Using Translation Tools for L2-Learning in a Self-Regulated Environment

Clinton Hendry

A Thesis

In the Department

of

Education

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Education) at

Concordia University

Montréal, Québec, Canada

January 2023

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Clinton Hendry

Entitled: Using Translation Tools for L2-Learning in a Self-Regulated Environment

and submitted in partial fulfillment of the requirements for the degree of

Doctor Of Philosophy Education

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

______________________________ Chair
Dr. Mitchell McLarnon

______________________________ Thesis Supervisor
Dr. Walcir Cardoso

______________________________ Examiner
Dr. Laura Collins

______________________________ Examiner
Dr. Sarita Kennedy

______________________________ Examiner
Dr. Denis Liakin

______________________________ External Examiner
Dr. Eva Kartchava

Approved by ____
Dr. Sandra Martin-Chang, Graduate Program Director

January 13, 2023 ____
Dr. Pascale Sicotte, Dean of Arts and Science

Abstract

Using Translation Tools for L2-Learning in a Self-Regulated Environment

Clinton Hendry, Ph.D.

Concordia University, 2023

The goal of this dissertation is to investigate the use of Google Translate (GT), a free online translation software that includes translation, Text-to-Speech (TTS), and Automatic Speech Recognition (ASR) functionalities, for online self-regulated learning of Mandarin Chinese within an interactionist approach. It also explores how GT can be used for pedagogical purposes within a complex learning environment that combines computer-assisted language learning (CALL), online self-regulated learning (SRL), and informed by interactionist theories. This dissertation begins with a review of the literature around the importance of interaction in language learning, followed by how a fully online interactionist approach using GT as a language partner and interlocutor allows learners to practice language use whenever and wherever they please (Chapter 1). Next, the dissertation contains three manuscript-based chapters (Chapters 2, 3, and 4) and a concluding chapter (Chapter 5). Each manuscript explores one aspect of using GT for pedagogical purposes by addressing the following overarching research questions: 1) Can GT provide the necessary computer-assisted interaction including input, output, and feedback to promote second language (L2) learning (Manuscripts 1 and 2), and 2) Are learners willing and able to use GT in an online, self-regulated environment for the learning of Mandarin and its associated tones? (Manuscript 3).

The first manuscript investigates the use of GT's TTS as a source of Mandarin Chinese input when compared with a native speaker in terms of Intelligibility (how much is understood), Comprehensibility (how challenging something is to be understood), and Naturalness (how much does a synthesized voice differ from a human speaker). The second manuscript further explores GT's ability to interact with a human interlocutor by investigating how much Mandarin Chinese speech can GT recognize at various language levels (intermediate, advanced, and native) and whether it can provide transcriptions accurate enough to be used as feedback by the language learner. The third and final manuscript investigate the pedagogical feasibility of using GT and its built-in features (translation, TTS, ASR) in an online, self-regulated environment by exploring how a small group of participants acquire language features, develop self-regulated learning strategies, and perceive the GT-enhanced pedagogical environment as a venue for language learning.

This dissertation will contribute to the literature around using translation, TTS, and ASR software for language learning, as well as interaction theories, SRL, CALL, and the acquisition of Mandarin Chinese in general. This research innovates on existing online language learning research and interactionist approaches by positioning GT as an interlocutor despite its intrinsic limitations (it is after all, not a human). This dissertation will further help guide future research into how human beings can interact with computers for language learning and will only become more relevant as translation, TTS, and ASR software becomes more intelligent, capable, and life-like.

Acknowledgements

When I began my doctorate, I realized it would take awhile, but I knew I was going to finish one day. What I did not realize was how much of myself I needed to invest in it to make it to the end. Sometimes, it felt like I was asking too much of myself and it beyond my ability, and in some ways, that was true. I know I could not have made it this far without my colleagues, friends and, especially, my family. They encouraged me, inspired me, and pushed me, and I love them all for it. These few pages are dedicated to them, but in truth, this whole dissertation was only possible because of their support.

First, I would like to thank my supervisor and friend, Dr. Walcir Cardoso. From the moment I told him I intended to get my doctorate, he was supportive and encouraging. He provided guidance throughout every step of my dissertation, helped me find teaching positions when I needed work, and never lost his temper or let me down. He gave essential and critical feedback on every single piece of writing I shared with him no matter the size or importance. I am the scholar and instructor I am today very much due to his expertise, professionalism, and kindness. My committee members, Dr. Laura Collins and Dr. Sarita Kennedy, offered their own expertise, encouragement, and patience throughout my doctorate, and my dissertation would not have been possible without their patience and constructive feedback.

I also want to thank my friends who have been there throughout my doctorate. June Ruivivar, Chloe Garcia, Ross Sundberg, Anne-Marie Sénécal, Emily Sheepy, and Michael Barcomb have all been there for me through the good times (which were many) and the bad times (which were blessedly few). We spent endless hours researching and working side by side, we co-authored multiple works, we travelled around the world together, and we helped each other out whenever and wherever we could. I am grateful for every minute we spent together. I also wish to give a special thanks to Yue He and Honglin Li for all of their support throughout my doctorate. I could not have done it without their friendship, expertise, and patience.

I would also like to thank my family. My father, Neil Hendry, for asking me how my doctorate was going when most people had completely forgotten about it. My mother, Wendy Wallace, who is no longer with us, for her belief in education. My brother and sister, Travis and Alex Hendry, for believing in me. Thank you all.

However, more than anybody else, I would like to thank my wife, Danfeng Gao, and my son, William Gao Hendry. Without Danfeng's love, assistance, support, and inspiration, I would have given up before I even started. I love you, Danfeng. Now and forever. Thank you for sharing your life with me. To my son, William Gao Hendry, it is you I dedicate this dissertation to. You inspire me to be a better academic, husband, father, and person. I love you, son, and I hope I make you proud.

Contribution of Authors

This manuscript-based dissertation consists of three manuscripts, along with a general introduction (Chapter 1) that states the research rationale and objectives and a concluding chapter (Chapter 5). Given the nature of this format, it should be noted that there is some overlap in the content between manuscripts.

This manuscript-based dissertation was conceptualized over the course of my studies and meetings with my supervisors, Dr. Walcir Cardoso. As the first author and principal investigator for the included three manuscripts, I was the major contributor to the manuscripts' conception, content, design, data collection, and write-up, under the guidance of Dr. Cardoso. This contribution is reflected in my status as first author in all three manuscripts.

Table of Contents

List of Figures

List of Tables

Chapter 1

General Introduction

This dissertation explores an interactionist approach, as conceptualized by Chapelle (2005), for the development of an online self-regulated environment for the acquisition of Mandarin and its tonal features using a popular translation tool (Google Translate - GT) for input and output practice. This manuscript-based dissertation consists of three related studies that explore key issues within this autonomous learning environment. It combines research and theoretical approaches from several pertinent areas, including computer-assisted language learning (CALL), online self-regulated learning (SRL), and pronunciation instruction by examining the interactions between the key interlocutors (a person who takes part in a dialogue): GT and second language (L2) Mandarin learners.

The Interaction Hypothesis (Gass, 1997; Hatch, 1978; Pica, 1994) states that interaction, especially *modified* interaction accompanied by negotiation for meaning (Long, 1983; 1996) can facilitate language acquisition. To date, most research using this approach focuses on human interlocutors. I argue, along the lines of Chapelle and Jamieson (2008) and Egbert and Shahrokni (2018), that the only essential interlocutor for second language acquisition (SLA) is the L2 learner themselves, and that within an interactionist approach, software such as GT can provide the modified interaction triggers necessary for learning. GT, although essentially a free translation tool, affords two speech technologies that can be used to enhance pronunciation learning and practice within an interactionist approach and consequently be used as an interlocutor within this approach: Text-to-Speech (TTS), which converts textual information to audio, and Automatic Speech Recognition (ASR), which converts audio input into text.

In this first chapter, I begin by reviewing the literature surrounding interactionist perspectives within a CALL and SRL approach, with a focus on how they can inform pronunciation instruction. I then focus on how these approaches can accommodate pronunciation instruction, specifically targeting Mandarin and its associated tones, and how GT can be used within a combined interactionist, CALL, SRL, and pronunciation instructional approach for language learning. I finish this chapter by discussing how this dissertation addresses real-world challenges regarding pronunciation instruction, and how it can contribute to the body of research around online language instruction.

