecutive interpreting and simultaneous interpreting taken and years of conference interpreting working experience, if any.

Afterwards, they were given a glossary sheet, which contained a brief biography of the speaker, the definition of key terms and those terms' equivalents in Chinese. Participants were allowed to study the glossary until they felt that they were well prepared enough, and they were told that they were allowed to use the glossary when they were interpreting. When they notified the researcher that they were ready, they were asked to first perform SI on a warm-up speech. The warm-up speech was an interview with the original speaker of the speech, Mr. Jerry Mintz, on his views on traditional education and alternative education. The recording of the warm-up speech lasted for three and a half minutes.

A Macbook Pro laptop was used as the apparatus for the experiment. The source speech of the warm-up exercise was saved into an Audacity audio file with a second audio track created for the recording of the target speech. The participants listened to the source speech through

doi:10.6342/NTU202000181

earphones connected to the laptop, and spoke through the microphone of the earphones. A random segment in the middle of the warm-up speech was first played for the participants to ensure that the volume was suitable. The participants were also allowed to adjust the volume themselves any time during the warm-up speech, if they felt uncomfortable with the volume. When the volume was confirmed to be comfortable for the participants, they were told to click on a start icon that would start playing the source speech and recording the target speech on to the second audio track in the same audio file. Afterwards, the recordings of the warm-up speech were saved and played immediately, so as to ensure that the recordings were intelligible.

After the warm-up speech, participants were told that they were allowed to read through the glossary again. When they were finished, the same procedures for the warm-up speech was used for the formal speech.

Finally, each of the recordings were saved into an Audacity project file and an mp3 file, labeled with the dates and participant number. The files are dual-track audio files, with the English source speech in the first soundtrack and the Chinese target speech in the second soundtrack and synchronized with the English source speech.

3.5 TTS Measurement

After collecting the interpretation of participants, the TTS length of their trigger sentences were measured and recorded. Past studies have adopted a variety of methods for measuring time lag between the source speech and target speech in simultaneous interpreting. The methods varied mainly in the points of measurement and the unit of measurement.

doi:10.6342/NTU202000181

Table 2 below integrates the lists of measurement methods of EVS in past studies by Timarová, Dragsted, & Hansen (2011) and Chen (2012), joined by a number of later studies. For points of measurement, past studies have adopted time intervals, syntactic units, semantic units or a mixture of the latter two units (Diaz-Galaz, Padilla, & Bajo, 2015) as their point of measurement. Using fixed time intervals or clearly defined syntactic units such as a complete clause, sentence or phrase as reference points for measurement has the benefit of intra- and inter-study consistency. On the other hand, using semantic units as reference points for measurement circumvents difficulties in matching source speech and target speech when the interpreter has summarized ideas or made substantial changes to the sequence of information presented (A. L. Chang, 2009). For units of measurements, measuring EVS in words or syntactical units was the prevalent unit for lag measurement in earlier days. In more recent studies, time measurement has become the more prevalent unit for lag measurement.

doi:10.6342/NTU202000181

Table 2 Methods of lag measurement adopted by past studies

Author	Year	Point of measurement	Unit of measurement
Treisman	1965	every 5 seconds	words converted to seconds
Oleron and Nanpon	1965	words with literal translation	seconds
Gerver	1969	every 5^{th} word of source text	words
Goldman-Eisler	1972	unclear	syntactical units
Barik	1973	every 5 seconds	seconds
Anderson	1994	words	seconds
Lamberger-Felber	2001	beginning of segments where at least one interpreter omitted more than 15 words	seconds
Lee	2002	beginning of sentence	seconds
Kim	2005	beginning of sentence	seconds
Podhajska	2008	unit of meaning	seconds
Ono, Tohyama, & Matsubara	2008	words	seconds
Chang	2009	beginning of primary message and secondary message	seconds
Chen	2012	beginning of sentence	seconds
Diaz-Galaz, Padilla and Bajo	2015	source text segment (phrases or clauses)	milliseconds

doi:10.6342/NTU202000181

For TTS measurement methods, not many past studies could be cited. Lee (2003), who pioneered the research of TTS, measured TTS using sentences as reference points and seconds as unit of TTS length. In a research on the effects of advance preparation in simultaneous interpretation, Díaz-Galaz (2012) mentioned using “source text speech segments” as a point of measurement and seconds as unit of measurement for measuring the lag at the end of the sentence, which might be the same as or similar to TTS. The segements in Díaz-Galaz’s research were determined first by selecting difficult phrases or clauses and then forming segments around the selected phrases and clauses according to syntactic units.

Since this research aims to continue upon Lee’s (2003) findings and also to allow for future comparison between TTS and EVS, the most prevalent measurement methods are adopted. Sentences will be used as the point of measurement, and seconds will be used as the unit of measurement.

Audacity was used for calculating the TTS of each trigger sentence. The onset and end of source speech trigger sentences and target speech trigger sentences were manually identified on the audio tracks in Audacity and the timecode was recorded and rounded to the first decimal point. Figure 9 below is a screenshot of the software interface with a multi-track audio file. The source speech is the first track and target speech is the second and third track. The timecodes were then entered into an Excel sheet. TTS is then calculated by subtracting the timecode of the end of the source speech trigger sentence from that of the target speech trigger sentence.

doi:10.6342/NTU202000181

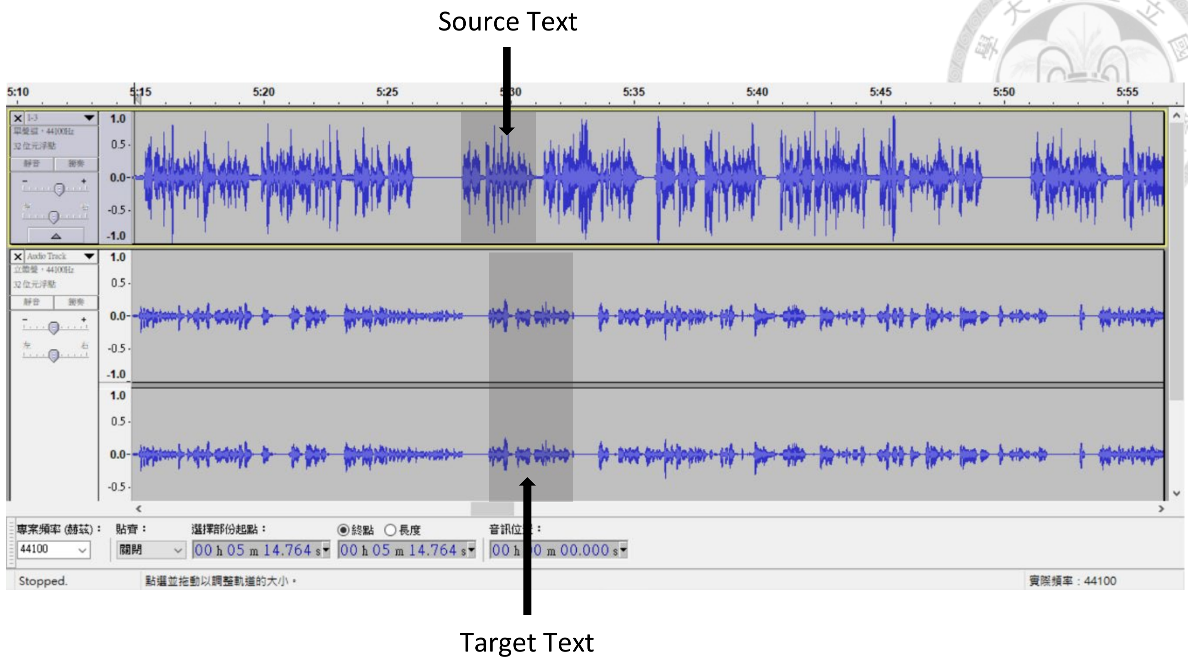


Figure 9 Screenshot of the dual track

3.6 Rating of Interpretation Quality

The two fellow interpreters who participated in the selection of critical units acted as co-raters. The researcher acted as the third rater. The two co-raters rated the first half of the participants' performance (Participant #1 to #7) and the second half of the participants (Participant #8 to #14) respectively; the researcher rated all of the participants' performance. Therfore, each participant's performance was rated by two raters.

To ensure that the criteria for rating was consistent amongst all three raters, the researcher and the two co-raters conducted an inter-participant and an intra-participant comparison of their respective rating results before rating all of the recordings. In the inter-participant comparison, three out of the fourteen sentences from the source text were randomly selected

doi:10.6342/NTU202000181

as the sentence for rating criteria comparison. Then, for each trigger sentence, three participants' interpretations were randomly selected to be rated by the three raters respectively. The raters then compared the results of their scores, discussed the possible reasons that might have caused major discrepancies in the scores given, and agreed upon the resolution of key issues that have caused score discrepancies. Next, the raters rated all fourteen sentences of a participant, compared the scores, and again discussed and reached an agreement for major discrepancies. After ensuring that the raters had reached a concensus on the rating criteria, the raters rated the sentences independently.

The design of quality rating aims to reflect the overall quality of the interpretation. Lee's research looked at how TTS affected accuracy in SI (Lee, 2003). This study adds a second dimension to quality assessment: fluency. Past studies on interpreting quality frequently included fluency as important dimensions of interpreting quality, whether the study focused on the perspective of users, interpreters, trainers or researchers (Kurz, 2001; Amini, Ibrahim-González, Ayob, & Amini, 2015; Zwischenberger, 2010; Macías, 2006). Since conference interpreting is a highly goal-oriented activity, which means the achievement of its goal – bridging language gaps and facilitating communication – is quintessnetial to the task, it is sensible to take into account a quality critereon considered important by all major stakeholders involved in the activity.

In addition to assessing the overall quality of SI in relation to TTS length, an additional hypothesis is also tested – the recovery effect hypothesis. Recovery effect in this study refers to the phenomenon that interpreting quality improves as interpreters are gradually relieved from the heavy cognitive load or the negative effects thereof caused by long TTS. The

doi:10.6342/NTU202000181

hypothesis that interpreting quality is affected by the length of TTS of its previous sentence is based on the presumption that long TTS imposes heavier burden on the interpreter; as interpreters finish off producing the rendition of the previous sentence, they are relieved from the heavy cognitive load or the accompanying negative effects. Proceeding forward and distancing themselves from the turbulence caused by long TTS, interpreters' performance should gradually recover. To examine whether this recovery effect exists in interpreters' performances, quality assessment is designed to includes three parts: the critical unit score, the score of the first seven seconds, and the score of the second seven seconds. The three scores correspond to three different parts of the target speech of the critical sentence in chronological order. The critical unit score refers to whether the participant correctly rendered the critical unit of the critical sentence, which occurs at the beginning of the critical sentence. The first seven seconds score corresponds to the first seven seconds after the end of the trigger sentence's target speech. The second seven seconds refers to the eighth to fourteenth sentence after the end of the trigger sentence, following immediately after the first seven seconds. The purpose and measurement for each quality score is explained in details below.

3.6.1 Critical Unit Score

The first part is whether the participant rendered the critical unit correctly. For the sake of this research, this score is referred to as the "critical unit score" hereinafter. The participant receives one point for correctly rendered critical units and zero points for erroneously rendered ones. Since the critical units are mostly, if not always, a single unit of meaning, which means there is relatively little space for ambiguity in the correctness of its rendition, it was decided that the scores given by the two raters must be unanimous. In circumstances

doi:10.6342/NTU202000181

where the scores given by the two raters for the same sample were not unanimous, the raters underwent deliberation to reach a consensus. The samples either had 1 point or 0 points as their final critical unit score.

3.6.2 The First Seven Seconds Score and the Second Seven Seconds Score

The second part is the participant's overall performance in the seven seconds after the end of the source speech trigger sentence. This score is referred to as the "0-7th second score". Figure 10 below illustrates where the 0-7th rating interval takes place.

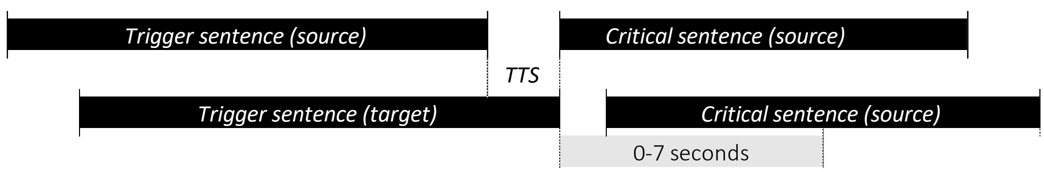


Figure 10 Graphic illustration of the position of the 0-7th second rating

Participants are rated for the fluency and accuracy of their interpretation, with each accounting for five points. If the rater determines that the participant has delivered a perfectly accurate and fluent rendition of the source speech, the participant receives ten points in total (five points for fluency and five points for accuracy) from that rater. The scores from the two raters are then added up and averaged. As a result, each critical sentence receives a final score that ranges from zero to ten for the 0-7th second score. For example, if one rater gave a total score of eight points, with five points for fluency and three points for accuracy, and another rater gave a total of six points, with four points for fluency and two points for accuracy, then the final score would be seven.

doi:10.6342/NTU202000181

The raters listened to the audio recording of the first seven seconds, and judged its accuracy and fluency. To ensure that the intervals in which participants were rated for remained of consistent length, the start and end time of each critical sentence sets of the interpretation were calculated and used as the standard timecode for snipping the recording corresponding to the first seven seconds in each participant's recording into separate audio files. The raters listened to these isolated audio recordings of the first seven seconds, while using the transcripts of the source speech and the participants' interpretation as aid. The transcript of the source speech allowed the raters to assess the accuracy of the target speech. The transcript of participants' interpretation served as reference material when raters felt that the seven seconds in the audio recording alone was not sufficient for quality assessment. In one instance, the transcript of a participant's interpretation was used to check whether a pronoun that appeared in the interpretation in the first seven seconds referred to the correct antecedent. In another instance where the target speech was structurally very different from the source speech, the transcript was used to check whether the items presented in the source speech during the first seven seconds but not present in the interpretation were rendered afterwards or not.

The rating of the second seven seconds ("8-14th second score") was handled in the same way as the 0-7th second score was handled. For the sake of convenience, the 0-7th second score and the 8-14th second score may together be referred to as the "interval scores". Figure 11 below illustrates where the 8-14^{th} second rating takes place in relation to the progression of the source and target speeches.

doi:10.6342/NTU202000181

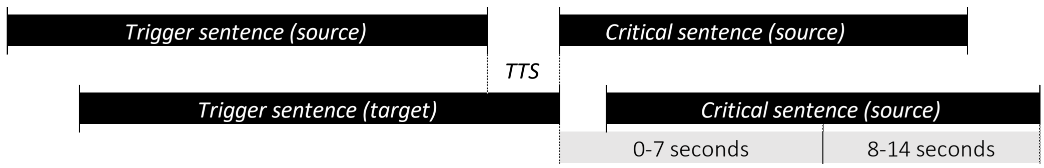


Figure 11 Graphic illustration of the position of the 8-14th second rating

Finally, a master list of data containing TTS length, critical unit score, interval scores of each sample was produced. Table 3 below is a sample of the master list. The data is processed in SPSS. Results of the statistical analysis will be presented in the next Chapter.