The three manuscripts (Chapters 2-4, respectively) focus on self-regulated technology-enhanced pronunciation learning. Manuscript 1 examines GT's TTS suitability as a learning tool in terms of providing appropriate input for learning. That is, I assessed its ability to be intelligible, comprehensible, and natural sounding in Mandarin with non-native (intermediate and advanced levels) and native speakers. Manuscript 2, on the other hand, addresses GT's ASR's effectiveness as a reliable L2 interlocutor by investigating its ability to understand L1 and L2 Mandarin speakers, and whether its transcriptions (a measure of intelligibility) are accurate enough for learners to identify their mistakes (e.g., if the learner says a word incorrectly, will GT transcribe the word as stated, incorrectly?). Last, Manuscript 3 uses an exploratory qualitative analysis to probe learner's perceptions of their experience while using GT for Mandarin learning combined with a customized instructional website - Moodle (an online learning management system or LMS, similar to Blackboard, D2L, and Canvas).

Interactionist Perspectives

Although not coined as the Interaction Hypothesis until much later (Long, 1996), the first sojourn into this area began with Hatch (1978). In her seminal work, she argued that discourse analysis could be applied to second language learning. At that time, language acquisition research focused largely on the productions of the learner. Hatch argued that research should also begin to focus on both the productions of the learner and the other interlocutor, that is, the discourse as a whole: "The important thing is to look at the corpus as a whole and examine the interactions that take place within conversations to see how the interaction itself determines frequency of forms and how it shows language functions evolving" (Hatch, 1978, p. 403). She further highlighted specific areas of interest such as clarification requests that motivate focus on form and strategies for topic highlighting. She ended her work by drawing attention to the challenges of this approach, including data collection and analysis of something as complex as discourse.

Later research began to develop what we now understand as the Interaction Hypothesis and interactionist approaches (e.g., Long, 1983; 1996; Pica, 1994). Long (1983) began by focusing on how interaction can be modified to affect language input and facilitate learning. He argued that speakers modify interactions to avoid conversation trouble or to repair discourse if a communication breakdown had already occurred. These strategies included comprehension checks (e.g., "Do you understand?"), repeating oneself or repeating after the interlocutor, relinquishing control of the conversation to the less proficient speaker, and accepting sudden changes or ambiguity of topic. Data collected from first language (L1) and L2 speaker interactions provided evidence that L1 speakers modify both speech and interactions when conversing with L2 speakers, but that modified interactions were specifically more common when two L2 speakers are interacting (Long, 1983). Later, Long revisited his theories, newly coined the Interaction Hypothesis, wherein he would re-examine modified interactions as negotiation for meaning (Long, 1996).

Negotiation for meaning is the process in which learners and/or proficient speakers of a language interpret signals of perceived comprehension, and then adjust their speech (content or linguistic form) and their interaction until the speakers believe they have reached an acceptable level of comprehension (Long, 1996). Specifically, Long (1996) defined it as “semantically contingent speech of various kinds” (p. 452) that was packed with a number of modified interactions such as repetitions and reformulations in response to a learners’ utterance, and it often refers to communication in which there is some focus on resolving communication challenges (Gass, 1997). These negotiations of meaning or modified interactions can broadly fit into a two-part model: Trigger → Resolution (Pica, 1994; Gass, 1997). A trigger is the initial stimulus from the first interlocutor that signals to the second to begin negotiation or modified interaction. The resolution is the attempt to mitigate or resolve the trigger/perceived problem in communication. These resolutions often take the form of modified interactions, and thus the interlocutors engage in negotiation for meaning.

Modified Interactions

These modified interactions that take place during negotiation for meaning appear any time there is the possibility for misunderstanding or communication hurdles, even between two L1 speakers (Pica, 1994). The key requirement for negotiation to take place is miscommunication (Gass, 1997). Miscommunication is whenever there is a mismatch between what the speaker intends to say, and what the hearer interprets (trigger). Negotiation is the attempt by both interlocutors to rectify this mismatch (resolution). These negotiations can then lead to language acquisition by providing what Pica (1994) calls the three learner-oriented and language-oriented conditions for language acquisition: comprehensible input (Krashen, 1982), comprehensible output (Swain, 1985, 1995), and attention to L2 form or noticing (Schmidt, 1990). The language-oriented conditions include positive (target-like) input, enhanced L2 input to increase salience, and feedback on negative (non-target like) input (Pica, 1994). How these conditions interact with CALL will be discussed in a later section.

Krashen (1982) argued that comprehensible input is required for language acquisition to take place. Specifically, Krashen, as part of his input hypothesis, argues that learners are only able to acquire new language that is near their current level of competence (that is, if it is comprehensible). Therefore, if input is within one level of the interlocutor's current language ability (*i* + 1, where *i* represents their current level), it is considered comprehensible, and learners can acquire language from this input. Other theories such as Pienemann's Processability Theory (2015) also argue that instruction is more effective within a zone close to the learner's current level. That is, from a cognitive perspective, a learner is only able to produce what they are able to process, and that we can predict language acquisition (for any language) based on the expected processability of a given linguistic structure. However, there are cases wherein the unmodified input is too advanced to be comprehensible. Modified interactions and negotiation for meaning can then be used to modify input through lexical repetition, isolation, replacement, or simplification, which can make input more comprehensible, and enable learners to internalize target forms within their interlanguage (Pica, 1994). Assuming success, modified output is then necessary for L2 mastery especially for the less salient features in the input (Swain, 1985). Comprehensible input is seen as the entrance requirement to access a new form, and following that, the learner may modify their output to attend to their interlanguage and learn to manipulate these new structures in target-like ways, i.e., comprehensible output (Pica, 1994). Last, the learner may be guided through these interactions to attend to language forms that might be challenging to learn inductively, as argued in Schmidt's noticing hypothesis (1990). Schmidt argues that for a learner cannot advance their language ability without consciously becoming aware or "notice" a linguistic structure or feature in the input they are receiving. That is, they must attend a new structure in their input for learning to take place.

These learner-oriented conditions (forms) are important for language learning (Long, 1996). Further, throughout any given negotiation, there are opportunities for all three conditions. The stronger interlocutor (e.g., more proficient speaker) may often choose to modify their speech to make it more comprehensible. The less proficient interlocutor can then take the opportunity to practice output, and over time, structures become more salient either due to explicit feedback or frequency of forms (e.g., Goldschneider & DeKeyser, 2001). It should be noted, however, that not every miscommunication will lead to negotiation, and even so, negotiation itself does not guarantee learning (Gass, 1997). What is known is that learning improves with negotiation for meaning, and it can happen with any possible interlocutor whether they are L1 speakers or L2 speakers of the target language (e.g., Long, 1996; McDonough & Mackey, 2008). The requirements to be an interlocutor so that they can support learning are that they can provide comprehensible input, are able to interpret comprehensible output, and allow for modified interactions to foster focus on form. An interlocutor would be further beneficial if they allowed as much negotiation as the learner wants.

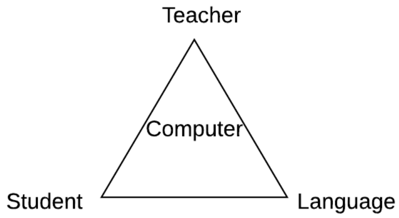
Computer Assisted Language Learning

In agreement with Chapelle and Jamieson (2008), I argue that using computer-based tools in an interaction meets and exceeds all requirements outlined above. CALL research has primarily been motivated by addressing learners' individual needs through the use of computer-mediated learning (Chapelle & Jamieson, 2008). Classrooms are physically and temporally limited, and students in foreign language settings often only receive language input within the classroom (Collins & Muñoz, 2016). Even then, only a small percentage of teacher-talk time is spent on speaking and pronunciation instruction (Foote et al., 2016). This is particularly challenging for the learner, as frequencies in the input are one of the strongest predictors of language acquisition (Ellis, 2002, 2012; Goldschneider & DeKeyser, 2001). Power differences such as between teacher and student may also increase stress and inhibit negotiation further (Gass, 1997).

Chapelle (2005) argues that interaction with language needs not be limited to people such as speakers of the target language or teachers. She proposes that computers can be used to mediate learning by giving students greater control, and then allowing for more opportunities for interaction and negotiation for meaning in and out of the classroom (see Figure 1).

Figure 1

Interaction with language in CALL (adapted from Chapelle & Jamieson, 2008)



However, this model may indicate a more complex relationship than at first glance. Egbert and Shahrokni (2018) propose a social interactionist approach for how interactions can take place between learner and computer with three components: learners may engage *around* the computer (e.g., to discuss a topic researched by a group of learners); *through* the computer (e.g., via computer-mediated communication such as videoconferencing and texting); and *with* the computer (e.g., interacting directly with computer based software such as GT) (Egbert & Shahrokni, 2018). By combining these approaches, I argue that within Chapelle and Jamieson's

(2008) model, "Teacher" and "Computer" can change places, and that the teacher can mediate the negotiation between students *around*, *with*, or *through* the computer. From here, the teacher can facilitate the learner from afar as they become more responsible for their learning, allowing more flexibility in how interactions may unfold between the living and non-human interlocutors.