Table 3 Sample master list of collected data

Participant	Critical sentence set	Trigger ST-S	Trigger ST-E	Trigger TT-S	Trigger TT-E	Trigger TTS	Critical CU	Critical First7	Critical Second7
1	1	00:07.4	00:13.8	00:09.5	00:16.6	2.8	1	8.5	10
1	2	00:31.4	00:36.1	00:32.6	00:38.2	2.1	1	8	7.5
1	3	00:56.6	01:01.9	00:58.0	01:04.2	2.3	1	8.5	7.5
1	4	01:22.1	01:28.7	01:25.0	01:31.3	2.6	0	3.5	9
1	5	02:18.1	02:24.6	02:20.4	02:25.3	0.7	0	3	1

(ST = source text, TT = target text, S = start, E = end)

doi:10.6342/NTU202000181

IV. Results

This chapter reports the results of the experiment and statistical analysis. The results of descriptive statistics will be reported first, followed by the results of inferential statistics. The statistical analyses aim to examine the research hypothesis in three levels – a general analysis, at inter-participant level and at intra-participant level. The general analysis examines whether the collected data in its entirety supports the research hypothesis. The inter-participant level analysis looks at whether participants with shorter TTS have higher quality scores, and participants with longer TTS have lower quality scores. The intra-participant analysis probes whether the same participant's sentence sets with shorter TTS have higher quality scores than his/her sentence sets with longer TTS. In each level of analysis, the relationship of TTS and the three quality scores will be presented in chronological sequence of the quality scores – critical unit score, the 0-7th second score and the 8-14th second score.

A total of 266 critical sentence sets are collected from the experiment. Each participant contributed 19 critical sentence sets. Amongst the 266 critical sentence sets, 19 were excluded because the trigger sentences were not perceivably addressed by the participants, and the TTS of these trigger sentences were thus unattainable. A total of 247 critical sentence sets remained for statistical analyses, with each participant contributing 15 to 19 valid critical sentence sets. Table 4 below lists the number of critical sentence sets each participant contributes to the final set of data analyzed.

doi:10.6342/NTU202000181

Table 4 Contribution of critical sentence sets from each participant

Participant	Count of valid critical sentence sets
1	19
2	19
3	15
4	19
5	17
6	16
7	15
8	19
9	18
10	18
11	18
12	17
13	19
14	18
Grand Total	247

4.1 General Analysis

Table 5 below summarizes the distribution of TTS of all 247 trigger sentences. The mean length of TTS is 4.1 seconds; the median TTS is 3.7 seconds. The range of TTS distribution is 15.8 seconds, with the shortest TTS being -0.2 seconds, and the longest being 15.6 seconds. The 25th percentile is 2.6 seconds, and the 75th percentile is 5.0 seconds.

Table 5 Distribution of TTS

TTS (in seconds)							
Sentence Count	Mean	Median	Minimum	Percentile 25	Percentile 75	Maximum	Standard Error of Mean
247	4.1	3.7	-0.2	2.6	5.0	15.6	0.1

doi:10.6342/NTU202000181

Figure 12 below shows the overall distribution ofE TTS in a histogram. The distribution is positively skewed, suggesting that, while most participants' TTSs are shorter than the mean, some participants' TTSs are considerably longer than the mean.

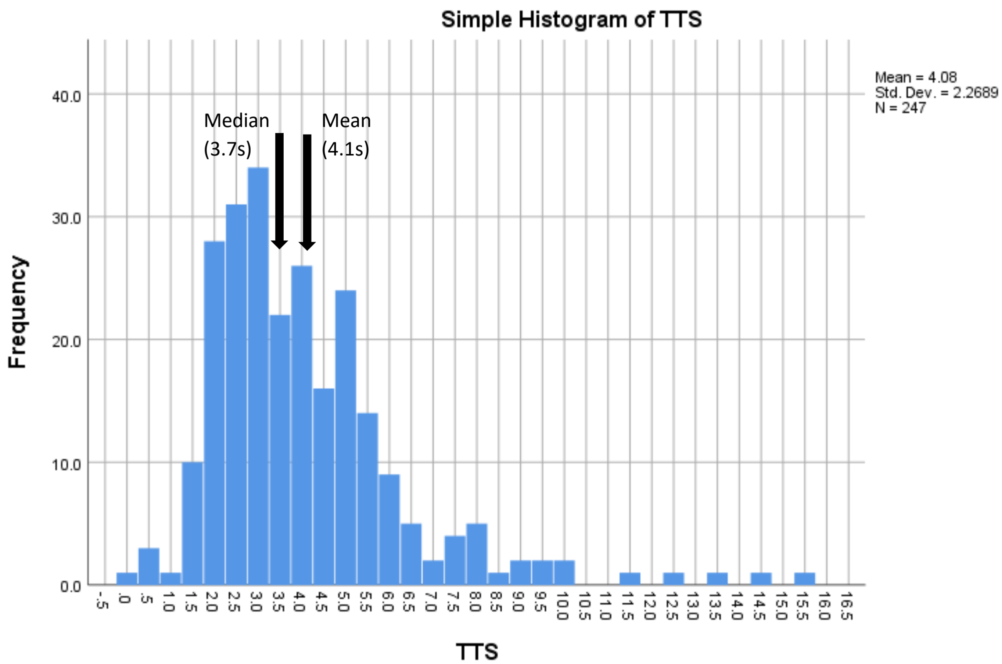


Figure 12 Histogram of TTS distribution

4.1.1 TTS and critical unit score

Table 6 below shows the distribution of critical unit scores. Out of the 247 critical sentence sets, 111 sets (44.9%) had correctly rendered critical units ("Correct" in Table 7), and 136 sets (55.1%) had incorrectly rendered critical units ("Incorrect" in Table 7).

doi:10.6342/NTU202000181

Table 6 Distribution of critical unit scores

	Sentence Count	Percentage
Correct (CU score = 1)	111	44.9%
Incorrect (CU score = 0)	136	55.1%
Total	247	100%

To examine whether CU score and TTS have a negative correlation, two sets of analyses were conducted. The first set examines whether TTS length is a strong predictor of CU score, and the second examines whether CU score can in turn predict the length of TTS.

The critical sentence sets are first grouped into three groups according to their TTS lengths – the “Long” group, the “Middle” group and the “Short” group. Each group has either 82 or 83 sentences. Table 7 below provides a summary of the group statistics. The “Mean TTS” column shows the mean TTS of each group. The “Correct” column shows the count of sentences that have correctly rendered critical units and the percentage they represent within the group, and the “Incorrect” column shows the count of sentences with incorrectly rendered critical units and the percentage they represent within the group. The mean TTS and percentage of correctly rendered critical unit supports the research hypothesis that TTS has an adverse effect on quality. The group with the shortest TTS (2.2 seconds) has the highest percentage of correctly rendered critical units (76.8%), whereas the group with the longest TTS (6.5 seconds) has the lowest percentage of correctly rendered critical units (14.6%).

doi:10.6342/NTU202000181

Table 7 Group statistics by TTS length

		Count of sentences	Mean TTS	Correct (CU score = 1)		Incorrect (CU score = 0)	
				Count	Percentage of Group	Count	Percentage of Group
TTS Group	Short	82	2.2	63	76.8%	19	23.2%
	Middle	83	3.6	36	43.4%	47	56.6%
	Long	82	6.5	12	14.6%	70	85.4%

ANOVA and Chi-square tests were conducted to examine whether the mean TTS of the three groups are statistically different and whether sentences in the three groups had significantly different likelihoods of having a correctly rendered critical unit.

First, an ANOVA test was used to ensure that the mean TTS of the three groups are truly different. Table 8 below shows that the differences of mean TTS between the three groups are all highly significant ($p < .001$), suggesting that the three groups do differ in their mean TTS length.

Table 8 Results of the formal study – ANOVA (TTS between 3 groups)

ANOVA

TTS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	759.722	1	759.722	277.011	.000*
Within Groups	441.554	161	2.743		
Total	1201.277	162			

doi:10.6342/NTU202000181

Next, Chi-square tests were conducted two groups a time to compare the groups' likelihoods of correctly rendering the critical unit (i.e. receiving 1 for their critical unit score). The groups compared each time are: the Short and Middle groups (Table 9), the Short and Long groups (Table 10), and the Middle and Long group (Table 11). The results are all highly significant ($p < .001$), inferring that the three groups have significantly different likelihoods of correctly rendering the critical unit. On this account, the inverse relationship between mean TTS and percentage of correctly rendered critical units between the three groups shown in Table 9 is confirmed to be statistically powerful. TTS is thus negatively correlated with the likelihood of rendering the critical unit correctly.

Table 9 Chi-square ("Short" vs. "Middle")

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	19.237[a]	1	.000*		
N of Valid Cases	165				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 32.80.

doi:10.6342/NTU202000181

Table 10 Chi-square ("Short" vs. "Long")

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	63.905[a]	1	.000*		
N of Valid Cases	164				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 37.50.

Table 11 Results of the formal study - Chi-square ("Middle" vs. "Long")

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	16.516[a]	1	.000*		
N of Valid Cases	165				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 23.85.

After establishing the negative correlation that sentences with shorter TTS have a higher likelihood of having a high critical unit score, the research hypothesis is examined from the opposite direction – whether sentences with high critical unit score have shorter TTS. An independent samples *t*-test was conducted to explore whether the TTS of sentence sets with a critical unit score of 1 and those with a critical unit score of 0 are significantly different.

The 247 sentences are first divided into two groups according to their critical unit score. Table 12 provides a summary of the group statistics. The low score group (CU score = 0) had a longer mean TTS (5.3 seconds), and the high score group (CU score = 1) had a shorter

doi:10.6342/NTU202000181

mean TTS (3.9 seconds), which seems to support the research hypothesis that TTS and quality scores have a negative relationship.

Table 12 Group statistics (by critical unit score)

	CU Score	Sentence Count	Mean TTS
Low Score Group	0	136	5.3
High Score Group	1	110	3.9

An independent samples test was conducted to examine whether the mean TTS of the two groups are significantly different. Table 13 shows the results of the independent samples test. The mean TTS of the two groups are significantly different, indicating that the inverse relationship between critical unit score and TTS length shown in Table 12 is highly significant ($p < .001$). Sentence sets with incorrectly rendered critical units have longer TTS, and those with correctly rendered critical units have shorter TTS.

doi:10.6342/NTU202000181

Table 13 Independent samples test for comparing TTS (grouped by critical unit score)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
TTS	Equal variances assumed	13.297	.000	-7.037	244	.000	-1.8625
	Equal variances not assumed			-7.391	227.419	.000*	-1.8625

The two-way analyses conducted above clearly illustrates the inverse relationship between TTS and the critical unit score. The Chi-square analyses (Table 9, 10 & 11) showed that sentences with shorter TTS have a higher percentage of correctly rendered critical units while sentences with longer TTS have a lower percentage. The independent samples test analyses show that sentences that scored high (receive 1 for their critical unit score) have shorter TTS, and sentences with low critical unit score (receive 0 for their critical unit) have longer TTS. This corroborates the research hypothesis that TTS and quality are negatively correlated.

4.1.2 TTS and the 0-7th second score

Table 14 below summarizes the distribution of 0-7th second score. The mean 0-7th second score is 5.1, and the median is 6.0. The range of 0-7th second score is 10, with the lowest scoring sentence set obtaining 0, and the highest scoring sentence set obtaining 10. The

doi:10.6342/NTU202000181

overall distribution of the 0-7th second score is negatively skewed, suggesting that the majority of the participants scored higher than the mean, but the rest scored notably lower than the mean. The 25th percentile is 0, and the 75th percentile is 8.5. Figure 13 illustrates the distribution of the quality scores for 0-7th second, with the data points scoring 0 excluded.

Table 14 Distribution of the 0-7th second score

0-7th second score							
Count	Mean	Median	Minimum	Percentile 25	Percentile 75	Maximum	Standard Error of Mean
247	5.1	6.0	0.0	0.0	8.5	10.0	0.2

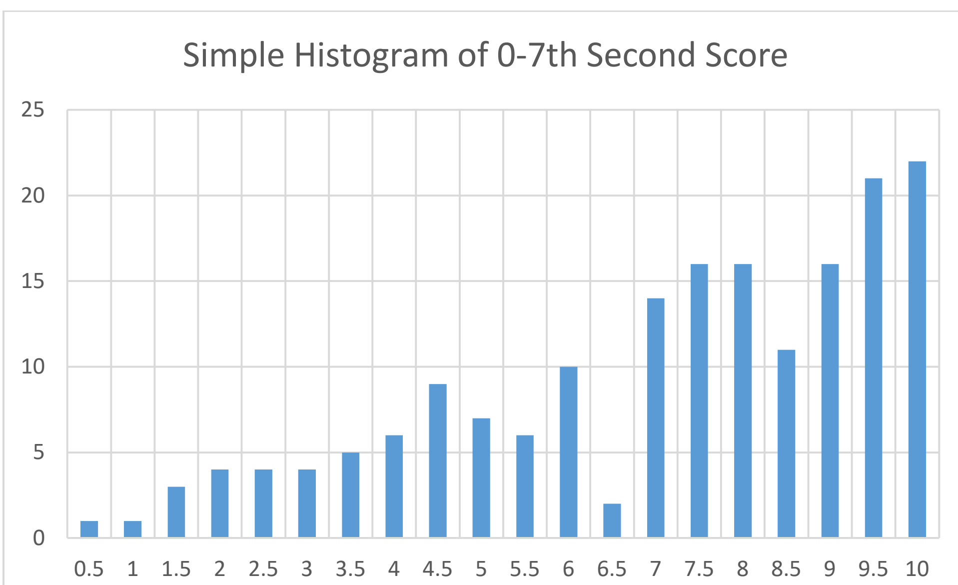


Figure 13 Simple histogram of 0-7th second score (excluding 0)

doi:10.6342/NTU202000181

A regression analysis is conducted to examine the relationship between TTS and the quality score for 0-7th second. The results are presented in Table 15 below. The results of the regression test show that TTS and the 0-7th second score are negatively correlated. The correlation between TTS and the 0-7th second score is highly significant ($p < .001$), and the coefficient between TTS and the 0-7th second score is -6.05, which is a large effect size.

Table 15 Regression analysis (TTS and 0-7th second score)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	9.269	.403		22.996	.000
	TTS	-1.022	.086	-.605	-11.854	.000*

a. Dependent Variable: 0-7th second

Figure 15 below is a scatter plot of TTS and 0-7th second score. The steep negative fit line shows the strongly negative correlation between TTS and the 0-7th second score. From the results obtained here, it can be inferred that longer TTS tend to have lower 0-7th second scores and shorter TTS tend to have higher 0-7th second score, which is in line with the research hypothesis.

doi:10.6342/NTU202000181

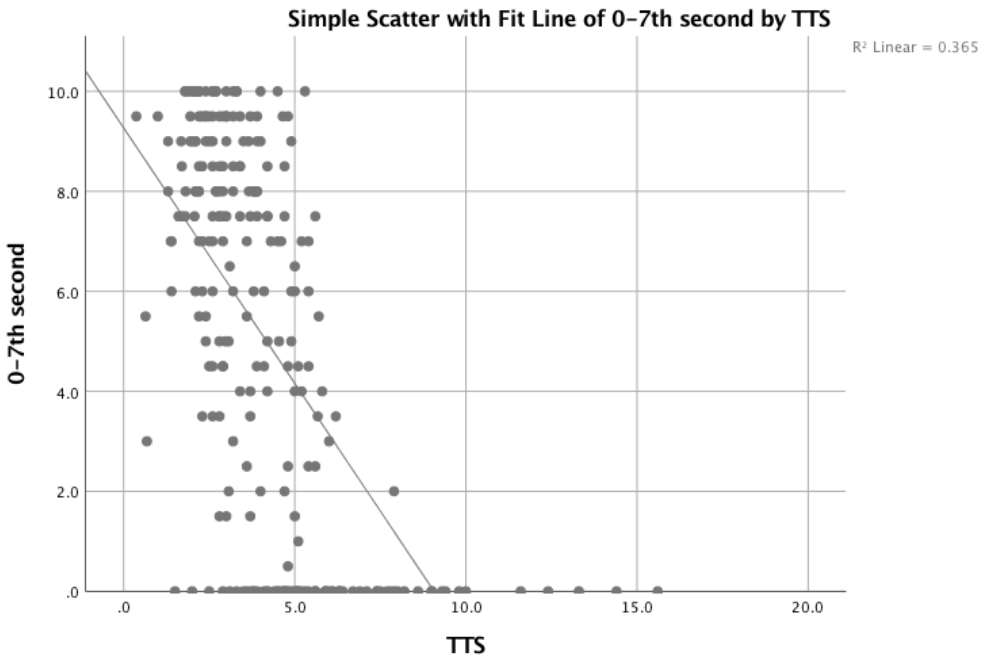


Figure 14 Scatter plot and fit line (TTS vs. 0-7th second)

4.1.2 TTS and the quality score for 8-14th second

Table 16 is a summary of the results for the quality score for 8-14th second. The mean quality score for 8-14th second is 6.6, and the median quality score for 8-14th second is 7.0. The range of 0-7th second score is 10, with the lowest scoring sentence set obtaining 0, and the highest scoring sentence set obtaining 10. The overall distribution of the quality score for 8-14th second is negatively skewed, suggesting that while most participants scored higher than the mean, some participants' scored much lower than the mean. The 25th percentile was 5.0, and the 75th percentile was 8.5. Figure 14 below illustrates the distribution of the 8-14th second quality score.

Table 16 Distribution of the quality score for 8-14th second

8-14th second							
Count	Mean	Median	Minimum	Percentile 25	Percentile 75	Maximum	Standard Error of Mean
247	6.6	7.0	0.0	5.0	8.5	10.0	0.2

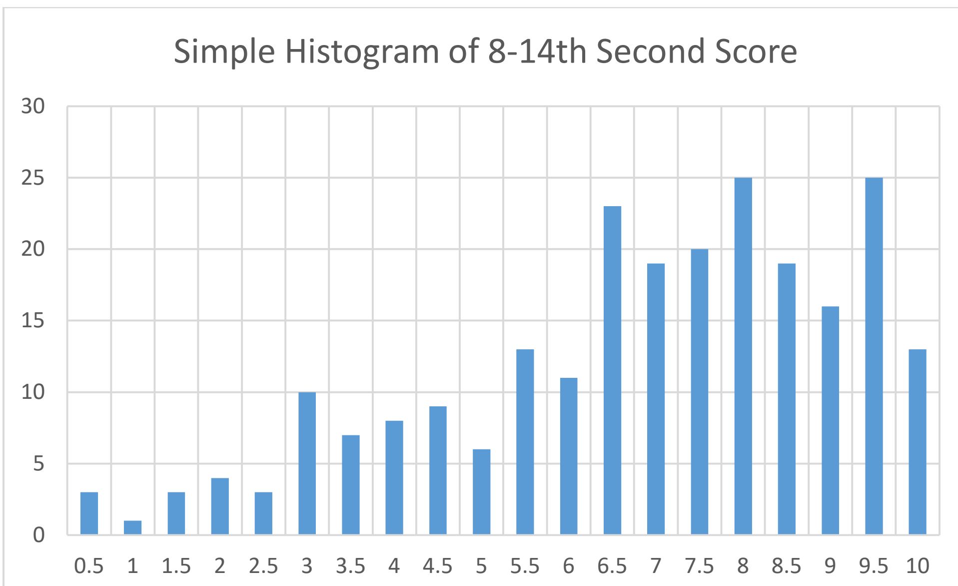


Figure 15 Simple histogram of the 8-14th second score

The mean, median and 25th percentile scores for the 8-14th second score are all higher than their respective counterparts in the 0-7th second score, which supports the recovery effect hypothesis. Table 17 below compares the 25th, 50th (mean) and 75th percentile of the two quality scores.

doi:10.6342/NTU202000181

Table 17 Overview of the 25th, 50th and 75th percentiles of the 0-7th and 8-14th second scores

	0-7th second	8-14th second
25th percentile	0	5
Mean	6	7
75th percentile	8.5	8.5

It is noteworthy that while the 75th percentile for both interval scores are 8.5, the 25th percentile score in the 0-7th second score is much lower than that in the 8-14th second score. In the quality score for 0-7th second, the 25th percentile is 0, meaning that 25% of the sentence sets obtained 0 points for their rendition for the 0-7th second score. However, in the 8-14th second score, the 25th percentile is 5 points. This suggests that even though participants still experienced substantial difficulties in the 8-14th second interval, participants were generally able to produce ~~usable~~ renditions of acceptable quality, whereas in the 0-7th second interval, at least 25% of the sentence sets did not have a rendition of acceptable quality for the critical sentence. In addition to the fact that the quality of rendition is better in the 8-14th second than in the 0-7th second, performance is also much more stable in the 8-14th second, judging from the narrower range of score distribution of the 8-14th second score.

A regression analysis is conducted to examine the relationship between TTS and the quality score for the 8-14th second. The results are shown in Table 18. TTS and the 8-14th second score demonstrate a negative correlation. The correlation between TTS and the quality score for 8-14th second is highly significant ($p < 0.001$). In addition, the coefficient between TTS and the quality score for 8-14th second is -.441, which is a medium effect size. This effect size, though still a substantial influence on the 8-14th second score, is much smaller than the effect size that TTS has on the 0-7th second score (-.605). This also corroborates the recovery

doi:10.6342/NTU202000181

effect hypothesis. However, it should be noted that although performance recovers, TTS still has an adverse effect on the performance in the 8-14th second.

Table 18 Regression (TTS vs. 8-14th second)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	8.661	.309		27.991	.000
	TTS	-.508	.066	-.441	-7.680	.000*

a. Dependent Variable: 8-14th second

Figure 16 is a simple scatter plot showing the negative correlation between the two variables. The fit line is flatter compared to that of the 0-7th score. In other words, the same unit of increase in TTS length would cause a stronger negative effect on quality in the 0-7th second than in the 8-14th second.

doi:10.6342/NTU202000181

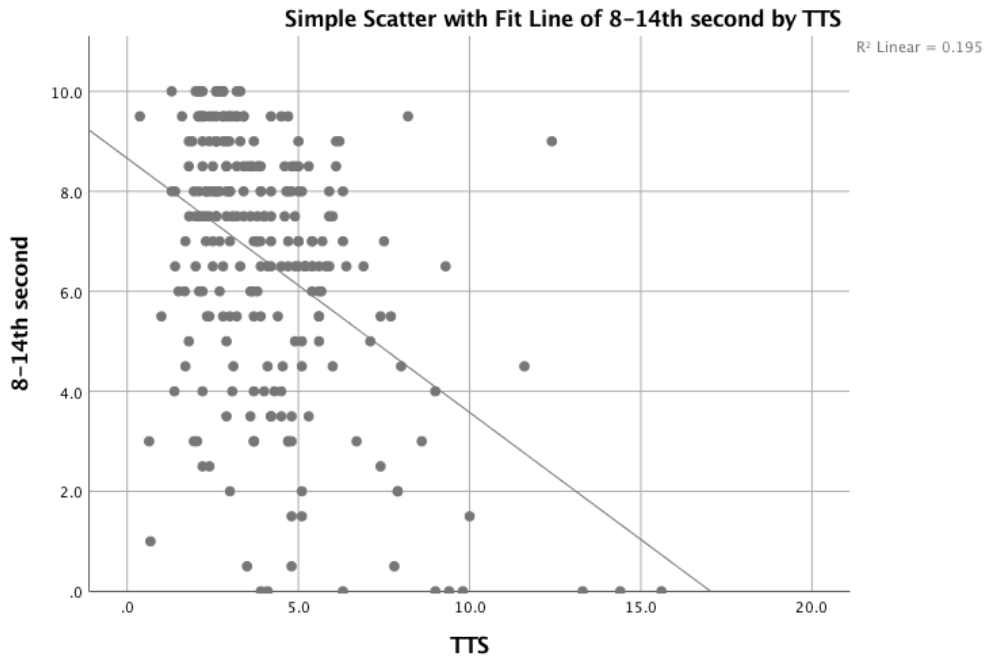


Figure 16 Scatter plot and fit line (TTS vs. 8-14th second score)

4.1.4 Summary of the general analysis

The results of the general analysis support the research hypothesis - a negative correlation exists between TTS and the critical unit score, TTS and the 0-7th second score and TTS and the 8-14th second score. In addition, a recovery effect can be observed from the effect size of TTS on the 0-7th second score and the 8-14th second score. It was not possible to conduct a similar comparison for the critical unit score and the two interval scores, because the critical unit score was assessed and scaled differently from the interval scores.

4.2 Inter-participant Analyses

The tests in the general analysis show that the overall data distribution is in line with the research hypothesis. This section moves on to examine whether the same conclusion can be

doi:10.6342/NTU202000181

applied to inter-participant comparisons, or whether participants with shorter TTS will have higher quality scores and those with longer TTS will have lower quality scores.

The main purpose of conducting an inter-participant analysis is to understand if idiosyncratic differences in TTS length is associated with interpreting performance. While the general analysis provides that sentence sets with long TTS have poorer quality scores and sentence sets with short TTS have better quality scores, the general analysis alone does not shed light on the possible causes of varying TTS. Looking into individual participant's pattern of TTS length and the associated performance helps establish or eliminate interpreter's personal style as a contributor to TTS length. If participants that have shorter TTS consistently outperform participants with longer TTS, then TTS length can be classed as a crucial element of personal style that affects interpreting performance. While personal style varies greatly across interpreters and one single element of personal style can hardly act as the sole predictor of one's interpreting performance, it is nevertheless worth using this element as a warning or indicator for performance deterioration. In practice, interpreter trainees, for example, might be warned against developing the habit of having long TTS. In addition, if the tests indicate that an interpreter's TTS length is a good predictor of his/her output quality, then it would be valuable to explore the strategies adopted by interpreters with shorter TTS. On the other hand, were this study to exclude inter-participant analysis, the general analysis alone would cripple this study's ability to explore the causes of varying TTS. The negative correlation between TTS and quality scores might be attributed to causes innate to the source speech such as delivery speed, which has been found to be a factor contributing to the variation of TTS length (Lee, 2003). Thereupon, this study proceeds to test the research hypothesis in an inter-participant level.

doi:10.6342/NTU202000181

Each participant's mean TTS and mean quality scores are calculated and shown in Table 19. It is to be noted that the "correct CU percentage" refers to the percentage of sentence sets with a correctly rendered critical unit, instead of an actual average score.

Table 19 Summary of inter-participant performance.

Participant	Mean TTS	Mean quality scores		
		0-7th second	8-14th second	Correct CU percentage
1	2.9	6.8	7.5	0.58
2	3.3	7.2	7.5	0.53
3	4.0	3.0	5.0	0.07
4	2.8	7.7	7.4	0.68
5	4.1	5.4	7.6	0.47
6	5.6	1.4	5.3	0.13
7	4.3	7.8	7.5	0.60
8	2.9	6.8	8.2	0.58
9	4.7	2.1	4.1	0.28
10	4.2	3.3	7.0	0.28
11	7.2	1.4	5.3	0.33
12	6.2	2.3	4.4	0.35
13	2.7	8.2	7.6	0.79
14	3.4	7.0	7.1	0.50

The mean of all participants' mean TTS is 4.1 seconds and the median mean TTS is 4 seconds. The mean TTS ranges from 2.7 seconds to 7.2 seconds. The mean of mean quality score for

doi:10.6342/NTU202000181

0-7th second is 5 and the median is 6.1. The mean quality scores for 0-7th second range from 1.4 to 8.2. The mean of mean quality score for 8-14th second is 6.5 and the median is 7.3. The mean quality scores for 8-14th second range from 4.1 to 8.2. The mean average critical unit score is 0.44 and the median is 0.49. The average critical unit scores range from 0.07 to 0.79.

Participants with short TTS tend to have higher average quality scores. Participant #13 had the shortest TTS (2.7 seconds), and the highest average critical unit score (0.79), as well as the highest quality score for the 0-7th second (8.2). Participant #4 had the second shortest TTS (2.8 seconds), and scored the third highest in the quality score of 0-7th second and the second highest average critical unit scoring rate. Participant #8 and participant #1 both had the third shortest TTS (2.9 seconds); participant #8 had the highest score for the interval of 8-14th second, and participant 1's performance for all three quality scores were above average.

On the other hand, participants with long TTS also had low average quality scores. Participant #11 had the longest TTS, and had the lowest score for 0-7th second. Participant #12 had the second longest TTS, and had the second lowest score for 0-7th second and 8-14th second. Participant 6 had the third longest TTS, and shares the lowest score for 0-7th second along with participant 11, and the second lowest score for average critical unit scoring.

The observations from the overview of participant performance appears to agree with the research hypothesis that TTS and quality scores are negatively correlated. The following sections will attempt to provide statistical proof for the observations. Except for the test for TTS and critical unit score, the remaining statistical methodologies used for the inter-participant analysis are in principal similar to those used for the general analysis. However,

doi:10.6342/NTU202000181

different tests were adopted to suit the small sample size of the inter-participant analysis (N = 14).

4.2.1 TTS and critical unit score

The correlation between participants' mean TTS and average critical unit score is examined first by a Kendall's Tau test, which was not applicable to the overall analysis due to the limitation of the critical unit score's data type. However, the same statistical tests used in 4.1.3 for examining the relationship between the two variables is replicated, so that the results for inter-participant analysis and that for the general analysis can be compared without interference from differences in statistical methodologies.

Table 20 shows the results of the Kendall's tau test. The results show that mean TTS and average critical unit score are negatively correlated ($p < .01$). The correlation coefficient is -.629 (marked in double asterisk in Table 20), which is a large effect size. These results suggest that participants with shorter TTS have a higher mean average critical unit score, or in other words, a higher likelihood of rendering the critical units correctly. This is in line with the research hypothesis that TTS and interpreting quality are negatively related.

doi:10.6342/NTU202000181

Table 20 Inter-participant Kendall's tau correlation (TTS vs. average critical unit score)

Correlations

			Mean TTS	Mean Average CU Score
Kendall's tau_b	Mean TTS	Correlation Coefficient	1.000	-.629**
		Sig. (1-tailed)	.	.001*
		N	14	14
	Mean Average CU Score	Correlation Coefficient	-.629**	1.000
		Sig. (1-tailed)	.001	.
		N	14	14

**. Correlation is significant at the 0.01 level (1-tailed).

4.2.2 TTS and the quality score for 0-7th second

In order to examine whether the negative correlation between TTS and the quality score for 0-7th second established in the general analysis also holds true for inter-participant analysis, a Kendall's tau test, which accommodates small sample sizes better than regression tests do, was conducted on participants' mean TTS and mean score for 0-7th second. The results are shown in Table 21 below. The results show that there is a significant correlation between the mean TTS and mean quality score for 0-7th second ($p < .05$). The coefficient is -.697, which is a large effect size. This suggests that participants with shorter TTS are likely to have a higher score for 0-7th second, and participants with longer TTS are likely to have a lower score for 0-7th second.

doi:10.6342/NTU202000181

Table 21 Inter-participant Kendall's tau correlation (TTS vs. 0-7th second)

Correlations

			Mean TTS	Mean 0-7th second
Kendall's tau_b	Mean TTS	Correlation Coefficient	1.000	-.697**
		Sig. (1-tailed)	.	.000*
		N	14	14
	Mean 0-7th second	Correlation Coefficient	-.697**	1.000
		Sig. (1-tailed)	.000	.
		N	14	14

**. Correlation is significant at the 0.01 level (1-tailed).

Figure 17 below is a scatter plot of participant's mean TTS and quality score for 0-7th second. The dots cluster very closely to the fit line, suggesting that a participant's mean TTS is a strong predictor for his/her mean 0-7th second score.