Creating Interactions to Promote Pronunciation Learning

As mentioned above, interactions facilitate language learning. With their roots in discourse analysis (Hatch, 1978), interactionist approaches also fit well when applied to pronunciation and speaking instruction, as they highlight the importance of input exposure, output practice, and feedback, but focusing more on the importance of negotiating for meaning. Concerning pronunciation instruction, Celce-Murcia et al.'s (2010) recommendations outline four general steps that similarly highlight the importance of input, output, and feedback: 1) developing the ability to become aware and consequently discriminate (i.e., *perceive*) the target sounds (e.g., through listening discrimination tasks), 2) controlled production tasks with feedback, 3) guided production practice (more advanced practice) with feedback, and 4) communicative or unguided instruction with further feedback. Both interaction and pronunciation instructional approaches agree that we need input practice, output practice, and feedback for learning to take place.

On the other hand, although input practice exists whenever there is speaking and interaction, many interactionist proponents argue that certain types of input foster language learning more effectively. Specifically, the availability of both positive and negative evidence within the input can correlate with success (Long, 1996; Gass, 1997). That is, positive evidence (evidence of what is acceptable) within modified interactions can lead to learner uptake and can be combined with enhanced input to further make those target structures more salient (Pica, 1994). However, negative evidence (evidence of what is not acceptable in a language) should be combined with feedback, which provides information on accuracy, comprehensibility, and may even improve the learner's ability to identify non-target like forms (Pica, 1994; Lightbown & Spada, 1990). In addition to what types of input foster language learning within an interactionist approach, Gass (1997) also highlights different negotiation strategies for modifying input such as adding repetitions, or segmenting portions of each production to improve saliency or frequency of input. Gass further argues that listener comprehension is improved as a result of these interactional modifications. However, in addition to input and the various ways it aids language learning listed above, output is also essential within an interactionist approach.

Output practice provides opportunities for learners to attempt new structures and make mistakes (Long, 1996, Swain, 1985). Further, according to Swain's comprehensible output hypothesis (1985), speakers modify their speech to improve understanding by testing the limits of their speaking ability and finding new ways to express themselves. This output can then be used to provide feedback including triggering a resolution whenever there has been a miscommunication (Celce-Murcia et al., 2010; Gass, 1997).

Consistent with the interactionist hypothesis, feedback has been shown to benefit pronunciation instruction (Celce-Murcia et al., 2010; Long, 1996). Research into corrective feedback, which is normally defined as a learner receives either explicit or implicit information to correct their language use, but which also can be described as a feedback-oriented resolution of a trigger (Gass, 1997; Lyster et al., 2013), has shown it to be effective in both laboratory and classroom contexts (Mackey & Goo, 2007). Further, the more triggers that present themselves that lead to corrective feedback, the more likely there will be uptake of L2-target like structures (Mackey & Goo, 2007; Lyster et al., 2013). Corrective feedback itself can be in the form of recasts (repetition of the trigger but without the error, e.g., a student says “I did homework” and the teacher responds with “I did *my* homework”), explicit correction (clear indication of what was incorrect, e.g., same example, but this time the teacher responds with “you need to add ‘my’ to the sentence to indicate whose home was finished”), and prompts (explicit or implicit, e.g., same example, the teacher says “*whose* homework?” [explicit] or “homework?” [implicit]), all of which urge the learner to come to the correct L2 target-like production (Lyster & Ranta, 1996; Nassaji & Kartchava, 2017). Within an interactionist approach, recasts and explicit correction provide positive evidence, while prompts provide negative evidence. Although it is possible that the above types of feedback may provide incorrect information depending on the interlocutors (e.g., two beginner L2 speakers discussing advanced target forms), the expectation is that, over time, the majority of positive and/or negative evidence will lead the learner to correct L2 target-like productions (Pica, 1994).

Teaching Pronunciation with CALL

One way of motivating learners to continue to be engaged in input and output practice accompanied by feedback is via the use of technology. When teaching pronunciation with CALL, using appropriate tools, the focus remains on input, output, and feedback despite the changes in medium (Chapelle, 2001; Neri et al., 2002). Chapelle’s (2001) criteria for selecting and designing CALL tasks include potential for positive impact with attention to form and meaning; opportunities for feedback that relies on personalized experiences with authentic interaction; and the use of accessible tools.

Neri et al. (2002) outlines other specific recommendations for Computer-assisted Pronunciation Teaching (CAPT), including the presence of a stress-free environment, exposure to meaningful input, motivational oral output exercises, and opportunities for immediate feedback. Specifically, input should include both small and larger linguistic units that pertain to real-world situations (e.g., going to a restaurant, asking for directions, etc.) to increase motivation. Output should be elicited with realistic material combined with exercises that allow for different individualized learning styles. Last, feedback should be available immediately after each practice (Neri et al., 2002).

Neri et al.'s (2002) approach for CAPT resonates with Chapelle's (2001) criteria for selecting CALL tasks: each task should have the potential for positive impact on learning with opportunities for feedback and personalized learning. Further criteria also have interactionist roots, as Chapelle emphasizes the importance of opportunities for interaction and argues that CALL tasks should focus on both form and meaning, using interactional modifications such as repetition or reformulations. There is also a strong emphasis across these two abovementioned approaches on increasing the frequency and salience of target structures (Chapelle, 2001; Ellis, 2002). Recall that within an interactionist framework, frequency and salience can be enhanced through negotiation for meaning (Pica, 1994). However, although GT may have the potential to address all of the criteria above, and the above model combining CALL and interactionist theories sounds reasonable, there is still a very large question whether learners could effectively use GT or other TTS and ASR software for language acquisition successfully within this learning environment. One of the goals of this thesis is to address this, and to do that, the focus will be on the acquisition of Mandarin pronunciation using GP, and specifically, Mandarin tones.

Acquiring Mandarin Pronunciation

Mandarin, the target language of this dissertation, can be difficult to acquire due to its phonemic inventory like many other languages, but the most challenging factor for many learners is its lexical tone system (Chen et al., 2013; Song, 2021). Tones are suprasegmental structures in which pitch change affects lexical meaning. Consequently, tone use in Mandarin has a direct effect on intelligibility (how much is understood) and comprehensibility (how challenging something is to understand). Learners from non-tonal languages struggle to acquire tones due to a lack of L1 reference (Halle et al., 2004; Saito & Wu, 2014). For example, pitch in English is used at the sentential level only and only for pragmatic purposes such as when a speaker raises pitch to indicate a question. However, this same pitch change used in Mandarin words would likely render a word or sentence unintelligible. Consequently, vocabulary acquisition, even at beginner levels, is complicated by Mandarin's tonal system.

Mandarin has four tones and one neutral tone (characterized by no pitch change), which is only used in specific situations such as reduced syllables or certain particles. The four tones all carry lexical meaning. To illustrate, the word "ma" can mean mother (Tone 1), hemp (Tone 2), horse (Tone 3), or it could be used to curse or swear (Tone 4). There is often no clear connection in terms of meaning between which tone applies to which word. Chao (1968) uses a 5-point pitch scale (where 5 is high and 1 is low) to identify the four tones, as illustrated below:

1st Tone: 5 → 5 (a high, even pitch) = ma[T55]

2nd Tone: 3 → 5 (a rising pitch) = ma[T35]

3rd Tone: 2 → 1 → 5 (a falling and then rising pitch) = ma[T215]

4th Tone: 5 → 1 (a falling pitch) = ma[T51]

For this dissertation, tones will be labeled based on their pitch change (e.g., ma[T55], ma[T215], etc.), as indicated above.

Likely due to their complexity, learners acquire tones separate from their accompanying phonemic productions (Wan & Jaeger, 1998), and then acquire them in a developmental sequence predicted by the learner's L1 (Liu, 2000; Halle et al., 2004; Hendry, 2017). Maddieson

(1978) identified a tonal implicational hierarchy based on markedness that can predict this sequence across tonal languages. That is, if a language contains the marked dipping tone (rising and falling, T215), then it must also contain the less marked rising, falling, and level tones. If it contains a rising tone (T35), then it must also contain a falling (T51) and level tone (T55), and so forth. Research into L1 Mandarin tone acquisition has mostly confirmed this hierarchy (Li & Thompson, 1976, Hao, 2012), but L2 acquisition can be complicated by a learner's first language (Hendry, 2017).