doi:10.6342/NTU202000181

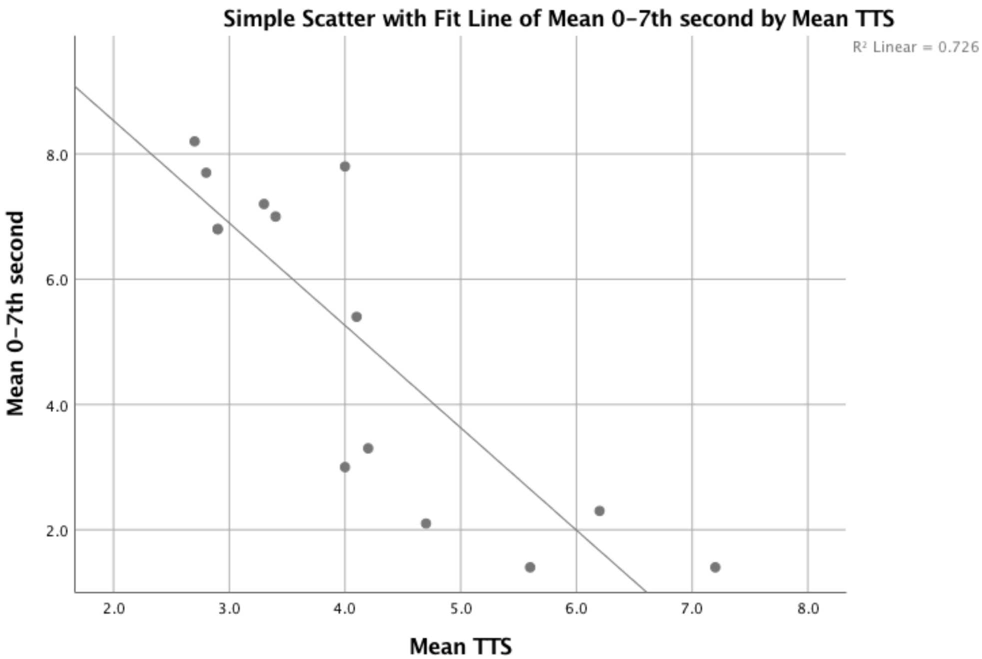


Figure 17 Inter-participant scatter plot and fit line (TTS vs. 0-7th second)

4.2.3 TTS and the quality score for 8-14th second

For the relationship between participants' mean TTS and mean quality score for 8-14th second, the Kendall's Tau correlation test used in 4.2.1 is adopted again. The results are shown in Table 22 below and Figure18 below.

doi:10.6342/NTU202000181

Table 22 Results of the formal study – inter-participant Kendall's tau correlation (TTS vs. 8-14th second)

Correlations

			Mean TTS	Mean 8-14th second
Kendall's tau_b	Mean TTS	Correlation Coefficient	1.000	-.526**
		Sig. (1-tailed)	.	.005
		N	14	14
	Mean 8-14th second	Correlation Coefficient	-.526**	1.000
		Sig. (1-tailed)	.005	.
		N	14	14

**. Correlation is significant at the 0.01 level (1-tailed).

The results show a significant relationship between the mean TTS and mean quality score for 8-14th second ($p < .05$). The coefficient is -.526, which is a large effect size. This suggests that participants with shorter TTS are likely to have a higher score for 8-14th second, and participants with longer TTS are likely to have a lower score for 8-14th second. The recovery effect found in the general analysis is also observed in the inter-participant analysis. The effect size TTS has on 0-7th second in the inter-participant analysis is also stronger than that on 8-14th second.

doi:10.6342/NTU202000181

Figure 18 below is a simple scatter plot with fit line of participant's mean TTS and the mean quality score for 8-14th second. The dots scatter less clustered around the fit line compared to the scatter plot for mean TTS and the mean quality score for 0-7th second.

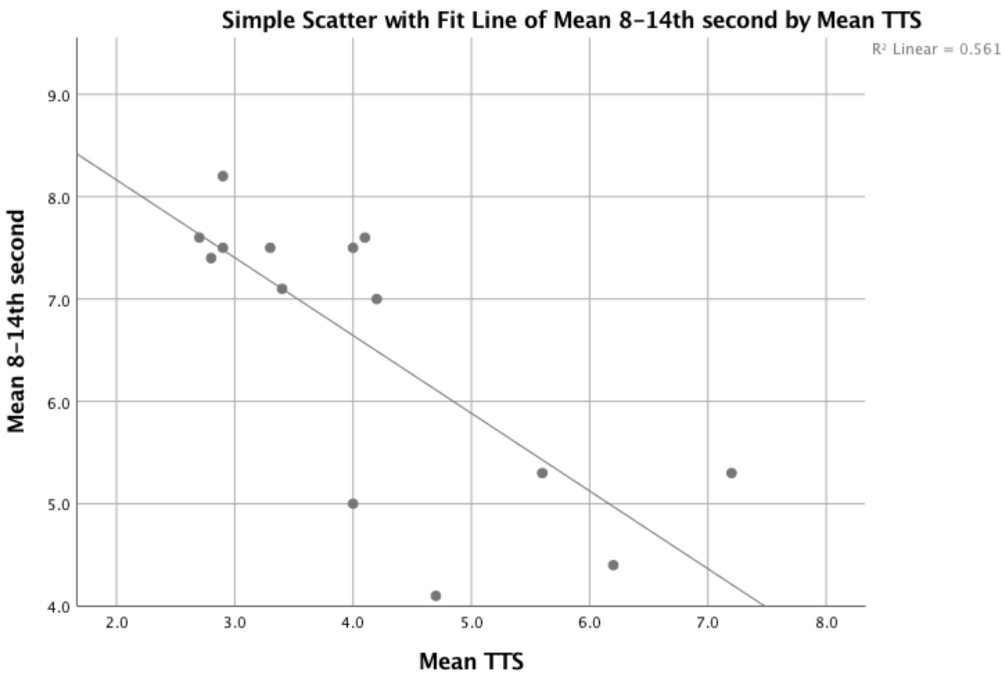


Figure 18 Inter-participant scatter plot and fit line (TTS vs. 8-14th second)

4.2.4 Summary of Inter-participant Analysis

The results of inter-participant analysis of all three quality scores also support the research hypothesis that TTS and quality are negatively correlated. In other words, participants who have shorter TTS tend to obtain higher quality scores. The significance of the negative correlations also tends to occur with the same participants. That is, some participants are far more prone to the effects of varying TTS lengths.

doi:10.6342/NTU202000181

4.3 Intra-participant Analysis

The third level of analysis is at an intra-participant level. This level of analysis looks at whether the research hypothesis also holds true within a participant's performance. The general analysis results have established that a negative correlation exists between TTS and quality scores. Results of the inter-participant analysis offer individual TTS habits as an explanation of the variation in TTS, which in turn is associated with quality score variations. The analysis at an intra-participant level further examines whether factors other than idiosyncratic style may contribute to TTS variation.

The statistical methodologies adopted in this level of analysis is largely similar to those adopted in the general analysis and inter-participant analysis, with slight differences in the actual statistical test adopted to suit the type of data.

4.3.1 TTS and critical unit score

Point bi-serial tests are conducted to examine whether the same participant's likelihood of rendering the critical units correctly is associated with the length of TTS. Table 23 lists out the results for all 14 participants. The results show that the significant negative correlation between TTS and critical unit score found in the general analysis and inter-participant analysis is present in four out of the fourteen participants (#5, #10, #12, #13) at a significance level of $p < .05$. The coefficient of these participant's point bi-serial tests all show a large effect size, suggesting that these four participants are especially prone to the effects of TTS length.

doi:10.6342/NTU202000181

Table 23 Intra-participant Kendall's tau correlation (TTS vs. critical unit score)

Participant	*p* value (one-tailed)	Coefficient
1	0.161	-0.228
2	0.072	-0.336
3	0.077	-0.368
4	0.291	0.128
5	0.002*	-0.805
6	0.484	0.011
7	0.076	-0.47
8	0.082	-0.32
9	0.096	-0.308
10	0.032*	-0.437
11	0.187	-0.491
12	0.006*	-0.599
13	0.028*	-0.437
14	0.401	-0.06

4.3.2 TTS and 0-7th second quality score

Kendall's Tau correlation test is again used to examine the correlation between TTS and the quality score for the 0-7th second. Table 24 lists out the test results for all fourteen participants. Out of the fourteen participants, the results of seven participants are significant (#4, #5, #7, #9, #11, #12, #13) at a significance level of $p < .01$ or a more significant level. It

doi:10.6342/NTU202000181

is noteworthy that three out of the four participants who had a significant negative correlation between TTS and critical unit score also had a highly significant negative correlation between TTS and the quality score for 0-7th second.

Table 24 Intra-participant Kendall's tau correlation (TTS vs. 0-7th second)

Participant	*p* value (one-tailed)	Coefficient
1	0.444	-0.024
2	0.089	-0.234
3	0.081	-0.289
4	0.038*	-0.321
5	0.000*	-0.658
6	0.134	-0.228
7	0.012*	-0.455
8	0.069	-0.258
9	0.008*	-0.454
10	0.274	-0.111
11	0.012*	-0.447
12	0.000*	-0.794
13	0.003*	-0.492
14	0.323	-0.082

4.3.2 TTS and 8-14th second quality score

Kendall's Tau correlation test is again used to examine the correlation between TTS and the quality score for 8-14th second. Table 25 lists out the results for all fourteen participants. Out

doi:10.6342/NTU202000181

of all fourteen participants, the results of two participants are significant at a level of $p < .05$ (#5 and #12). There are, nevertheless, four other participants whose correlations are close to a significant value (#6, #7, #10, #11). Amongst these six participants whose results for TTS and the quality score for 0-7th second are significant, five of them also have a significant or close to significant result for the correlation between TTS and the quality score for 8-14th second. Participant #5 and participant #12 have a significant negative correlation between TTS and all three quality scores.

Table 25 Intra-participant Kendall's tau correlation (TTS vs. 8-14th second)

Participant	*p* value (one-tailed)	Coefficient
1	0.207	-0.143
2	0.135	-0.193
3	0.344	0.08
4	0.349	.067
5	0.012*	-0.419
6	0.056**	-0.298
7	0.067**	-0.296
8	0.486	-0.006
9	0.111	-0.217
10	0.057**	-0.285
11	0.057**	-0.281
12	0.012*	-0.427
13	0.284	-0.1
14	0.119	-0.208

doi:10.6342/NTU202000181

4.3.4 Summary of Intra-participant Analyses

The results of intra-participant analysis show that significant or marginally significant correlations between TTS and all the three quality scores exist in some participants' performance. Table 26 summarizes the results of intra-participant analyses.

Table 26

Participant	TTS vs. critical unit score		TTS vs. 0-7th second		TTS vs. 8-14th second	
	p value (one-tailed)	Coefficient	*p* value (one-tailed)	Coefficient	*p* value (one-tailed)	Coefficient
1	0.161	-0.228	0.444	-0.024	0.207	-0.143
2	0.072	-0.336	0.089	-0.234	0.135	-0.193
3	0.077	-0.368	0.081	-0.289	0.344	0.08
4	0.291	0.128	0.038*	-0.321	0.349	0.067
5	0.002*	-0.805	0.000*	-0.658	0.012*	-0.419
6	0.484	0.011	0.134	-0.228	0.056**	-0.298
7	0.076	-0.47	0.012*	-0.455	0.067**	-0.296
8	0.082	-0.32	0.069	-0.258	0.486	-0.006
9	0.096	-0.308	0.008*	-0.454	0.111	-0.217
10	0.032*	-0.437	0.274	-0.111	0.057**	-0.285
11	0.187	-0.491	0.012*	-0.447	0.057**	-0.281
12	0.006*	-0.599	0.000*	-0.794	0.012*	-0.427
13	0.028*	-0.437	0.003*	-0.492	0.284	-0.1
14	0.401	-0.06	0.323	-0.082	0.119	-0.208

Significance occurred most frequently in the correlation between TTS and the quality score for 0-7th second; followed by that between TTS and critical unit score; that between TTS and 8-14th second occurred most infrequent. It is notable that the significance or marginal significance occurred almost always with the same group of participants (#5, #10, #11, and

doi:10.6342/NTU202000181

#12). This may imply that some participants are more prone to the effects of varying TTS lengths. The possible reasons are discussed in the following chapter.

doi:10.6342/NTU202000181

V. Discussion

The results gathered from the experiment and the statistical analyses from the previous chapter confirm the research hypothesis that TTS has an adverse effect on the output quality in SI. The effect is present in all three levels of analysis - the general analysis, inter-participant analysis and some of the participant's intra-participant analysis. This chapter provides qualitative illustration of the impact of TTS on output quality on a local level – within the scope of two or three adjacent sentences or idea units, and also probes into the factors that may exert an influence on the length of TTS.

5.1 Comparing the results of the present study with Lee's 2003 study

The results found in the present study corroborates Lee's findings that TTS length and accuracy are negatively correlated(Lee, 2003), while providing even further evidence for the negative correlation between TTS and output quality through incorporating an extra dimension "fluency" into quality judgment. The results of this study also expands the applicability of the TTS-quality correlation from the English-Korean language pair to the English-Chinese language pair, and provides ground for future studies to examine the relationship between TTS and quality in more language pairs. In addition, this study's results together with Lee's results suggest that a target language's relative syllabic-verbose nature does not necessarily mean longer TTS. In Chapter II, Korean was identified as a syllabic-verbose language as opposed to English. When performing SI from English into Korean, interpreters often had to use more syllables than the English source text to express the idea in Korean. Chinese on the other hand is relatively concise and compact when compared to English. This difference between English and Chinese yields room to hypothesize that

doi:10.6342/NTU202000181

English to Korean SI may have a longer TTS than English to Chinese SI. However, the mean TTS found in the present study, 4.1 seconds, is almost identical to the 4.11-second mean TTS found in Lee's study. The high proximity of the mean TTS between the two studies suggest that TTS length is influenced by more than just the syllable count of the target speech. A possible explanation is that the length of TTS has more to do with the amount of information contained within approximately 4.1 seconds of the source speech, since both studies have English as the source language of the experiment. It can be postulated that the information load contained in 4.1 seconds plus or minus a small range is the average information load interpreters feel comfortable to retain in working memory. This explanation, however, cannot be proved through comparing the results of the present study and Lee's study because different materials were used as the source speech, and there is no way to examine whether 4.1 seconds in Lee's source speeches contains a similar information load with 4.1 seconds of source speech in the present study.

5.2 The effects of TTS on SI quality

The overall result of this study suggests that TTS has a significant negative correlation with all three quality scores at a level of $p < .001$. Table 27 summarizes the results of the general analysis.

Table 27 Summary of general analysis

	P (one-tailed)	**Coefficient**
TTS vs. Avg. CU	.000	N/A
TTS vs. 0-7th Second	.000	-.605
TTS vs. 8-14th Second	.000	-.441

doi:10.6342/NTU202000181

For the scores for the 0-7th second and the 8-14th second, the effect of TTS is larger on the quality of the 0-7th second (r = -.605) than the effect on the quality of the 8-14th second (r= -.441). This decrease in effect size implies that participants recover gradually from the effects of the imported cognitive load from the previous sentence. The overall mean score for the data also supports this implication. The mean score, median and 25th percentile of the quality score for the 8-14th second are consistently higher than that for the 0-7th second, as illustrated in Table 28 below.

Table 28 Mean, median and 25th percentile scores for 0-7th second interval & 8-14th setntence intervals

	0-7th	8-14th
Mean	5.1	6.6
Median	6.0	7.0
25th percentile	0	5.0

This suggests that the negative effect of long TTS diminishes gradually as the interpreter is relieved from the pressure of having to finish up the imported load from the trigger sentence while listening to the critical sentence and retaining the message in memory.

Another factor that might contribute to this result is that at the start of a sentence interpreters could still be gauging the direction of the sentence, and thus pause, stall or backtrack more

doi:10.6342/NTU202000181

frequently, which affects the perception of quality. Take the below instance as an example (participant #8, sentence #13):

ST: The founder had run away from home at a young age. He went to India to educate himself, and came back to Nepal.

TT: 那這個創辦人呢其…呃其實呢翹家呢跑到印度自學後來又回到尼泊爾然後創立了一個孤兒院。

(Translation: This founder act[ually]… um actually ran away from home to educate himself in India and then came back to Nepal and then started an orphanage.)