L2 Mandarin learners who speak non-tonal languages struggle because they have no experience processing tones for lexical information (Halle et al., 2004). In Halle et al.'s (2014) study, French speakers strained to identify individual tones when presented in isolation but were able to discriminate differences in pitch change when presented with two tones. They were most accurate when discriminating between level (T55) and contour (T215) tones, which are on opposite ends of the tonal hierarchy (Maddieson, 1978). This indicates that the more drastic a change in pitch is, the more likely a non-tonal L1 speaker is to perceive it (see also Hao, 2012). In this dissertation, depending on the scope of the specific study, participants will be either L1 Mandarin speakers or speakers of non-tonal languages such as English or French.

Proper production of tones is required for comprehensibility and intelligibility. One only needs to call their friend's mother (ma[T55]) a horse (ma[T215]) to realize how important tones can be. In order to address the instructional focus of this dissertation (i.e., Mandarin tones – Chapter 4), Mandarin language instruction follows Celce-Murcia et al.'s (2010) framework for pronunciation instruction (outlined previously), combined with a CALL-based interactionist approach, discussed in the next section.

Teaching Mandarin Pronunciation via CALL within an Interactionist Approach

For a computer to be a successful interlocutor, it needs to be able to both provide and receive signals to trigger negotiation for meaning. As a computer cannot easily initiate practice, CALL tasks should be created to motivate the learner to use the computer in the desired way. Recall that Chapelle's (2001) criteria for CALL tasks include potential for positive impact, attention to both form and meaning, feedback, personalizing the experience to the learners' needs, and language interactions, while packaged in a practical, easy-to-use system. For pronunciation practice in a CALL environment, the computer needs to be able to create learner input (speak) as well as receive output (listen). It then needs to allow the learner to control their experience, making decisions such as when, where, and what they might want to learn, while promoting opportunities for authentic language use (Neri et al., 2002). Two technologies that allow for authentic language use are TTS and ASR. While the former allows learners to practice listening and reading (input), the latter provides opportunities for speaking (output). In addition, both provide ample opportunities for feedback, as will be discussed next.

Computer as Interlocutor: TTS and ASR

ASR and TTS are technologies that allow for users to speak (ASR) or listen to their computers (TTS), meeting the above requirements (i.e., it can create learner input, receive output, and allows the user to control their experience). While TTS software converts textual information to audio, ASR does the reverse: it listens to speech productions, interprets them, and transcribes them for the user to read. Together, these technologies provide a practical method for users to speak and listen to their computers. Software such as GT bundles TTS and ASR with their translation software, allowing users to control their input and output with little or no target-language knowledge, consequently giving them control over their L2 learning experience.

Previous research into using GT's combined software package for autonomous language learning has shown it to be effective in laboratory settings (Van Lieshout & Cardoso, 2022). Interestingly, much of the research into TTS and ASR has looked at their use separately, as will be discussed below.

TTS converts textual input to speech, thus providing learners the opportunity to adjust input frequency, or make certain targets more salient, determinants for all language learning (Ellis, 2002). It allows for theoretically unlimited perception practice for students (Cardoso et al., 2012), and has been shown to be effective with language learning when compared with similar in-class work (Bione & Cardoso, 2020; Cardoso et al., 2012; Liakin et al., 2017). Bione and Cardoso (2020) examined TTS's ability to perform in comparison with a human on four measures of pronunciation: comprehensibility, naturalness, intelligibility, and the user's ability to distinguish between past and non-past forms. Results showed there were no significant differences between comprehensibility and intelligibility between the target synthesized voices and those produced by humans, but participants did find the TTS significantly less natural sounding. Participants were also able to identify the target feature (past -ed) effectively in both the TTS and non-TTS group. Overall, TTS has been shown to be similarly effective as more traditional classroom-based language input, but it has the added benefit that gives learners full control over their input practice in anytime-anywhere settings, which may motivate learners to practice more frequently.

ASR's primary pedagogical affordance is that it provides users with opportunities to practice oral output, a component of interaction that is normally limited to the classroom. This practice has been shown to benefit language acquisition (Liakin et al., 2015). However, for ASR to be useful for SLA, Derwing et al. (2000) put forth two criteria: 1) it must recognize ESL speech at an acceptable level, and 2) it must be able to identify errors in a way similar to L1 language speakers. This means that the ASR should transcribe a target user's speech accurately, including any errors. Derwing et al. (2000) found that their ASR software, Dragon Systems' Naturally Speaking, accurately transcribed 90% of L1 speech, but only 70% of non-L1 speech. From an interactionist perspective, this creates several problems. If the software fails to transcribe target-like speech as accurate as an L1 speaker, learners may grow frustrated or learn to distrust their interlocutor. However, if the software automatically corrects for errors in speech, it will not signal to the learner that there may be an error in output, and so the learner may not have the opportunities necessary to attend to these errors (see Swain, 1995). However, it should be noted that Derwing et al.'s (2000) study is over 20 years old, and speech technologies such as ASR and TTS have advanced since then.

In a more recent study, McCrocklin et al. (2019) recreated Derwing et al.'s (2000) study with 20 advanced L2 English speakers using two popular ASR technologies: Google Voice Typing and Windows Speech recognition. Following the same protocol, McCrocklin et al. (2019) found that Google Voice Typing could be up to 90% accurate, but Windows Speech Recognition had only between 55-75% accuracy, which is even lower than Dragon System's Naturally Speaking used in Derwing et al.'s (2000) study. However, despite the variability in these results and the focus on advanced learners in both studies (Derwing et al., 2000; McCrocklin, 2019), other research has found ASR to be effective for SLA even when used autonomously (McCrocklin, 2016), and especially if combined with feedback (Penning de Vries et al., 2015).

Together, both ASR and TTS offer learners the ability to have unlimited input and output practice. Within this dissertation's approach, GT is in effect the *second* interlocutor necessary for negotiation for meaning. The learner can use the TTS function an unlimited number of times for input, focusing on target sounds, words, and phrases, providing an analogous experience to an interlocutor providing corrective feedback. The ASR function, assuming it operates at a similar level to a native speaker per Derwing et al.'s (2000) criteria, can be used for output practice and signal to the learner whether their production is correct or flawed. This can prompt the learner to repeat the target speech, correct production errors, or use the TTS function to listen again to the input. Gass (1997) argues that negotiation for meaning begins when there is a clear indication between interlocutors that understanding has not been reached. I argue that GT's TTS and ASR *together* allow a single learner practice and negotiate for meaning with GT, thereby mimicking a group activity, but in an autonomous environment.

Learning autonomously: Online Self-regulated Learning

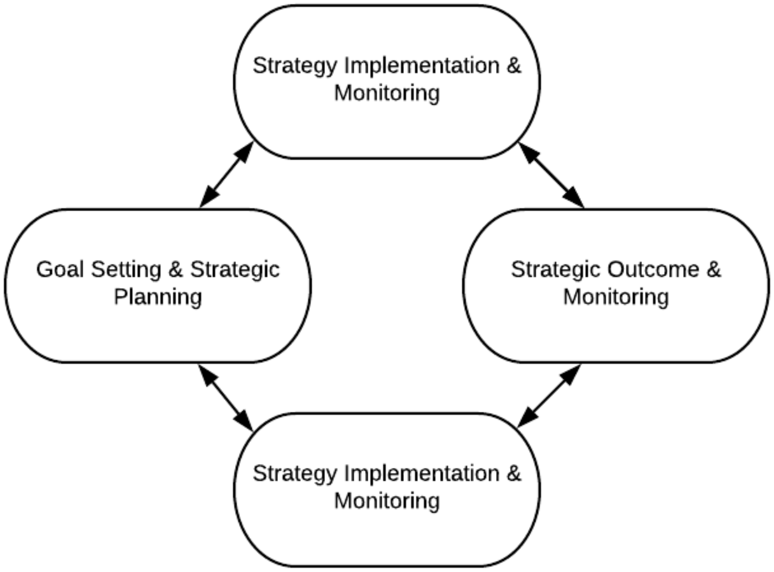
Google Translate's position as interlocutor within an interactionist approach lends itself well to one of CALL's primary benefits: to provide opportunities for anytime-anywhere instruction. However, GT's flaw as an interlocutor is apparent to any learner or teacher who has been involved with motivating their students to work autonomously; that is, it requires that the student become responsible for their learning (Andrade & Bunker, 2008). GT's TTS and ASR have the capacity to provide input and output, and signal miscommunications, but the nature of anytime-anywhere learning requires that learners choose if or when to learn.

Autonomous learners need to be in control of some or all aspects of their learning such as planning it (e.g., what language, amount of instruction, learning strategies, time, place). Strategies for learning autonomously are often referred to as self-regulated learning (SRL), so much so that autonomous learning and SRL are sometimes used interchangeably within the literature (e.g., Andrade & Bunker, 2008; Van Lieshout & Cardoso, 2022). However, for clarity, this dissertation will use SRL to refer to the strategies that lead to autonomous learning (Andrade & Bunker, 2008), while autonomous learning is on a spectrum and happens whenever the learner takes a degree of responsibility for their learning.