(0-7th second score: 4.5; 8-14th second score: 8.5)

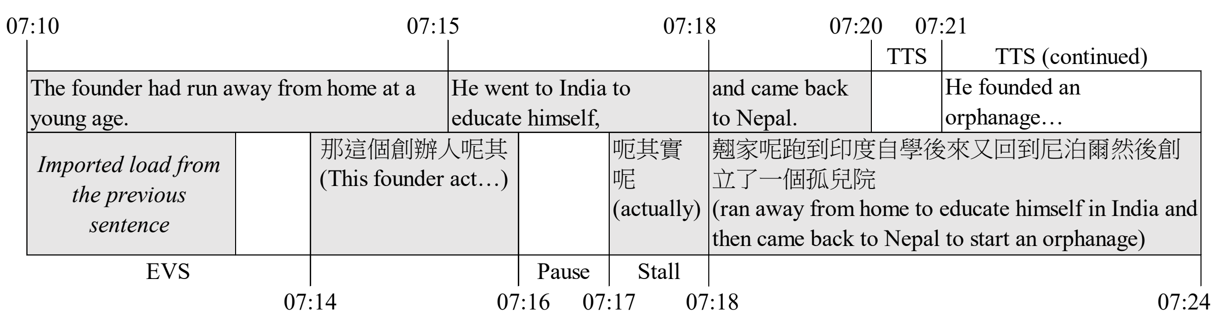


Figure 19 Temporal relationship between ST & TT for sentence #13 of participant #8

In this instance, the participant began his/her utterance with “那這個創辦人其” (This founder act[ually]…), and then paused for one second, then stalled with “呃其實呢” (um actually) for one more second, and finally translated the rest of the source speech. By the time the participant began the stalling utterance (“um actually”), the information following

doi:10.6342/NTU202000181

"this founder" had already become available in the source speech ("ran away from home at a young age"). However, the participant stalled a little more, and then produced a rendition that incorporated "ran away from home" with the next segment "educate himself in India…" It is likely that the participant wanted to be more certain about where the source speech is going, and therefore paused and stalled to better comprehend the source speech, and then reformulated the comprehended information all at once.

Another possible explanation is that the results were affected by the sentence sets that scored poorly in 0-7th second because the participant's TTS or pause took up the entire 0-7th second interval, and thus received 0 for this interval score. A statistical test was thus conducted to examine whether the significance of the relationship between the TTS and the quality score for 0-7th second would change if those sentences were removed and the results showed that the relationship remains significant.

The distribution of TTS collected from the experiment shows considerable deviations from the mean TTS span. The shortest TTS recorded is -0.2 seconds, which means that the interpreter finished the sentence earlier than the speaker did.

A -0.2 second TTS means that in this instance the interpreter finished earlier than the speaker. The short TTS was followed by an accurate and fluent rendition for the 0-7th second that received a quality score of 9.5 out of 10 points. However, a closer look into the participant's performance reveals that the participant failed to generate any output for the final two clauses in the trigger sentence. It was because of this omission that this participant was able to have such a short TTS. The trigger sentence and a transcript of its target speech is provided below, with the chunking of the trigger sentence marked with the "//" sign:

doi:10.6342/NTU202000181

ST: I encountered a mother and her 9-year-old son//, who looked a bit bored, //while I was waiting for the elevator.

TT: 那時候呢，我有遇到一對母子。
(Translation: At that time, I encountered a mother and her son.)

The rendition of the trigger sentence only accounts for the idea unit "*I encountered a mother and her 9-year-old son*". The remainder of the trigger sentence was not rendered. Given this apparent omission, the TTS-quality relationship in this sample should be examined in a finer scope. Lee (2003) has found in his research that source speech sentence length (measured by the speaker's syllables) is positively correlated with TTS length, which means the longer the source speech sentence is, the longer the TTS length. In this sample, the portion of source speech that was rendered by the participant is less than half the length of the entire trigger sentence (13 syllables vs. 29 syllables). According to Lee's findings, the TTS for interpreting the first idea unit (up to "son") should be shorter than for interpreting the entire sentence (up to "elevator"), so interpreting the first unit only should yield a TTS that is shorter than the TTS for interpreting the entire sentence. However, while the other thirteen participants had an average of 3.2 second TTS for interpreting the entire sentence, this participant had a 3.4 second for interpreting only up to "母子" (mother and son), while the source speech ended with "9-year-old-son." On the other hand, the average TTS of the other thirteen participants for the entire trigger sentence was only 3.2 seconds. This suggests that a failure to interpret the latter two idea units of the trigger sentence could be caused by an excessively long TTS for the first idea unit. The 3.4-second lag coincides almost entirely with the latter two idea units. Since the experiment did not include a retrospective interview

doi:10.6342/NTU202000181

to discuss with the participants their performance, the reason for the omission is unclear. It is possible that the participant had comprehended the second and third idea units but decided to discard those units because (s)he regarded the units as less relevant, or (s)he sensed that lagging further behind could cause more harm to the output. It is also possible that the participant's comprehension of the second and third idea units were simply not sufficient for generating an output. In either scenario, the omission shortened the lag for the entire trigger sentence, which likely allowed the participant to focus on the message of the critical sentence and contributed to the participant's outstanding performance in the 0-7th second quality score.

The longest TTS recorded was 15.6 seconds from participant #11's target speech of trigger sentence #9. This participant's TTS is consistently longer than other participants. The participant had a mean TTS of 7.2 seconds, the longest mean TTS amongst all participants (mean TTS of the other 13 participants ranged from 2.9 to 6.2 seconds). Even for a participant with a long-TTS tendency, 15.6 seconds is still longer than twice the mean TTS of this participant. The TTS for this sample had reached this unusual length because the participant appeared to have comprehension difficulties when (s)he was interpreting the two sentences preceding the trigger sentence. The participant made frequent pauses that added up to 8.1 seconds over those two sentences, and also repeated an entire sentence in his/her target speech, costing an extra 3.5 seconds. As a result, the EVS for the trigger sentence amounted to 13.7 seconds.

(The two sentences preceding the trigger sentence)

doi:10.6342/NTU202000181

ST: Once the school meeting had made a decision that games could not be played in the computer room in the morning. So they passed the computer out the window, so they could play computer games.

TT: 他們學校有一次 [0.5 seconds]決定 [0.5 seconds]早上在 [1.5 seconds]電腦教室不可以玩遊戲 [1.8 seconds]所以小孩子就想出了一個方法 [1.8 seconds]繞過這樣子的規定 [2 seconds]嗯，就想出一個方法，來執行這樣子的規定。

(Translation: Their school once [0.5 seconds] made a decision that, in the morning [0.5 seconds], in the computer room, games could not be played [1.8 seconds]. So the kids came up with a way, to circumvent this rule [2 seconds], um, came up with a way, to enforce this rule.)

After this long lag, the participant still managed to produce a brief interpretation for the trigger sentence, suggesting that (s)he was able to divert attention to comprehending new incoming source speech while generating the interpretation for the trigger sentence.

(The trigger sentence)

ST: This may sound funny. But these students were able to follow the decision they made themselves, even when it means less game time.

TT: 所以即使這樣子的規定會減少他們玩遊戲的時間，他們還是願意執、執行這樣子的規定。

doi:10.6342/NTU202000181

(Translation: So even though a rule like this would reduce their game time, they were still willing to enforce it, enforce a rule like this.)

The two instances above unveil possible ways that TTS length affects the output quality in SI. While the relationship between TTS and quality score of the two examples above are in line with the research hypothesis, the two examples demonstrate that the TTS-quality relationship should not be taken at face value. Short TTS and good performance of one segment could be the result of long TTS and omissions in the previous segment.

5.3 Factors influencing TTS length

Since TTS plays a part in influencing SI quality, it is worth exploring in further details what influences the length of TTS. Based on inferences on temporal causal-relationships and findings of earlier studies, a couple of factors may contribute to the variance of TTS length. These factors are: (a) TTS of the previous sentence (imported TTS), (b) EVS, (c) speech rate, and (d) succinctness of the interpreter's rendition. The TTS of the previous sentence and the current sentence's EVS determine when interpreters start their target speech sentence. The longer this initial lag is, the longer TTS tends to be (Lee, 2003). The latter two, speech rate (with frequency and duration of pauses included) and succinctness determine how much time interpreters need to complete conveying the message of the source speech. Factor (a), (b), (c) and (d) at play together determine the length of TTS of a sentence. If shorter TTS is correlated with better quality and the four factors are associated with the length of TTS, it is worth using the samples collected in this study to explore and illustrate how the aforementioned factors influence the length of TTS.

doi:10.6342/NTU202000181

5.3.1 Imported TTS & EVS

The lag of any given sentence or idea unit in SI is influenced by the imported TTS of the previous sentence and the EVS of the current sentence, except for the initial sentence of the entire discourse, the lag of which is only subject to the length of EVS. Therefore, for the majority of sentences in an SI discourse, the length of the imported TTS directly adds to the length of EVS, assuming that the speaker does not suddenly makes a pause longer than the interpreter's TTS.

Table 29 shows the mean ear-voice span of all participants. The participants with the shortest tail-to-tail span and ear-voice span are marked bold and underlined; the participants with the longest TTS and EVS are framed in a square.

doi:10.6342/NTU202000181

Table 29 Summary of participant's TTS and EVS

Participant	N	Mean TTS	Mean EVS
1	19	**2.9**	**2.5**
2	19	3.3	4.0
3	15	4	2.6
4	19	**2.8**	**1.7**
5	17	4.1	4.1
6	16	5.6	3.9
7	15	4.3	5.4
8	19	**2.9**	**2.0**
9	18	4.7	4.9
10	18	4.2	3.1
11	18	7.2	6.4
12	16	6.2	5.6
13	19	**2.7**	3.2
14	18	3.4	4.3

Except for participant 13, the other 3 participants (#4, #1 and #8) with the shortest tail-to-tail span happen to be the participants with the second shortest tail-to-tail span (both participant #1 and participant #8 had the same TTS).

Seeing the intertwined relationship between EVS and TTS and the clear trend between the lengths of EVS and TTS found in inter-participant analysis, the chicken-and-egg conundrum is invoked. Although temporally EVS occurs first, the lag length determined by the initial EVS can be toppled by problem triggers in the source speech or interpreters' strategy uses

doi:10.6342/NTU202000181

anytime throughout the interpreted discourse. On the other hand, imported TTS, which is the part of TTS that overlaps with the source speech of the next segment, directly adds to the length of TTS. The intertwined relationship between EVS and TTS makes it hard to say whether the effect TTS has on SI quality is in fact really from TTS alone or is the effect mainly related to the length of EVS and examinations of TTS turn out to be positive simply because TTS lengths are partly influenced by EVS.

In order to test whether the length of EVS has a stronger effect on interpreting quality than TTS does, statistical tests are conducted to examine their effects on the interpreting quality scores. A three-phased method is adopted to compare the effects of EVS and TTS. Firstly, a statistical test was conducted to examine whether EVS alone and TTS alone are significantly correlated with the three quality scores. If the results show the both EVS and TTS are significantly correlated with the quality scores, a second statistical test is carried out to check if a model using both EVS and TTS as predictors turned out to be significantly correlated with the quality scores. If the results are positive, a third test is carried out to examine whether EVS and TTS still remain significant when the other is taken into account as a contributing factor (i.e. if EVS remains significant when TTS is also taken into account, and vice versa), and how significant are they, to determine which one contributes more to the variances in quality scores.

5.3.1.1 0-7th second

Correlation test results show that both EVS and TTS are significantly related to the quality score for 0-7th second (Table 30) and 8-14th second (Table 31) at a level of $p < .000$ (for 0-7th second, $p = .000$ for both EVS and TTS; for 8-14th second, $p = .000$ for both EVS and

doi:10.6342/NTU202000181

TTS). The coefficients for both EVS and TTS are negative for both 0-7th second and 8-14th second. TTS has a stronger effect size than EVS does on both 0-7th second (TTS: -.607 vs. EVS: -.402) and 8-14th second (TTS: -.431 vs. EVS: -.322).

Table 30 Correlation between EVS & TTS at 0-7th second interval score

Correlations

		First7 score	EVS	TTS
Pearson Correlation	First7 score	1.000	-.402	-.607
	EVS	-.402	1.000	.685
	TTS	-.607	.685	1.000
Sig. (1-tailed)	First7 score	.	.000*	.000*
	EVS	.000	.	.000
	TTS	.000	.000	.
N	First7 score	247	247	247
	EVS	247	247	247
	TTS	247	247	247

doi:10.6342/NTU202000181

Table 31 Correlation between EVS & TTS at 8-14th second interval score

Correlations

		Second7 score	EVS	TTS
Pearson Correlation	Second7 score	1.000	-.322	-.431
	EVS	-.322	1.000	.685
	TTS	-.431	.685	1.000
Sig. (1-tailed)	Second7 score	.	.000*	.000*
	EVS	.000	.	.000
	TTS	.000	.000	.
N	Second7 score	247	247	247
	EVS	247	247	247
	TTS	247	247	247

The results from simple correlation tests indicate that TTS has a stronger effect on the two interval score than EVS does. However, since EVS has been found to influence the length of TTS, further tests are needed to examine how the contribution of the two lags to SI output quality compare, for which a multiple linear regression is calculated to predict the quality scores based on EVS and TTS. The results are shown in Table 32, 33 and 34 below.

Table 32 shows that the regression of the model using both TTS and EVS as predictors is highly significant at $p < .001$ ($p = .000$) and that TTS and EVS account for 36.8% of the variance in the quality score for 0-7th second (R square = .368).

doi:10.6342/NTU202000181

Table 32 Regression of the model using both TTS and EVS as predictors (0-7th second)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1321.919	2	660.959	71.127	.000[b]
	Residual	2267.401	244	9.293		
	Total	3589.320	246			

a. Dependent Variable: First7 score

b. Predictors: (Constant), TTS, EVS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.607[a]	.368	.363	3.0484	.368	71.127	2	244	.000	1.392

a. Predictors: (Constant), TTS, EVS

b. Dependent Variable: First7 score

Table 33 below shows the respective contribution of EVS and TTS. Once TTS is taken into account, EVS's contribution to the variance of the quality score for 0-7th second is no longer significant ($p = .710$). On the other hand, when EVS is taken into account, TTS's contribution remains significant ($p = .000$).

doi:10.6342/NTU202000181

Table 33 Respective contribution of EVS and TTS to SI output quality (0-7th second)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	9.224	.412		22.364	.000
	EVS	.041	.110	.026	.372	.710
	TTS	-1.050	.118	-.624	-8.938	.000*

a. Dependent Variable: First7 score

To ensure that the regression model used above is not biased by collinearity between EVS and TTS, the tolerance and VIF of the model are examined. Table 34 shows that VIF is smaller than 10 (VIF = 1.885), and tolerance is larger than .2 (Tolerance = .53). Therefore, EVS and TTS do not share a collinearity that is strong enough to bias the model.

doi:10.6342/NTU202000181

Table 34 Collinearity test results for EVS & TTS (0-7th second)

Coefficients[a]

Model		B	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound	Correlations Zero-order	Correlations Partial	Correlations Part	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	9.224	8.411	10.036					
	EVS	.041	-.175	.257	-.402	.024	.019	.530	1.885
	TTS	-1.050	-1.282	-.819	-.607	-.497	-.455	.530	1.885

a. Dependent Variable: First7 score

5.3.1.2 8-14th second

The same tests are performed for the relationship between EVS and the quality scores for 8-14th second and that between TTS and the quality score for 8-14th second.