From its inception (Mlott, 1976), SRL has been of considerable interest to the education field (Winne, 2018). SRL requires that the learners apply themselves, including goal setting, strategic planning, and self-monitoring during learning, as seen in Figure 2 (Zimmerman, 1998). This model assumes that the learner will take agency over their learning and create individualized strategies to accomplish that goal, followed by implementation, and self-monitoring of each strategy via feedback. These assumptions overlap with those recommended by Celce-Murcia et al. (2010) for pronunciation instruction, as discussed earlier (i.e., learners move towards unscaffolded spontaneous speech as they navigate through the stages of pronunciation instruction).

Figure 2

Zimmerman's (1998) Cyclical SRL model


Strategy Implementation & Monitoring
Goal Setting & Strategic Planning
Strategic Outcome & Monitoring
Strategy Implementation & Monitoring

SRL research often draws from sociocultural approaches (e.g., Hadwin & Winne, 2001). Specifically, Vygotsky's Zone of Proximal Development (ZPD; Vygotsky, 1978) is often applied. The ZPD is a zone within which a person with assistance from outside sources becomes able to learn beyond what they were able to learn on their own (Lightbown & Spada, 2013).

This scaffolding can be implicit (e.g., via the careful ordering of activities) or explicit (e.g., via strategies for learning). It is considered essential within SRL research because, without assistance, students may become less likely to succeed in their learning (Winne, 2001).

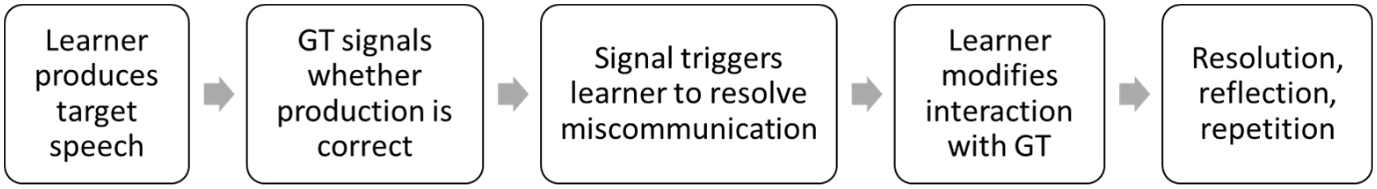
Applying SRL Scaffolding to an Interactionist, CALL, Pronunciation Instructional Design

Although a highly motivated learner could learn a target language without assistance, most learners will require some level of scaffolding to improve their ability to self-regulate their learning and to enhance aspects of the experience (Winne, 2001). Within this dissertation's design, scaffolding should accommodate SRL, interactionist perspectives, CALL, and pronunciation instructional design with the formal goal of fostering Mandarin learning (specifically acquisition of Mandarin tones) in an online, autonomous environment.

GT is ideal for this environment because it can provide the user with full control over their learning. Learners can choose the words and phrases they would like to learn in their L1, translate them, and then proceed to use GT as their interlocutor. From there, following Chapelle and Jamieson (2008), I argue that negotiation for meaning is possible, beginning whenever either interlocutor signals a miscommunication, thus triggering a resolution (Gass, 1997). These negotiations will operate as a form of scaffolding, making input and output more comprehensible as signals lead to corrective feedback either from the ASR or TTS functionalities. For example, if the learner produces a target Mandarin word incorrectly in their interaction with GT, the tool's ASR should provide a transcription that does not match the learner's target production, signalling a miscommunication. From there, the user can either repeat the target, or move to the TTS for corrective feedback and/or additional input. Figure 3 breaks down what this negotiation may look like in GT-based learning.

Figure 3

Example Process of Negotiation for Meaning with GT


Learner produces target speech
GT signals whether production is correct
Signal triggers learner to resolve miscommunication
Learner modifies interaction with GT
Resolution, reflection, repetition

It should be mentioned, however, that this GT-based process could be also understood as providing what Lyster (2002) coined *negotiation on form* as a related by distinct term to *negotiation of meaning*, discussed earlier. Lyster argues that elicitation, repetition of error, clarification requests, and metalinguistic cues (that is, comparatively implicit forms of corrective feedback such as what GT's ASR may provide, instead of more explicit forms of feedback or recasts) are more likely to provide corrective feedback to encourage self-repair instead of repairing a miscommunication. However, as the focus of this dissertation is using GT within a CALL based interactionist approach, I will continue to refer to it as *negotiation for meaning*, as this is the term used in CALL for similar types of interactions (e.g., Chapelle, 2007).

This design fits within the CALL and pronunciation instructional approaches highlighted in the previous sections, as it allows users to choose their own targets for learning. Further, the target environment should be relatively stress-free as the learner has control over when and what they learn, an ideal factor for pronunciation instruction in CALL (Neri et al., 2002). Most importantly, SRL, CALL, and pronunciation instructional approaches all highlight the importance of immediate feedback (Celce-Murcia et al., 2010; Neri et al., 2002; Zimmerman, 1998), and this design has the potential to give limitless opportunities for output practice with immediate feedback. This instant feedback should foster learners to both efficiently move through the SRL cycle (Zimmerman, 1998) and acquire new target phonological structures (Celce-Murcia et al., 2010).

The Learning Environment. One last key to this dissertation is where the learners were scaffolded to improve their SRL and learn to use GT for Mandarin language learning. Much of SRL and online language learning research proposes the use of LMS such as Moodle for their use as organizational learning tools (Dabbagh & Kitsantas, 2005; Godwin-Jones, 2015). They allow instructors to organize information for learners to easily access and use learning materials, an important aspect of CALL task design as well (Chapelle, 2001). For example, Moodle gives instructors the ability to create *books*, similar to PowerPoint presentations, where the learner must follow a series of pages in order. It further allows for assignments to be accepted via text or audio, which provides the instructor with the ability to collect learner production for assessment. Last, as LMSs such as Moodle, Blackboard, D2L, Canvas, and more have become the norm at the post-secondary level, most language learners will have experience with at least one of the above platforms. For this dissertation, Moodle was used to provide initial instruction in how to use GT, and the most scaffolding so that learners could start learning Mandarin. It was further used to collect data for research purposes.

Learning Mandarin in an Online Autonomous Environment

To analyze GT's pedagogical potential within this learning setting, I used Cardoso's (2022) chronological framework which, among other things, describes how technologies with pedagogical value are researched in CALL. It frames the exploration of GT as a language learning tool within the complex environment that arises from incorporating CALL, pronunciation instruction, SRL, and interactionist elements to create an autonomous L2 learning system. The first stage of the framework is the conceptualization and development of the tool, followed by exploration of how it can be used for pedagogical purposes. Next, the tool is assessed for its suitability, including testing its usability and learners' attitudes towards it. Last, it is tested for overall pedagogical effectiveness with a pre-post test design. Following this chronological (and organizational) framework, this dissertation explores how GT and its associated ASR and TTS capabilities can be used for pedagogical purposes within the intended learning environment, focusing on the L2 acquisition of aspects of Mandarin vocabulary and its tones. As far as I have been able to determine, there has been no research on how this language can be learned (including its tones) using the types of GT-based interactions discussed above.

To examine the pedagogical use of GT (in combination with its speech capabilities) and its affordances for the learning of Mandarin pronunciation and its associated tones, this dissertation has two overarching research questions:

1) Can GT provide the necessary computer-assisted interaction (including opportunities for input, output, and feedback) to promote L2 learning? (Chapters 2, 3)
2) Are learners willing and able to use GT in an online, self-regulated environment for learning Mandarin and its associated tones? (Chapter 4)

To answer these questions, I have performed three studies, each evaluating whether the proposed design could be used for Mandarin language and tone acquisition by:

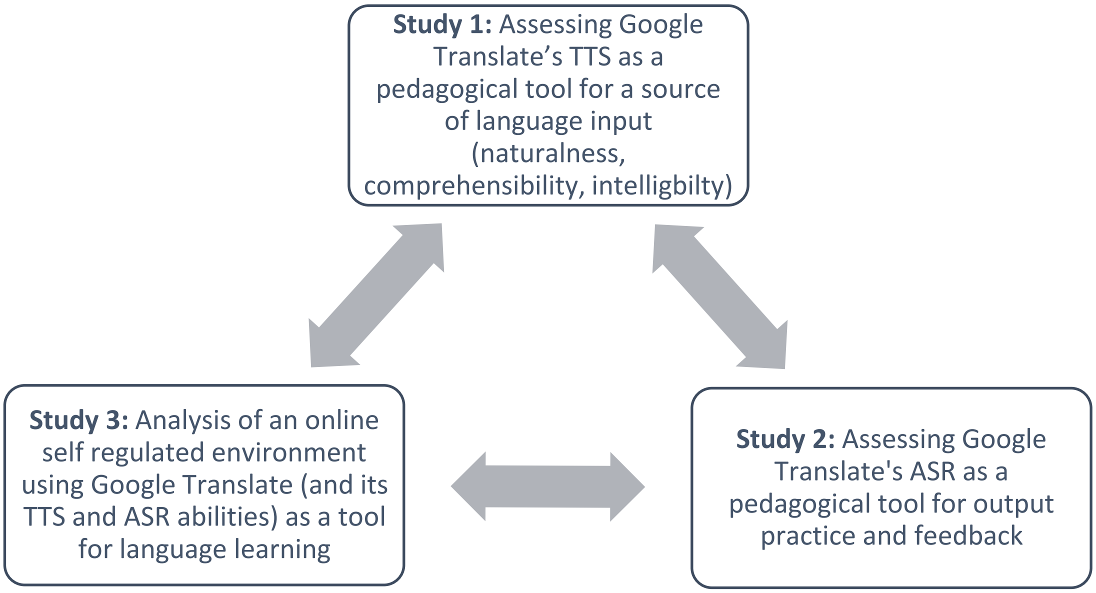
1) Assessing Google Translate's TTS functionality in a previously untested language, Mandarin Chinese, for naturalness, comprehensibility, and intelligibility. (Chapter 2)
2) Assessing Google Translate's ASR's pedagogical effectiveness and suitability for output practice and feedback in Mandarin. (Chapter 3)

3) Analyzing whether the proposed self-regulated online environment can lead to effective language learning and learner satisfaction. (Chapter 4)

A visualization of the dissertation's structure can be seen in Figure 4. Study One evaluates Google Translate's TTS ability to provide target input for Mandarin language learners by comparing the tool's productions with that of an L1 Mandarin speaker in terms of naturalness, comprehensibility, and intelligibility, adapting Bione and Cardoso's (2020) and Cardoso et al.'s (2015) methodology. Intermediate, advanced, and L1-Mandarin speakers listened to a series of Mandarin sentences read by both GT and an L1 speaker, and then compared these two sources of input in terms of comprehensibility, intelligibility, and naturalness. Beginner language learners were excluded from the study as they were deemed unlikely to have the language experience necessary to understand the prompts; they are also ill equipped to judge intelligibility and comprehensibility. To assess intelligibility, all three groups completed a dictation task read by both GT and a native Mandarin speaker. For comprehensibility and naturalness, participants were asked to rate how comprehensible and how "natural" each sentence is on 9-point Likert scales. Comprehensibility, accentedness, and intelligibility scores were compared between (intermediate, advanced, and native Mandarin speakers) and within groups (GT and native speaker input) using a Mixed Methods ANOVA to determine if there were significant differences between the ratings of the artificial (TTS) and human productions.

Figure 4

Dissertation Design


Study 1: Assessing Google Translate's TTS as a pedagogical tool for a source of language input (naturalness, comprehensibility, intelligbilty)
Study 3: Analysis of an online self regulated environment using Google Translate (and its TTS and ASR abilities) as a tool for language learning
Study 2: Assessing Google Translate's ASR as a pedagogical tool for output practice and feedback

Study Two measures GT ASR's pedagogical suitability for language learning concerning output practice and feedback. First, I created a bank of Mandarin phrases from Mandarin speakers at different levels (intermediate, advanced, and L1-like) following the methodology outlined in Derwing et al. (2000) and McCrocklin et al. (2019). To meet Derwing et al.'s (2000) criteria for ASR software to be useful for language learning, GT's ASR must be able to understand learner productions at a rate similar to a native speaker. To assess GT's ASR, each learner production was given a recognition rating derived from the percentage of the words and phrases that the ASR transcribed accurately. I then ran a One-way ANOVA to analyze whether GT's accuracy was affected by language level and a Mixed-Methods ANOVA to determine whether the ASR software struggled more or less with different proficiency levels when compared with three native speakers.

Finally, Study Three examines the pedagogical potential of GT and its built-in speech capabilities within an SRL environment by qualitatively analyzing how feasible it would be for beginner-level language learners to use this system, and whether it is possible to acquire aspects of L2 Mandarin pronunciation in this environment. The study focused on whether participants felt they were able to learn in this “anytime-anywhere” pedagogical environment. It also examines whether they perceived the GT-based learning environment as motivating, and usable, and if they are willing to continue to learn in this way on their own. Participants completed an online self-regulated Mandarin course on Moodle where input and output practice was done solely with GT. They were then asked to practice what they had learned by recording themselves completing five tasks or “quizzes”: a Translation quiz, Listening quiz, Speaking quiz, Sentence quiz, and “Introducing yourself” quiz. Participant perceptions of GT, the learning environment, and tone acquisition were evaluated based on participants responses in an interview regarding their learning, which included direct questions on their understanding of tones, and whether or not they could perceive and produce them, analyzed via comprehensibility measures by raters. The results of this manuscript indicate whether learners perceive the technology as something to embrace and whether it might lead them to success in their language learning.

These studies and their individual contributions are summarized in Table 1 and described in more detail in the following, respective chapters.

Table 1

Summaries of Study One, Two, and Three

Study	Goal	Contribution	Participants
1	Assess GT's pedagogical suitability for *input* practice in Mandarin	Determine whether GT TTS software can be used as an acceptable source of language input	Intermediate, Advanced, and L1 Mandarin language users
2	Assess GT's pedagogical suitability for *output* practice and feedback in Mandarin	Determine whether GT ASR software can be used as an acceptable interlocutor when negotiating for meaning	Intermediate, Advanced, and L1 Mandarin language users
3	Analyze how learning takes place in the proposed SRL, GT-based environment	Examine whether learners find the proposed online environment effective for SRL learning	True-beginner Mandarin language learners (self-reported no Mandarin language ability, no previous formal instruction)

Chapter 2

The Pedagogical Appropriateness of Using TTS for Mandarin Language Listening Practice

Classroom-based language learning has physical and temporal limitations that are difficult to overcome. There are limited instructional hours in a day, and most learners only have a few hours of class a week (Collins & Muñoz, 2016). These circumstances may make it difficult for learners to be exposed sufficiently to new structures for learning, a determinant in language acquisition (Ellis, 2002). Pronunciation instruction and speaking practice are particularly vulnerable in the language classroom as teachers may be the only source for language input for learners. For example, Foote et al. (2016) found that very little teacher-talk time in language classrooms is devoted to language instruction, and only 10% of that time is devoted to pronunciation instruction. There is therefore a need for learners to find opportunities for exposure and practice outside of the classroom.

Interactionist approaches argue that interaction is essential for language acquisition (Long, 1996; Gass, 1997; Nassaji & Kartchava, 2017): They provide learners with opportunities for negotiation for meaning or modifications to facilitate interaction such as slowing speech, comprehension checks, corrective feedback, and other strategies (Lyster & Ranta, 1997). These modified interactions can lead to more comprehensible input and output (see Krashen, 1982; Swain, 1995), and so learners are more likely to attend to or *notice* difficult or less salient structures (see Schmidt, 1990). However, there may be few opportunities for interaction in the classroom due to the aforementioned constraints; as a result, some researchers argue that when classroom time is insufficient for quality interactions, including language input and practice, computer-assisted language learning (CALL) may offer potential substitutes (Chapelle, 2005).

One way CALL can complement classroom instruction is through technologies such as text-to-speech (TTS) synthesis, which converts textual input to audio, giving learners the opportunity to both select their target input and listen to it repeatedly – a boon to language acquisition (Larsen-Freeman, 2009). The use of this technology can provide opportunities for input practice either in addition to classroom-based practice or on their own (Liakin et al., 2017; Moussalli & Cardoso, 2019). This allows instructors to increase target structure exposure and create customized listening discrimination tasks, both advantageous for pronunciation practice (Celce-Murcia et al., 2010; Ellis, 2002).

One challenge when considering TTS for autonomous language learning, however, is that it requires the learner to choose the target language input to be useful for language learning. That is, the learner is restricted by what they already know in the target language, and so it would be frustrating for beginners who have minimal language ability to learn and practice new words and their pronunciations without considerable teacher assistance. Translation software such as Google Translate (GT), which includes TTS functionality, offers a unique opportunity, as it allows a learner to translate L1 words and phrases into L2 targets, and subsequently listen to these L2 constructions converted to speech. For example, if a beginner Mandarin L2 learner wanted to learn something directly relevant to their current situation such as “Excuse me, where is the bathroom?”, they would have prompt access to target input from GT’s translation software and its built-in TTS capability. GT also has automatic speech recognition technology (ASR, addressed in Chapter 3), which can be used for output practice.