Table 35, 36 and 37 below show the results of multiple linear regression. Table 35 shows that the regression of a model using both TTS and EVS as predictors for quality scores for 8-14th second is significant at $p < .001$ ($p = .000$). Table 36 shows that this model accounts for 18.7% of the variance in the quality score for 8-14th second (R square = .187).

doi:10.6342/NTU202000181

Table 35 Power of the regression of a model using both TVS and EVS as predictors (8-14th second)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	311.578	2	155.789	28.107	.000[b]
	Residual	1352.432	244	5.543		
	Total	1664.010	246			

a. Dependent Variable: Second7 score

b. Predictors: (Constant), TTS, EVS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.433[a]	.187	.181	2.3543	.187	28.107	2	244	.000	1.648

a. Predictors: (Constant), TTS, EVS

b. Dependent Variable: Second7 score

Table 36 below shows the respective contribution of EVS and TTS. Once TTS is taken into account, EVS's contribution to the quality score for 8-14th second is no longer significant (p = .528). On the other hand, when EVS is taken into account, TTS's contribution remains significant (p = .000).

doi:10.6342/NTU202000181

Table 36 Respective contribution of EVS and TTS to SI output quality (8-14th second)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	8.636	.319		27.112	.000
	EVS	-.054	.085	-.050	-.632	.528
	TTS	-.455	.091	-.397	-5.008	.000

a. Dependent Variable: Second7 score

Once again, to ensure that the model is not biased by the collinearity shared by the two predictors, EVS and TTS, the collinearity statistics are examined. Table 37 shows that VIF is smaller than 10 (VIF = 1.885), and tolerance is larger than 0.2 (Tolerance = .530), which shows that the model is not biased by the collinearity shared between EVS and TTS, and the conclusions obtained from the model are also unbiased.

doi:10.6342/NTU202000181

Table 37 Collinearity test results for EVS & TTS (8-14th second)

Coefficients[a]

Model		B	95.0% Confidence Interval for B Lower Bound	Upper Bound	Correlations Zero-order	Partial	Part	Collinearity Statistics Tolerance	VIF
1	(Constant)	8.636	8.008	9.263					
	EVS	-.054	-.220	.113	-.322	-.040	-.037	.530	1.885
	TTS	-.455	-.633	-.276	-.431	-.305	-.289	.530	1.885

a. Dependent Variable: Second7 score

5.3.1.3 Critical unit score

To test if the EVS better predicts a participant's correct, a logistic regression test was carried out. Table 38 and 39 below show the results.

Table 38 shows that regression using a model that has both EVS and TTS as its predictors for variance in critical unit score (scoring 0 for incorrect rendition or 1 for correct rendition) is significant, suggesting that EVS and TTS together influence the chance of rendering the critical unit correctly.

doi:10.6342/NTU202000181

Table 38 Regression analysis of how EVS and TTS together influence the likelihood of correctness of the critical unit

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	61.075	2	.000
	Block	61.075	2	.000
	Model	61.075	2	.000

Table 39 further compares the contribution from EVS and TTS. Wald statistics show that while TTS makes a significant contribution to whether participants render the critical unit correctly (Wald = 20.866, $p < .000$), EVS's contribution is not significant at all (Wald = 3.385, $p > .05$).

Table 39 Comparison of EVS' and TTS' respective contribution to the likelihood of correctness of the critical unit

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	EVS	.162	.088	3.385	1	.066	1.176	.989	1.397
	TTS	.535	.117	20.866	1	.000	1.707	1.357	2.147
	Constant	-2.402	.420	32.759	1	.000	.091		

a. Variable(s) entered on step 1: EVS, TTS.

doi:10.6342/NTU202000181

The regression tests above show that in all three quality scores (0-7th second, 8-14th second and critical unit score), TTS is a stronger predictor than EVS is.

5.3.2 Speech rate

Higher speech rate may also contribute to shorter tail-to-tail span. Table 40 lists out the TTS and overall speech rate of all participants. Overall speech rate is measured in characters per minute (cpm), calculated by dividing the total character count of the target text by the total length of the target speech. The participants with the shortest TTS and highest speech rate are marked bold and underlined.

doi:10.6342/NTU202000181

Table 40 Summary of participant's TTS and speech rate

Participant	N	Mean TTS	Speech Rate
1	19	**2.9**	**190-200**
2	19	3.3	180-190
3	15	4	140-150
4	19	**2.8**	**220-230**
5	17	4.1	170-180
6	16	5.6	120-130
7	15	4.3	180-190
8	19	**2.9**	**190-200**
9	18	4.7	140-150
10	18	4.2	180-190
11	18	7.2	160-170
12	16	6.2	150-160
13	19	**2.7**	160-170
14	18	3.4	160-170

The results show that, excluding participant #13, whose TTS is the shortest among all participants, the other 3 participants (#4, #1 and #8) with the shortest TTS have the highest speech rate amongst all participants. On the other hand, not all of the participants with the longest tail-to-tail span have low speech rate. Participant 6 has a speech rate of 120-130 characters per minute, which is the lowest amongst all participants. The other two participants with long TTS spoke at an average speech rate.

doi:10.6342/NTU202000181

It is notable that the overall speech rate ranking used here is only a reference for what may contribute to TTS. The speech rate here does not take into account the style of the language and omitted sentences. Participant 3 and participant 6, for example, have lower speech rates, but they omitted more sentence sets (N = 15 and N = 16, compared to the average N = 17.7), which indicates that their target speech as an entirety likely includes large chunks of blankness. In addition, speech rate does not reflect the interpreter's pause patterns. Some participants pause longer between sentences and start later, but produce fluent and complete target texts afterwards, whereas other participants may have shorter pauses between sentences but more fragmented pauses within sentences.

Take participant 10 and #14 as examples. Participant 10's mean ear-voice span (3.1 seconds) is shorter than that of participant #14 (4.3 seconds), and participant 10's speech rate (180-190 cpm) was also higher than that of participant #14 (160-170 cpm). Based on the EVS and speech rate measurements, participant #10 should have a shorter TTS than participant #7. However, in sentence #7, participant #10 ended up with a longer TTS because of frequent pauses and fillers within his/her target speech sentence. Participant #10 started much earlier than Participant #14. Participant #10 had an EVS of 2.5 seconds, whereas Participant #14 had an EVS of 5.8 seconds.

Trigger ST: It was not until years later, when he went on to study in the University of Edinburgh that he began to have different ideas of education.

> Trigger TT (#10): 那麼在很多年以後呢…尼爾他…這位創辦人他到…愛丁堡大學去上呃上大學的時候他才開始發現教育應該是不同的。(After many years, Neil he… this found, he went to the University of Edinburgh for college, and when he was in college, that was when he realized that education should be different).

doi:10.6342/NTU202000181

Trigger TT (#14): 但是之後他到愛丁堡大學的時候，他就開始有些對教育不一樣的想法。(But after that he went to the University of Edinburgh, he started to have different thoughts on education).

(Participant #10 & #14, sentence 7)

Despite the short lag at the beginning of the sentence, Participant #10 had many fillers, pauses and backtracks in the middle of the sentence. Participant #14, on the other hand, delivered the output at a steady pace and was also more concise. Participant #10 paused 3 times, repeated "上" from "上大學", used filler once, and was much more verbose in conveying the same message compared to participant #14. Participant #14, on the other hand, only made a very short and natural pause at the junction of the two sentences. In this end, participant #14 finished the sentence with a 2.7 seconds tail-to-tail span, whereas participant 10's TTS was 5.7 seconds – 3 seconds longer.

5.3.3 Succinctness

While high speech rate may contribute to shorter TTS, low speech rate does not necessarily lead to longer TTS if the interpreter translates succinctly. Theoretically, a sentence composing of less words would require less time to utter. Therefore, under the same EVS length and speech rate, the more succinct the rendition is, the shorter the TTS should be. However, not much empirical evidence supporting the theoretical assumption above had been collected. It was found that the overall performance of Participant #13 indicates that his/her output is more succinct and complete than the output of other participants.

Table 41 below summarizes the number of trigger sentence sets each participant translated (represented by N), the total word count each participant used to translated the entire speech, each participant's mean TTS, average correct rate of CU and mean 0-7th and 8-14th second

doi:10.6342/NTU202000181

scores. Only 5 participants translated all 19 trigger sentence sets. Participant #13 was amongst them. In addition, (s)he used less words to translate the entire speech (total words used = 2298), while the entire cohort used an average of 2356 words. Participant #13 also had the shortest TTS and also was the highest or second highest scorer for all three quality scores. This suggests that Participant #13 was succinct, correct, complete, and fluent in his/her rendition.

doi:10.6342/NTU202000181

Table 41 Overview of participants' key results (sentence sets rendered, word count, TTS, EVS, critical unit score, 0-7th second score and 8-14th second score)

Participant	N	Word count	TTS	EVS	CU score	0-7th score	8-14th score
1	19	2614	2.9	2.5	58%	6.8	7.5
2	19	2553	3.3	4.0	53%	7.2	7.5
3	15	1976	4	2.6	7%	3	5
4	19	3072	2.8	1.7	68%	7.7	7.4
5	17	2352	4.1	4.1	47%	5.4	7.6
6	16	1617	5.6	5.4	13%	1.4	5.3
7	15	2555	4.3	3.9	60%	7.8	7.5
8	19	2652	2.9	2.0	58%	6.8	8.2
9	18	2009	4.7	4.9	28%	2.1	4.1
10	18	2474	4.2	3.1	28%	3.3	7
11	18	2210	7.2	6.4	33%	1.4	5.3
12	16	2052	6.2	5.6	35%	2.3	4.4
13	19	2298	2.7	3.2	79%	8.2	7.6
14	18	2265	3.4	4.3	50%	7	7.1
Average	17.6	2356	4.1	3.9	44%	5.0	6.5

However, detailed examination comparing Participant #13 and the performance of other participants did not yield specific instances or clear trends between succinctness and the length of TTS. One possible explanation is that no matter how succinct an interpreter is, unless the anticipation strategy is adopted, the output is still paced by the speaker. Therefore, the influence of succinctness on TTS and SI output quality may need to be examined at an even more local level to have a clearer view of the relationship. The lack of clear connection between succinctness and TTS and SI output quality may be associated with the language pair and directionality of the present study. In Chapter 2, the features of English and Chinese were compared. English is more syllabic-verbose and form-reliant than Chinese is. Therefore, it is possible that the English source speech tolerates a larger range of verbosity from the

doi:10.6342/NTU202000181

Chinese target speech. In addition, most of the participants of this experiment have Mandarin Chinese as their A language. It has been argued that lexical knowledge and retrieval speed of lexical items are much better in one's A language than in one's B language(C.-c. Chang, 2005). Combining the linguistic features of English and Chinese and the advantages of different directionalities, it is possible that the English-to-Chinese SI happens to be a more verbosity-tolerant pair. Testing Chinese-to-English SI on Chinese A participants could be a better field to examine how interpreters' succinctness affects the length of their TTS and SI output quality, as interpreters are more likely to encounter gaps in lexical knowledge when producing speech in their B language, and the relatively syllabic-verbose nature of English could further amplify the effects of wordy renditions.

5.4 TTS and performance variance within participants

The final sets of statistical analyses looked at whether a participant's performance would be affected by his/her own variation in TTS length. The results showed that some participants' performance was susceptible to the length of TTS, while other participants remained unaffected. In addition, the effects of TTS were present in the 0-7th second score interval more frequently than in the 8-14th second interval.

Table 42 shows the p-value and coefficient of the three intra-participant correlations examined in this study. Significant correlations are underscored and marked bold.

doi:10.6342/NTU202000181

Table 42 Summary of intra-participant analyses

Part-icipant	N	Mean TTS	0-7th		8-14th		Critical Unit	
			p*	r**	p*	r**	p*	r**
1	19	2.9	0.444	-0.024	0.207	-0.143	0.161	-0.228
2	19	3.3	0.089	-0.234	0.135	-0.193	0.072	-0.336
3	15	4	0.081	-0.289	0.344	0.08	0.077	-0.368
4	19	2.8	**0.038**	**-0.321**	0.349	.067	0.291	0.128
5	17	4.1	**0.000**	**-0.658**	**0.012**	**-0.419**	**0.002**	**-0.805**
6	16	5.6	0.134	-0.228	0.056	-0.298	0.484	0.011
7	15	4.3	**0.012**	**-0.455**	0.067	-0.296	0.076	-0.47
8	19	2.9	0.069	-0.258	0.486	-0.006	0.082	-0.32
9	18	4.7	**0.008**	**-0.454**	0.111	-0.217	0.096	-0.308
10	18	4.2	0.274	-0.111	0.057	-0.285	**0.032**	**-0.437**
11	18	7.2	**0.012**	**-0.447**	0.057	-0.281	0.187	-0.491
12	16	6.2	**0.000**	**-0.794**	**0.012**	**-0.427**	**0.006**	**-0.599**
13	19	2.7	**0.003**	**-0.492**	0.284	-0.1	**0.028**	**-0.437**
14	18	3.4	0.323	-0.082	0.119	-0.208	0.401	-0.06

p*: one-tailed significance; **r: coefficient

Table 42 shows that the performance in 0-7th second is most likely to be influenced by TTS, followed by critical unit, and 8-14th second. A recovery effect can also be observed here. In the interval of the first 7 seconds, more participants were prone to the effects of TTS, whereas in the second interval, the effect has disappeared for most of the participants. The lack of consistent correlation across individual participants' results may be caused by the small sample size (N = 15, 16, 17, 18, 19). It is also possible that some participants appeared to be

doi:10.6342/NTU202000181

less susceptible to the variation in TTS length because they have developed strategies to maintain a steady lag.

The infrequent correlation between TTS and the critical unit score was also surprising. A possible explanation aside from the lack of a large sample size is that the critical unit score itself is a rather limited aspect to reflect the participant's performance. Because the critical unit score only looks at whether participants had comprehended and rendered the critical unit, and does not allow scores other than 0 or 1, the score may not reflect participants' performance in a very granular manner. For example, in instances where the participant may have failed to catch the critical unit but managed to come up with an educated guess, the participants would still receive a zero for the critical unit score. In the interval scores, however, the participants may not receive the full 10 points for such renditions, but would also not receive a zero as long as they have managed to convey a portion of the source speech message correctly.

The following is an example:

> Trigger: Now I am not saying that people who grew up from the system are always education bureaucrats.
>
> Critical: I've visited the Summerhill School; it's in England. (Critical unit: visit)
>
> Trigger: 不是說接受傳統教育的人就是這種想法，就是很傳統。(It's not that people who receive traditional education all think like this, like very traditional)

doi:10.6342/NTU202000181

Critical: 那英國有個叫夏山學校的(地方)。(地方 was delivered in the 8-14th second interval). (There was a place in the UK called the Summerhill school).

(Participant #14, sentence #6)

The TTS for the trigger sentence was 3.2 seconds. The rendition scored 0 for its critical unit, because "visit" was not present in the rendition. For the 0-7th second interval, rendition scored 8 points, with 3 points for accuracy and 5 points for fluency. The 2 points for accuracy had been deducted because the participant did not convey the information that the speaker has visited this place. The reason why this piece of information was omitted cannot be confirmed, as the study did not include retrospective interviews. It is possible that the participant might not have caught it at all, or have caught it but chose not to render it as a strategy. Either ways, the raters believe that the lack of this piece of information has undermined the accuracy of the interpretation. But despite missing this piece of information, the participant still rendered and delivered the other information in the critical sentence well. Therefore, the participant still obtained a fairly good score for this rendition. The score for critical unit is less flexible compared to the interval scores.

The second reason is that the critical unit score only takes the accuracy component of SI output quality into account, as opposed to the interval scores, where the participants could earn credit from good delivery. In extreme cases, a participant may score high in fluency but low in accuracy. The case below is an example:

Trigger: In fact, a couple had to apply for asylum in the USA to escape Germany to give their child home education.

doi:10.6342/NTU202000181

Critical: Legal homeschool centers also have to deal with a lot of hardships from the authorities from time to time. (Critical unit: legal homeschool centers)

Trigger: 比如說在德國、瑞典，在家自學是非法的，甚至有一些在家自學的人，必須要到美國來申請政治，政治庇護，因為這樣子才能在家自學。(For example, in Germany and Sweden, home education is illegal. Some people who try home education even have to apply for asylum from the US so that the children can learn at home).

Critical: 在家自學的小孩，也受到很多挑戰，(尤其是政府方面)。(尤其是政府方面 was delivered in the 8-14th second interval) (Home-schooled children also faced lots of challenges)

(Participant 2, sentence 18)

The TTS of the trigger sentence is 5 seconds. The rendition scored 0 for its critical unit, because "legal home school centers" was not translated. The rendition scored 6.5 for the 0-7th second interval, in which 1.5 points are for accuracy and 5 points are for fluency. The score for accuracy was very low because the gist of the message was completely absent. However, the delivery part of the rendition was flawless in the first 7 seconds. Therefore, both raters gave 5 points for fluency. The raters have discussed this phenomenon where the participant's product was quite irrelevant to the source text but delivered in a fluent manner. The raters then decided that in these very rare cases, highest possible score the participant would get is 5 points.

doi:10.6342/NTU202000181

5.5 Conclusion

The results of the study support the research hypothesis: a statistically significant negative relationship with large size of effect exists between TTS and interpreting quality. This can be explained with Gile's effort models. When the interpreter's TTS is longer, then longer portions of the production of the trigger sentence will be imported into the critical sentence and divert processing resource away from the critical sentence. As a result, interpreters have less resource for comprehension of the critical sentence, and also bear higher risks of overloading his/her memory effort.

In the results section, the research hypothesis was further examined in three aspects: the overall data trend, inter-participant analyses and intra-participant analyses. Tests for general analysis and inter-participant analysis both support the research hypothesis. However, tests for intra-participant analyses did not turn out significant for all participants.

Analyses with case illustrations were further carried out on factors that could potentially influence the length of TTS, which include the length of imported TTS, the length of EVS, the interpreter's speech rate and succinctness. All four factors could potentially influence the length of TTS based on temporal inferences, and except for succinctness, data collected from the study supported the factors. The reason why data supporting the relationship between succinctness of rendition and TTS length was lacking in this study may be associated with the language pair and directionality of the experiment.

doi:10.6342/NTU202000181

VI. Conclusion

This research sets out to explore the relationship between TTS and quality in simultaneous interpreting from English to Chinese. An experiment was designed to examine whether the TTS of a sentence would affect the quality of the following sentence. The research hypothesis postulated that TTS and quality have a negative relationship. Statistical analysis results of the data collected from the experiment indicate that a negative relationship between TTS and quality does exist, and that the length of TTS does have a strong effect on SI output quality.

The results of this study has four major implications. Firstly, the results of this study is in line with the findings of Lee's study on English-Korean SI in 2003. The agreement between the results of the two studies provide strong evidence that the relationship between TTS and SI output quality does exist, and the relationship exists not just in English-to-Korean SI but also in English-to-Chinese SI.

Secondly, compared to Lee's research, this study adds an extra dimension to interpreting quality – the fluency aspect. Fluency is an integral part of SI quality. Lee's research looked at the relationship between TTS and the accuracy dimension of SI output quality, whereas this research looks at the relationship between TTS and both accuracy and fluency of SI output quality. The results of this study show that SI output quality as a whole, with both accuracy and quality included in it, is adversely affected by the length of TTS.

Thirdly, this study also looked at the effect of TTS on SI output quality at inter-participant level and intra-participant level. The analysis at inter-participant level is conducted to understand whether participants' TTS patterns have an effect on their SI performance. The

doi:10.6342/NTU202000181

results show that participants with shorter TTS do tend to have better SI performance. This makes the case for interpreting trainers and trainees to place focus on their TTS.

The intra-participant level of analysis looks at whether the same participant's performance would vary as his/her TTS length changes. The results turned out to be mixed. Some participants are more inclined to be affected by their TTS lengths, whereas other participants do not seem to be affected much.

Although the results of the experiment are in line with the research hypothesis, the present study still has many limitations and insufficiencies.

Firstly, while the sample size for the general analysis (N = 247) is sufficient, the sample size for the inter-participant and intra-participant analyses are rather small (N = 14 for inter-participant analysis and N = 15, 16, 17, 18, 19 for intra-participant analysis). The small sample size may affect the power of the results of the statistical tests for the inter-participant and intra-participant analysis. The results could be better supported with a replicated experiment with a larger sample size. However, due to time and budget constraints of the research, the present study is limited to the current sample size.

Secondly, many of the factors that influence the length of TTS were not explored in-depth, apart from EVS. For example, speech rate and succinctness were both analyzed at a macro-level instead of a local level. To better understand how speech rate and succinctness influence the length of TTS, a much more in-depth analyses of pause and stall patterns, speech rate at local level, the use of anticipation strategy and the relationship between syllable-count of the target speech at a local level and TTS.

doi:10.6342/NTU202000181

Thirdly, quality assessment used in this research differs slightly from the assessment methods used in previous researches which use propositions(Ding, 2017), errors and omissions(Pio, 2003), idea units (Liu, 2001) or entire sentences as assessment units. The purpose of using a specific phrase at the start of a sentence and time intervals was to examine whether a recovery effect can be observed from the data. As discussed in previous chapters, using a time-interval-based rating method had indeed caused some challenges. Although a statistical test was used to examine whether those challenges had an effect on the overall relationship between tail-to-tail span and the other quality scores, the possibility that such challenges have caused different results for inter-participant and intra-participant analyses cannot be ruled out.

Fourthly, the participants of this study were all interpreting trainees at the time they attended the study. The results may be different with an experienced group of interpreters. Perhaps the relationship between TTS length and SI performance at the intra-participant level will be clearer. Even though previous studies rarely touched upon TTS and its effect on SI output quality, books and research have mentioned the importance of finishing rendering a sentence as soon as possible (Jones, 2014; Chang & Schallert, 2007). Even without the support of empirical evidence, experienced interpreters may have already developed their own set of strategies for maintaining an appropriate lag time behind the speaker. It would be interesting to include experienced interpreters as participants and compare the experienced professional group and the trainee group.

Finally, this study is heavily focused on quantitative analysis, incorporating qualitative analysis only when necessary. But even the rare qualitative deep-dives show that the statistical relationship between TTS and quality found in this study should not be taken at

doi:10.6342/NTU202000181

simple face value. As some of the examples in this study have shown, a short TTS and high quality score could mean that the participant omitted a lot at the first place, and a short TTS was simply a result of the omission, instead of conscious strategy use. A more qualitative-focused analysis would help distinguish between consciously-maintained short TTS and short TTS as a result of omission, and also provide more information into successful strategies for short TTS.

Despite the limitations, the results of this study still serves as a reference for the conference interpreting community. The findings corroborate the experiences of practicing interpreters and previous studies that did not refer directly to TTS but indicated that finishing the sentence in a timely manner is crucial to interpreting quality. The findings also serve as evidence for conference interpreting instructors to coach students to deploy appropriate strategies to reduce their TTS.

doi:10.6342/NTU202000181

VII. References

Ahrens, B. (2005). *Analysing prosody in simultaneous interpreting: Difficulties and solutions.* Retrieved from CiteSeerX: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.667.2016&rep=rep1&type=pdf

Al-Khanji, R., El-Shiyab, S., & Hussein, R. (2000). On the Use of Compensatory Strategies in Simultaneous Interpreting. *Meta*, 548-557.

Amini, M., Ibrahim-González, N., Ayob, L., & Amini, D. (2015). Users' Quality Expectations in Conference Interpreting. *International Journal of Multicultural and Multireligious Understanding*, 1-17.

Barik, H. (1973). Simultaneous Interpretation: Temporal and Quantitative Data. *Language and Speech*, 237-270.

Bovair, S., & Kieras, D. (1985). A Guide to Propositional Analysis for Research on Technical Prose. University of Arizona.

Cecot, M. (2001). Pauses in Simultaneous Interpretation: A Contrastive Analysis of Professional Interpreter's Performances. *The Interpreter's Newsletter*, 63-85.

Chang, A. L. (2009). Ear-Voice-Span and Target Language Rendition in Chinese to English Simultaneous Interpretation. *Studies of Translation and Interpretation,*, 177-217.

Chang, C.-c. (2005). Impact of Directionality in Chinese/English Simultaneous Interpreting: Impact on Performance and Strategy Use. The University of Texas at Austin.

Chang, C.-c., & Schallert, D. L. (2007). The impact of directionality on Chinese/English simultaneous interpreting. *Interpreting*, 137-176.

Chen, C.-h. N. (2012). Ear-voice Span and Performance in English to Chinese Simultaneous Interpretation.

Chong, S. (2009). *Chinese Translation and Interpretation: An Overview.* Chinese Translator: S. Chong.

Crookes, G. (1990). The Utterance, and Other Basic Units for Second Language Discourse Analysis. *Applied Linguistics*, 183–199.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language* , 84-101.

Díaz-Galaz, S. (2012). La influencia del conocimiento previo en la interpretacion simultanea de discursos especializados: Un estudio empirico.

Diaz-Galaz, S., Padilla, P., & Bajo, M. T. (2015). The role of advance preparation in simultaneous interpreting: A comparison of professional interpreters and interpreting students. *Interpreting*, 1-25.

doi:10.6342/NTU202000181

Ding, Y. (2015). The Effect of Domain Knowledge on Consecutive Interprering Performance. University of Auckland.

Ding, Y. L. (2017). Using Propositional Analysis to Assess Interpreting Quality. *International Journal of Interpreter Education*, 17-39.

Gile, D. (1995). *Basic Concepts and Models for Interpreter and Translator Training.* John Benjamins.

Giles, D. (2008). Local Cognitive Load in Simultaneous Interpreting and its Implications for Empirical Research. *Forum*, 59-77.

Greene, A., & Turner, E. (1977). The construction and use of a propositional text base. University of Colorado.

Gumul, E., & Lyda, A. (2007). The time constraint in conference interpreting: simultaneous vs. consecutive. *Research in Language, 5*, 165-183.

Kirchhoff, H. (1976). Simultaneous Interpreting: Interdependence of Variables in the Interpreting Process, Interpreting Models and Interpreting Strategies. In F. Pöchhacker, & M. Shlesinger, *The Interpreting Studies Reader* (pp. 110-119). London and New York: Routledge.

Kraviarová, Z. (2013). Assessing Source Speech Difficulty for Research Purposes. *Forum*, 51-75.

Kriston, A. (2012). The Importance of Memory Training in Simultaneous Interpretation. *Professional Communication and Translation Studeis*, 79-86.

Kurz, I. (2001). Conference Interpreting: Quality in the Ears of the User. *Meta*, 394-409.

Lee, T.-h. (1999). Speech Proportion and Accuracy in Simultaneous Interpretation from English to Korean. *Meta*, 260-267.

Lee, T.-h. (2002). Ear Voice Span in English into Korean Simultaneous Interpretation. *Meta*, 596-606.

Lee, T.-h. (2003). Tail-to-tail Span: A New Variable in Conference Interpreting Research. *Forum*, 41-62.

Li, J. (2013). An empirical study into categorisation and cause of omission in simultaneous interpreting. POLYTECHNIC UNIVERSITY.

Lian, X. (2015). The Influence of Cultural Differences on English-Chinese Interpreting. *Contemporary English Teaching and Learning in Non-English-Speaking Countries*, 1-23.

Liu, M. (2001). Expertise in Simultaneous Interpreting: A Working Memory Analysis.

Liu, M., & Chiu, Y.-h. (2009). Assessing source material difficulty for consecutive interpreting. *Interpreting*, 244-266.

Liu, M., Schallert, D. L., & Caroll, P. J. (2004). Working Memory and Expertise in Simultaneous Interpreting. *Interpreting*, 19-42.

doi:10.6342/NTU202000181

Macías, M. P. (2006). Probing quality criteria in simultaneous interpreting: The role of silent pauses in fluency. *Interpreting*.

Moser, P., & AIIC. (1995). *Survey on Expectation of Users of Conference Interpretation.* Vienna: AIIC.

Ono, T., Tohyama, H., & Matsubara, S. (2008). Construction and Analysis of Word-level Time-aligned Simultaneous Interpretation Corpus. *Proceedings of the Sixth International Conference on Language Resources and Evaluation.* Paris: ELRA.

Pio, S. (2003). *The Relation Between ST Delivery Rate and Quality in Simultaneous Interpreting.* Retrieved from The Interpreter's Newsletter: https://www.openstarts.units.it/bitstream/10077/2475/1/04.pdf

Pym, A. (2008). On omission in simultaneous interpreting: Risk analysis of a hidden effort. In G. Hansen, A. Chesterman, & H. Gerzymisch-Arbogas, *Efforts and Models in Interpreting and Translation Research: A tribute to Daniel Gile* (pp. 83-105). John Benjamins Publishing Company.

Repovs, G., & Baddeley, A. (2006). The multi-component model of working memory: explorations in experimental cognitive psychology. *Neuroscience*, 5-21.

Seeber, K. G. (2013). Cognitive Load in Simultaneous Interpreting: Measures and Methods. *Target, 25*(1), 18-32.

Seeber, K. G. (2015). Simultaneous Interpreting. In H. Mikkelson, & R. Jourdenais, *Routledge handbook of interpreting* (pp. 79-95). Taylor and Francis.

Shlesinger, M. (2000). Interpreting as a Cognitive Process: How Can We Know What Really Happens. In S. Tirkkonen-Condit, & R. Jääskeläinen, *Tapping and Mapping the Processes of Translation and Interpreting: Outlooks on empirical research* (pp. 3-15). John Benjamins Publishing Company.

Timarová, Š., Dragsted, B., & Hansen, I. G. (2011). Time lag in translation and interpreting: A methodological exploration. In C. Alvstad, A. Hild, & E. Tiseliu (Eds.), *Methods and Strategies of Process Research: Integrative approaches in Translation Studies* (pp. 121-146). John Benjamins Publishing.

Tóth, A. (2011). Speech disfluencies in simultaneous interpreting: A mirror on cognitive processes. *SKASE Journal of Translation and Interpretation*.

Wang, B., & Tao, L. (2015). An Empirical Study of PAuses in Chinese-English Simultaneous Interpreting. *Perspectives: Studies in Translatology*, 124-142.

Wu, S.-c., Liu, M., & Chang, C.-c. (2008). Interpretation Evaluation Practices: Comparison of Eleven Schools in Taiwan, China, Britain and the USA. *Compilation and Translation Review*, 1-42.

Yagi, S. M. (2000). Studying Style in Simultaneous Interpretation. *Meta*, 520-547.

Yeh, S.-p. (2015). 口譯錯誤之辨認與分類 . National Taiwan Normal University.

doi:10.6342/NTU202000181

Zamanian, M., & Heydari, P. (2012). Readability of Texts: State of the Art. *Theory and Practice in Language Studies*, 43-53.

Zheng, W. (2014). The Principle Of Brevity in Simultaneous Interpreting. *2nd International Conference on Education, Management and Social Science (ICEMSS 2014)*, (pp. 136-138).