There are, however, questions regarding the efficacy of GT’s TTS in terms of target language input, and whether learners can use its artificial productions for practice as a replacement for L1 speakers such as those found in the classroom (Bione & Cardoso, 2020).

Further, the research around using TTS for language learning has so far focused on Western languages such as French (Liakin et al., 2017), English (Bione & Cardoso, 2020), and Dutch (Van Lieshout & Cardoso, 2022).

To address these concerns, this study evaluates GT's TTS functionality and appropriateness for increasing target language input by comparing it with an L1 speaker using three metrics: intelligibility (the extent in which a message is actually understood by a listener), comprehensibility (how difficult it is to understand an utterance), and naturalness (an analog for accentedness, operationalized as the extent to which the synthesized voice produced by TTS differs from that of a human speaker) (Munro & Derwing, 1995). GT's TTS was chosen due to GT's translation functionality, accessibility, and anytime-anywhere nature which, combined, may give users unprecedented control over their learning (see Chapter 4 for the implementation of this hypothesis). The target language for this study will be Mandarin Chinese, whose acquisition can be demanding due to its complicated tonal system (Halle et al., 2004). As such, learners will more likely benefit from additional opportunities for input practice to increase the salience of a challenging structure (i.e., lexical tones).

Mandarin Language Learning

Mandarin can be challenging for learners to acquire. Although some learners may struggle with Mandarin's phonetic inventory, the body of research on Mandarin language acquisition generally agrees that Mandarin tones can be particularly troublesome (Chen et al., 2013; Song, 2021). Tones are changes in pitch that affect lexical meaning and can be found in multiple South-East Asian languages such as Mandarin and Thai, many African languages such as Yoruba, or even some European languages such as Swedish (Yip, 2002).

Mandarin has four tones: a high level tone (T1), a rising tone (T2), a dipping tone (T3), and a falling tone (T4). It also has a neutral tone (T0; no pitch change), used only with specific words. Tonal languages are particularly challenging for learners from non-tonal L1 backgrounds, as suprasegmental information such as pitch change is processed at the perceptual level in their L1s (Halle et al., 2004). As those from non-tonal backgrounds have no previous mental categorization of lexical pitch change, tones can be difficult to both perceive and produce. For example, French speakers primarily use pitch change for intonation such as when asking a question, and so they struggle to perceive pitch change when presented with individual tones. Interestingly, however, the same French speakers are able to tell that two tones produced in tandem may be different (Halle et al., 2004). Consequently, Mandarin Chinese learners from non-tonal language backgrounds would benefit from additional exposure to tones (e.g., via TTS) to facilitate the perception and noticing of features.

Google Translate

Within an interactionist approach (e.g., Long, 1996; Gass, 1997), GT's translation, TTS, and ASR functions make it a possible contender as interlocutor. The combination of the three functionalities allows learners to practice listening (via TTS) and speaking (via ASR) without the need of a target language speaker for input and oral interaction. Learners can enter a language item in their L1, and GT will translate it to the target L2 (Mandarin, in this case). From there, TTS can be used to create infinite input opportunities (Cardoso et al., 2012; Liakin et al., 2017), and it even allows the learner to isolate target phonemes or add additional words to create phrases and sentences with ease.

One possible concern with any TTS is the artificial quality to the speech. However, research has shown that although listeners prefer authentic L1 speech, learning is equally possible from both L1 and synthesized speech (Cardoso et al., 2015, Liakin et al., 2017). Other research has found that, in general, learners using TTS were able to acquire target structures at a similar or higher rate than non-TTS groups, where input is provided from human speakers (Cardoso et al., 2012; Liakin et al., 2017).

The combination of GT's translation and TTS functionality also provides learners with a unique opportunity to take more responsibility for their learning. Learners can not only choose when and where they would like to practice, as GT is available on any platform with a browser, but they can also choose L2 learning items that interest them, followed by an amount of input practice commensurate with their personal motivation. Combined, these functionalities allow learners to use GT and self-identify strategies for learning, including the use of modified interactions.

TTS in Language Learning

TTS software produces artificial speech in an attempt to mimic a human speaker. This is an incredible boon for language learners who might struggle to find opportunities for input practice, as it allows for theoretically unlimited language practice (Cardoso et al., 2012). For example, in their study on how the pedagogical use of TTS can improve learners' performance in French liaison (when a word-final consonant is followed by a word-initial vowel re-syllabify in certain contexts; e.g., compare liaised *les avions* /lɛ.z̲a.viɔ̃/ with non-liaised *les trains* /lɛ.trɑ̃/), Liakin et al. (2017a) found that participants were able to improve their control of this phonological phenomenon using TTS. However, despite multiple studies indicating users' willingness to use TTS (Liakin et al., 2017b; Van Lieshout & Cardoso, 2022), only a handful examine TTS' ability to create effective and accurate input for language learning (e.g., Bione &

Cardoso, 2020), and to my knowledge, none specifically addresses TTS's ability to produce tones (i.e., appropriate to serve as L2 input).

Munro and Derwing (1995) outlined three constructs that can be used to evaluate pronunciation: intelligibility (the extent to which a message is understood measured by a evaluating a listener's transcriptions), comprehensibility (listener ratings of how easy or difficult a message is to understand), and accentedness (listener judgements of how closely a production matches that of a native speaker). In a study on advanced Montréal English learners, Cardoso et al.'s (2015) participants rated a TTS program (Natural Reader 13) as significantly lower in terms of intelligibility, comprehensibility, and naturalness (an analog of accentedness, defined as listeners' perceptions of the extent the TTS differs from authentic human speech). However, there was no significant difference in participants' ability to identify a target phonological feature, the English past tense morpheme *-ed*, indicating that the TTS software can still be a valuable form of language input.

In a more recent study designed to formally evaluate TTS for its language learning potential, Bione and Cardoso (2020) compared a freely available TTS voice named "Julie" (https://neospeech.com) to an L1 English speaker also in terms of intelligibility, comprehensibility, and accentedness. In a foreign English-learning context (Brazil), intermediate level EFL participants listened to several stories as well as a series of unrelated sentences in English produced by both Julie and an L1 English speaker. The researchers found that Julie and the L1 English speaker were similarly intelligible and comprehensible, but the synthesized voice was rated considerably less natural sounding. However, when further testing participants' ability to recognize *-ed*, they found no significant differences based on input, hinting that naturalness may play a very limited role in intelligibility (similar findings were reported in Munro and Derwing, 1995 concerning accentedness).

Cardoso et al. (2015), Bione and Cardoso (2020), and Liakin et al. (2017a) all indicate that TTS can be effective for instruction and that its oral output is target-like enough for learners to perceive a challenging structure such as past *-ed* morphophonemics. However, their mixed results concerning intelligibility and comprehensibility indicate that TTS is not consistently able to reach an L1-like level of intelligibility and comprehensibility. Further, none of the above studies specifically addressed whether proficiency level might impact TTS's effectiveness (i.e., whether TTS is more pedagogically appropriate for certain language levels than others). Last, the targets of all the above studies have so far been Western European languages such as English and French; accordingly, there is a dearth of research on whether TTS can effectively produce intelligible, comprehensible, and natural sounding oral productions in other languages, and specifically tonal languages, such as the target of this study, Mandarin Chinese.

The Current Study

This study explores the appropriateness of GT's TTS as a source of input for Mandarin Chinese by comparing TTS productions with that of an L1 Mandarin speaker. Mandarin was chosen as the target language due to the inherent complexity created by its tonal system, which may prove challenging for GT's TTS to adequately mimic, and because previous research into TTS has focused on non-tonal languages such as English, Dutch, and French. Participants were chosen from three proficiency levels in Mandarin: intermediate, advanced, and native speaker. This is because, as mentioned previously, there is insufficient evidence regarding the effect of proficiency on the pedagogical efficacy of TTS. Accordingly, it is assumed that proficiency may have an impact on the suitability of the tool in a real-world context. The research question that guided this study is provided below:

1) In comparison with human speech, is GT's TTS system capable of producing Mandarin speech at a level that can facilitate language learning in terms of:
   (a) Intelligibility
   (b) Comprehensibility
   (c) Naturalness
2) Will Mandarin speakers of different proficiencies (intermediate, advanced, and native speaker) find GT's TTS equally intelligible, comprehensible, and natural sounding?