Zwischenberger, C. (2010). Quality criteria in simultaneous interpreting: an internaional vs. a national view.

doi:10.6342/NTU202000181

VIII Appendix

8.1 Experiment material

Trigger sentence: Blue

Critical sentence: Red

Critical unit: Red + underlined

#	Source Text	Critical Unit
0	On my way here, I was at the airport in New York. The airport was packed. [2s] I encountered a mother and her 9-year-old son, who looked a bit bored, while I was waiting for the elevator.	
1	Our website was on my T-shirt, so the mother asked me what that was. I said I was an educational consultant. The boy then asked me what I did for work. I told him that I helped people start special schools. [2s] In these special schools, you can decide whether you want to go to class or not.	Our website
2	Before the mother could think of something to say, without hesitation, the boy shouted, sign me up! I always almost get these reactions. Children are super excited, and parents will smile politely and ask, “but how will children be able to learn this way?” [2s] On the other hand, young children know by heart that they are natural learners.	The mother
3	Preschoolers never worry about how they are going to learn. Almost no student under 11 years old will say, "but how will I learn this way?" Why is that? Traditional education kills that ability. But human spirit is resilient. [2s] It endures for another 6 to 7 years, before the natural ability gradually disappears.	Preschoolers

doi:10.6342/NTU202000181

4	150 years ago, the compulsory state government education system was first created. The curricula adopted by the large majority of schools at that time was largely similar to that in the 19th century. Education in the 19th century was authoritarian, and had a strong focus on preparing students for university studies. The belief behind is that children are naturally lazy, and that schools have to force them to learn. Over the 150 years, many important educational pioneers have disagreed with that approach. But why does true change not take place? The strongest resistance comes from the education bureaucrats. [2s] Since the system was established, they have continued to control it, and attempted to grow it.	compulsory state government education
5	They grew up in the system, and therefore the system's approach naturally seemed unproblematic; even when today modern brain research clearly shows that children are natural learners, and that the coercive way is not the most effective way of education. Behavioral experiments came to the same conclusion. But somehow, policy makers, the majority of whom are from the system, refuse to believe it. [2s] Now I am not saying that people who grew up from the system are always education bureaucrats.	grow up
6	I've visited the Summerhill School. It's in England. The founder Mr. Neill's story is written on one of the walls of the school. Mr. Neill was brought up in the traditional Scottish education system. He taught at the public school right after graduation, and taught in the way he had been taught. [2s] It was not until years later, when he went on to study in the University of Edinburgh, that he began to have different ideas of education.	visit
7	Freedom and fun, he found, had an effect that was largely misunderstood by the public. They actually facilitate learning. In order to foster an environment where students could show and pursue their interests, students had complete freedom in choosing classes to go to. All of the school's decisions are made democratically, and students had equal voting rights as the faculty. [2s] Now, many people scoff at that idea, thinking how could it be possible that a bunch of kids can actually practice democracy?	Freedom and fun

doi:10.6342/NTU202000181

8	Neill's daughter Zoe Redhead now runs the school. She told me a funny story during my visit. Once the school meeting had made a decision that games could not be played in the computer room in the morning. So they passed the computer out the window, so they could play computer games. [2s] This may sound funny, but these students were able to follow the decision they made themselves, even when it means less game time.	Neill's daughter
9	The network of this kind of learner-centered programs is growing rapidly. And there are programs for all age levels, because everywhere people are becoming more and more dissatisfied with an unchanging traditional system. In Europe, the US and Asia alike, more and more such schools have emerged. The location of the schools doesn't matter. These schools have attracted hundreds of thousands of parents and children who find the traditional educational approach unhelpful, and set out to find alternatives. [2s] These schools don't have tests or homework, because it really isn't necessary for people who want to learn.	learner-centered program
10	Libraries, for example, allow you to come and go as you like. They do not make you sit down and test you on your way out; they won't stop you from leaving to punish you for poor grades, [2s] because they assume that you are here to learn something, and that is your business alone. Then why at all should schools be any different?	Libraries
11	The age configuration is also very bizarre. Just think what an artificial situation it is, to be in a room every day, where there are 25 or 30 people of your exact age, not to mention that everyone is doing the exact same thing. This will never again happen in your life! Why should we let children be socialized to that bizarre configuration? [2s] Many schools now offer mixed age classrooms, which have been found to be a very good way to foster learner-centered environments.	age
12	Learner-centered schools can be found everywhere. In Nepal, there is a learner centered institution. I learned about it from Zoe, and was so curious about it that I had to visit it. By the way, Zoe knows a lot of interesting stories. But more on that later. [2s] So I flew to the capital of Nepal, Katmandu, to see for myself this education miracle.	Learner-centered schools
13	The founder had run away from home at a young age. He went to India to educate himself, and came back to Nepal. He founded an orphanage, and he and the volunteers there taught the children themselves. It became one of the learner centered institutions we help support. [2s]	The founder

doi:10.6342/NTU202000181

	And it is also the one that I mention the most often, because it truly shows the power of learner-centered education.	
14	What they do with the children there is unbelievably amazing. For example, at our 2008 conference, the founder brought along a 12-year-old boy. He was brought to the orphanage when he was only 3 years old. Now 10 years late, he is getting his doctorate in physics at a famous German university. [2s] There are hundreds of such schools, with every one of them being distinctly different from each other.	do with
15	Many survive in very untraditional ways. In New York, for example, the Free School operates on income from buildings they have bought at auctions and rehabilitated. In Israel, there is a network of over 25 public democratic schools. The starter Mr. Levy is now working with city mayors to change the education in entire cities. [2s] Mr. Levy is a very knowledgeable person, who has great visions for education. He invited me to join one of his projects.	survive
16	Many refugees have not been able to complete school. Seeing the influx of refugees, he gathered a group of young teachers to help organize home school education centers that are tailored for refugees. At this point, it might seem like I'm just randomly talking about passionate education initiatives that have nothing in common. [2s] What these schools have in common is that they all take a learner-centered approach. Home education is one of the fastest growing learner-centered education alternatives.	refugees
17	40 years ago, when John Holt's groundbreaking book "Teach Your Own" was published, there were about 20,000 being home educated in the USA. Today, there are more than two million home educated children in the USA. But home education as well as some other alternative educations are not legal everywhere. Home education is illegal in Sweden and Germany. [2s] In fact, a couple had to apply for asylum in the USA to escape Germany to give their child home education.	40 years

doi:10.6342/NTU202000181

18	Legal homeschool centers also have to deal with a lot of hardships from the authorities from time to time. But despite the hardship, many still choose to do so. Home education is a good way to start a learner-centered education. And it is definitely worth spending time to get through those hardships. Home education is one of our organization's top focuses, because it is affordable, and anyone can start their home school initiative. [2s] I have helped started at least 50 home education centers in the last few years, and had gained quite some experience.	Legal homeschool centers
19	One person was of incredible importance. In the 1960s, I was studying for my master's degree at Yellow Springs, Ohio. I wanted to start an interracial recreation center in the town, but kept hitting brick walls. I was lucky to have met that important person Mr. Arthur Morgan, who started a progressive elementary school in the 1920s in Yellow Springs.	One person

doi:10.6342/NTU202000181

8.2 TTS and quality scores

Participant #	Sentence #	TTS Seconds	CU Score	Critical Score (0-7)	Critical Score (8-14)
1	1	2.8	1	8.5	10
1	2	2.1	1	8	7.5
1	3	2.3	1	8.5	7.5
1	4	2.6	0	3.5	9
1	5	6.8	0	3	1
1	6	3	0	7.5	2
1	7	4	1	9	7.5
1	8	3.1	0	6.5	9.5
1	9	4.2	1	7.5	8
1	10	2.3	1	7	8
1	11	2.5	0	0	8.5
1	12	3.4	1	8.5	8
1	13	2.2	1	10	9
1	14	5	0	6	8
1	15	2.8	1	7.5	6.5
1	16	2.2	1	10	9.5
1	17	3.9	0	7.5	8
1	18	4.8	0	4.5	8.5
1	19	1.4	1	6	6.5
2	1	1.8	1	10	8.5
2	2	3.7	1	9.5	7
2	3	2.5	0	7	6.5
2	4	2.6	1	6	10
2	5	2.8	1	8	10
2	6	4.2	0	7.5	3.5
2	7	2.1	1	6	6
2	8	3.6	0	2.5	8.5
2	9	4.5	0	10	6.5
2	10	1.4	0	7	8
2	11	4.1	0	0	0
2	12	4.5	1	7	9.5
2	13	2.8	1	9.5	9.5
2	14	3.8	0	6	7.5
2	15	4.1	0	6	6.5
2	16	2.1	1	10	9.5
2	17	4.7	1	8.5	9.5

2	18	5	0	6.5	6.5
2	19	2.7	1	10	10
3	1	4.1	0	4.5	4.5
3	2	2.7	1	8	8
3	3	2.4	0	5	2.5
3	4	4.8	0	0.5	0.5
3	5	4.7	0	0	3
3	6	4.7	0	0	8
3	7	4.7	0	0	3
3	8	4.2	0	5	7.5
3	9	3.8	0	0	6
3	10	4.2	0	4	3.5
3	11	3.7	0	0	3
3	13	2.9	0	0	5
3	14	6.2	0	7	9
3	15	2.9	0	7	3.5
3	19	4.6	0	7	8.5
4	1	6.8	0	5	9
4	2	10.1	1	4.5	7
4	3	7.3	1	7.5	10
4	4	9.5	1	5	4.5
4	5	9.7	0	5	7.5
4	6	4.6	1	9.5	3
4	7	10.9	1	9	8
4	8	8.6	1	8	7.5
4	9	12	1	9.5	8
4	10	6.7	1	8	6
4	12	7	1	9	8.5
4	13	9	1	7.5	9.5
4	14	7	0	8	8.5
4	15	6.9	0	7	4
4	16	6.7	1	9	6
4	17	2.4	1	9	7.5
4	18	2.9	0	8	9
4	19	2.6	1	10	10
5	1	3.2	1	8.5	9.5
5	2	2.2	1	9.5	9.5
5	3	2.5	1	9.5	9.5
5	4	5.4	0	4.5	6
5	6	7.5	0	0	7

doi:10.6342/NTU202000181

5	7	1.7	1	7.5	7
5	8	5.9	0	0	8
5	9	6.1	0	0	8.5
5	10	2.9	1	7.5	8
5	12	6.4	1	9.5	9.5
5	13	4.9	0	0	6.5
5	14	4.7	0	7.5	6.5
5	15	9	0	3	4
5	16	3.3	1	10	9
5	17	2.6	1	9	8
5	18	8	0	0	4.5
5	19	3.9	0	8	8
6	1	3.3	0	0	6.5
6	2	6.1	1	0	9
6	3	11.6	0	0	4.5
6	6	4.8	0	0	1.5
6	7	5.1	0	0	5
6	8	5.6	0	0	5
6	10	7.8	0	0	0.5
6	11	4.2	0	0	7
6	12	3.2	0	3	8.5
6	13	2.6	0	4.5	7.5
6	14	5.4	0	7	6.5
6	15	7.8	0	3.5	6
6	16	5.3	1	0	3.5
6	17	5.6	0	0	5.5
6	18	9	0	0	4
6	19	4.9	0	5	5
7	1	1.57	1	10	10
7	2	2.3	1	9.5	9.5
7	3	2.22	1	10	9.5
7	6	8.44	1	9.5	8
7	7	4.98	0	4.5	4.5
7	8	3.97	0	2.5	6.5
7	10	2.24	1	9.5	5.5
7	11	4	1	8.5	6.5
7	12	3.31	0	9.5	9.5
7	13	2.75	1	8.5	7.5
7	14	4.26	0	2.5	7
7	15	3.65	0	5.5	6

doi:10.6342/NTU202000181

7	16	5.3	1	10	8.5
7	17	5.6	0	7.5	6
7	18	9	0	5.5	7
7	19	4.9	1	9	7.5
8	1	3.8	1	8	8.5
8	2	2.1	1	8	10
8	3	1.6	0	7.5	9.5
8	4	2.3	1	6	8
8	5	2.5	0	4.5	7
8	6	3.9	1	8	5.5
8	7	2	1	9	7.5
8	8	2.9	0	8.5	8.5
8	9	3	1	9.5	8
8	10	2.4	1	9.5	8
8	11	4.2	0	0	9.5
8	12	5	0	0	7
8	13	2.9	1	4.5	9.5
8	14	2.6	0	7	9
8	15	1	1	9.5	5.5
8	16	2.4	1	9.5	9.5
8	17	2.6	1	10	9
8	18	4.9	0	6	8.5
8	19	2.3	0	3.5	7
9	1	3.4	1	8.5	9.5
9	2	5.9	0	0	7.5
9	3	3.9	0	0	0
9	4	6.3	0	0	0
9	5	4.5	1	0	4
9	6	5.1	0	1	2
9	7	3.7	0	4	3
9	8	4.5	0	0	3.5
9	9	7.9	0	0	2
9	10	1.8	1	7.5	5
9	11	3.5	0	0	0.5
9	12	6.3	0	0	8
9	13	3.6	0	0	3.5
9	14	9.9	0	2	4
9	15	2.8	0	1.5	5.5
9	16	6	1	3	7.5
9	18	9.4	0	0	0

doi:10.6342/NTU202000181

9	19	3.2	1	10	7.5
10	1	2.1	1	9	10
10	2	2.1	1	9	8
10	3	3.8	0	0	7
10	4	5.1	0	0	1.5
10	5	4.7	0	2	7
10	6	5.2	0	7	6.5
10	7	5.7	0	5.5	7
10	8	2	0	0	6.5
10	9	5.9	0	0	6.5
10	10	5.8	1	4	6.5
10	12	5.4	0	6	7
10	13	2.2	1	8	8.5
10	14	5	0	4	8.5
10	15	1.5	0	0	6
10	16	3.4	1	4	8.5
10	17	4.6	0	0	7.5
10	18	5.6	0	0	5.5
10	19	5.1	0	0	8
11	1	2.5	1	9	8
11	2	12.4	1	0	9
11	3	7.4	0	0	5.5
11	4	5.4	1	0	6.5
11	5	6	0	0	4.5
11	6	6.9	0	0	6.5
11	7	13.3	0	0	0
11	8	5.2	0	0	6.5
11	9	15.6	0	0	0
11	10	2.2	1	5.5	2.5
11	11	8.6	0	0	3
11	13	1.3	1	9	10
11	14	7.4	0	0	2.5
11	15	7.9	0	2	2
11	16	4.4	1	0	5.5
11	17	8.2	0	0	9.5
11	18	6.4	0	0	6.5
11	19	9.3	0	0	6.5
12	1	-0.2	1	9	4.5
12	2	7.7	0	0	5.5
12	3	6.3	0	0	7

doi:10.6342/NTU202000181

12	4	7.1	1	0	5
12	5	8.9	1	5.5	3
12	7	10	0	0	1.5
12	8	9	0	0	0
12	9	6.7	0	0	3
12	10	4	0	2	4
12	12	5.2	0	4	6.5
12	13	2.8	1	5	9
12	14	4.8	0	2.5	3.5
12	15	14.4	0	0	0
12	16	2.4	1	5.5	5.5
12	17	2.9	1	4.5	7.5
12	18	9.8	0	0	0
12	19	5	0	1.5	9
13	1	3.4	0	7.5	7.5
13	2	1.8	1	10	9
13	3	2.2	1	10	10
13	4	2.6	1	9.5	9.5
13	5	3.1	0	0	4.5
13	6	3	0	9	5.5
13	7	2.2	1	7	6
13	8	3	0	1.5	7
13	9	3.7	1	3.5	5.5
13	10	3	1	9.5	9.5
13	11	2.9	1	9.5	8.5
13	12	3.4	1	9.5	9.5
13	13	2	1	10	10
13	14	2.4	1	10	9
13	15	2.04	1	10	3
13	16	2.2	1	8.5	4
13	17	2.7	1	10	7
13	18	3	1	10	9.5
13	19	1.9	1	10	9
14	1	1.3	0	8	8
14	2	3	1	9.5	8
14	3	3.9	1	9.5	6.5
14	4	3.8	1	8	7
14	5	1.7	1	8.5	4.5
14	6	3.2	0	8	5.5
14	7	2.7	0	8	6

doi:10.6342/NTU202000181

14	8	3.2	0	6	10
14	9	2.6	1	7.5	8
14	10	3.5	1	9	8.5
14	11	5	0	0	7
14	12	4	1	10	7.5
14	13	3.9	1	9	8.5
14	14	3.7	0	7.5	9
14	15	4.3	0	7	4
14	16	3.6	1	7	7.5
14	18	4.8	0	0	3
14	19	2.8	0	3.5	9.5

doi:10.6342/NTU202000181