Based on Bione and Cardoso's (2020) results for English, I predicted that there would be significant differences in terms of naturalness between GT and the L1 Mandarin speaker, but there would not be significant differences between intelligibility and comprehensibility for the advanced and L1 participants. Further, I predict that any concerns with GT's intelligibility and comprehensibility may be exacerbated by the lower proficiency of the intermediate level speakers. This research will help determine GT's TTS's pedagogical appropriateness for input practice by comparing its intelligibility, comprehensibility, and naturalness in the acquisition of a tonal language, Mandarin Chinese.

Methods

Participants

Sixty-four participants were recruited to form three groups based on self-reported language level: intermediate (n = 22), advanced (n = 20), and native Mandarin speaker (n = 22). Participants were recruited through word of mouth first targeting students and former students of Beijing Language and Culture University. Each participant interested in joining the study was first sent an email detailing the requirements of the study; if they agreed, they were then asked to fill out a consent form. Language level was self-reported, as the learner population being sampled from is very diverse in terms of daily language use, instruction history, and ability. Many intermediate and advanced speakers have received little traditional language education but have lived in China for decades while others may have studied Mandarin for years but have never been to China. In addition, self-reported language level has also been found to positively correlate with more objective measurements in other works (*r* > .5; Hakuta & D'Andrea, 1992; Luc & Bialystok, 2013). There is no beginner group, as we believe they would struggle to understand both the TTS voice and the native Mandarin speech. We believe the intermediate group is likely the earliest level at which a participant could complete the tasks in this study. Their demographic information can be seen below in Table 2. All participants across the three groups reported some familiarity with Google Translate. After completion of the study tasks, participants were renumerated $20 CAD for their time.

Table 2

Participant Demographic Information

Group	Gender	Ages	L1 Language(s)
Intermediate	9 Female, 13 Male	18-45	English (n=12), French (n=3), Spanish (n=2), Hiligaynon (n=2), Italian, Tok Pisin, and Tongan
Advanced	6 Female, 14 Male	18-35	English (n=12), Brazilian Portuguese (n=2), Italian (n=2) French, Spanish, Dutch, Hindi, and Samoan
Native	12 Female, 9 Male	18-56	Mandarin Chinese

Data collection

Data collection took place in anytime-anywhere environments (complete flexibility in choosing both time and place), using a custom Moodle website. Moodle is a learning management system that allows instructors and researchers to create secure environments for distributing course materials or performing research. It has a wide range of features and available plugins which can allow for a highly customizable and secure experience. Participants were first given a unique login and password to access their private Moodle environment. From there, they were guided to the testing phase, which was divided into two sections: intelligibility (involving transcription) and comprehensibility/naturalness (involving rating – see forthcoming discussion section).

Using GT and the Chrome Audio Capture extension (a plugin that records audio directly from the browser), 22 sentences were recorded and uploaded to the Moodle webpage. The L1 Mandarin speaker data (the same 22 sentences as used to generate the GT data) were recorded using the Audacity phone app. Another researcher verified that all L1 Mandarin speaker sentences were of comparable quality (e.g., in terms of loudness and pitch) to the audio captured from Google Translate. Further, the L1 Mandarin speaker was instructed to read the sentences at approximately the same speed as Google Translate, following Bione and Cardoso's methodology (2020).

To analyze intelligibility, participants were asked to transcribe 20 sentences, 10 produced by GT, and the same 10 produced by a native speaker. For each sentence, they were first prompted to listen to each sentence, and then asked to write it out in Mandarin Chinese (i.e., using Chinese characters) to the best of their ability. Each sentence was then given a recognition score, a percentage that indicates the number of characters the participants got correct compared with characters that were incorrect. Homophones (characters with the same phonemic representation and tone) were counted as correct.

For comprehensibility and naturalness, participants listened to the remaining 24 sentences (12 produced by GT and the same 12 again produced by a native speaker) and provided a Likert rating between 1-9 (1 being difficult to understand or unnatural sounding, and 9 being comprehensible or natural, respectively) using a multiple-choice tool in Moodle.

Before each task, participants were given instruction on how to use the system on the Moodle website and practiced each operation at least once although they could choose to practice additional times. During testing, however, they were instructed to listen to each sentence only once. The order in which the stimuli were presented to participants was randomized and counterbalanced to reduce testing effects.

The sentences used for analysis were adapted from Bione and Cardoso (2020) and can be seen in Appendix A. To adapt the sentences for this study, they were first translated directly into Mandarin by an advanced language user. Next, two L1 Mandarin speakers with experience teaching vocabulary highlighted words they believed would be too challenging for an intermediate speaker. Any vocabulary identified this way was replaced with a more frequent equivalent in meaning or with a word that does not interfere with the overall reading of the sentence (word frequency data were collected from the Leeds Mandarin Corpus; University of Leeds, 2021). For example, in Bione and Cardoso's study, several sentences use the word "parrot" (鹦鹉) such as "Last Christmas, Jimmy received the best present: it was a parrot." However, parrot was identified as rare by the L1 speakers and, as a result, the word was replaced with "video game" (电脑游戏), a considerably more frequent vocabulary item. Further, as transliteration may also cause difficulty for readers, English names such as "Jimmy" were replaced with common Chinese names such as "Zhao Jing" (赵婧).

Data Analysis

Each participant recorded their responses on Moodle as described above. To address the primary research questions, intelligibility was calculated based on the percentage of syllables in each sentence (0-100%) that the participants transcribed correctly, and comprehensibility and naturalness were measured using the Likert ratings given per sentence (1-9). The data were then analyzed with a Mixed Model ANOVA for each group (intermediate, advanced, and native), as there were between-groups data (GT and L1) and within-groups data (intelligibility, comprehensibility, and naturalness).

A methodological limitation of the study (see discussion) was that the same sentences produced by both the TTS and the native speaker were used to assess these pronunciation measures. To address this limitation and examine possible testing effects due to the participants listening to the same sentence twice (once from GT's TTS and once from the native speaker, or vice-versa), paired-sample t-tests were used to determine whether the results improved between the first time a sentence is presented and the second time it is presented, regardless of whether the input was from GT or the native speaker.

Results

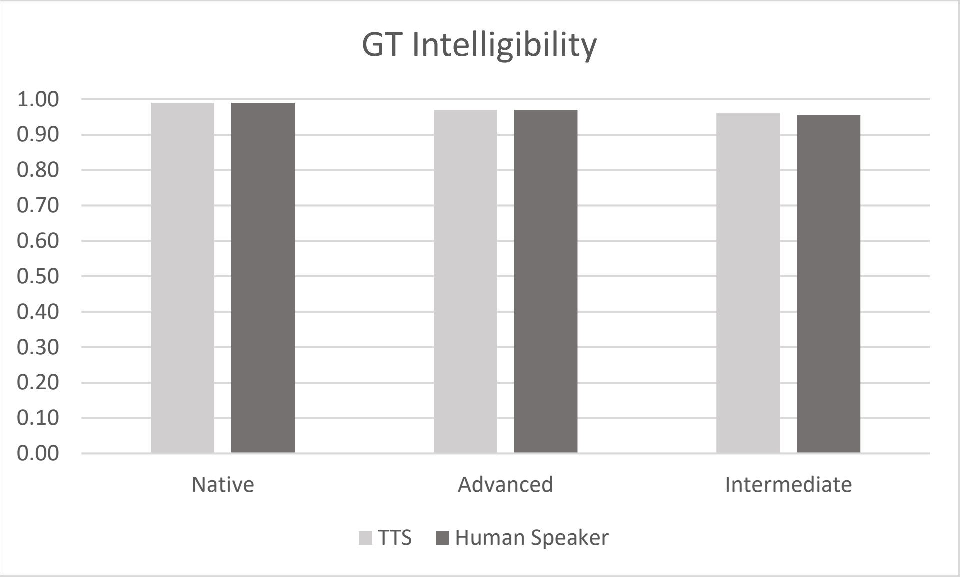
Intelligibility

A Mixed Model ANOVA was used to analyze within group effects (comparison of human- vs. TTS-produced input – also referred to as "input type") and between group effects (proficiency level). The results indicate that there was no effect for input type, *F*(1, 62) = .001, *p* = .98, and consequently there was no interaction between input type and proficiency level, *F*(2, 62) = 1.01, *p* = .37. However, there was a significant effect with a large effect size for proficiency level, *F*(2, 62) = 7.13, *p* = .002, η^2 = .188. The results can be seen in Figure 5.

Bonferroni adjusted post-hoc tests show a significant difference between native and intermediate speakers with a large effect size (*p* = .001, *d* = .6). However, there were no other significant differences between native speakers and advanced speakers (*p* = .062) and Intermediate and Advanced speakers (*p* = .51).

Figure 5

Intelligibility Results



Note: Higher is more intelligible

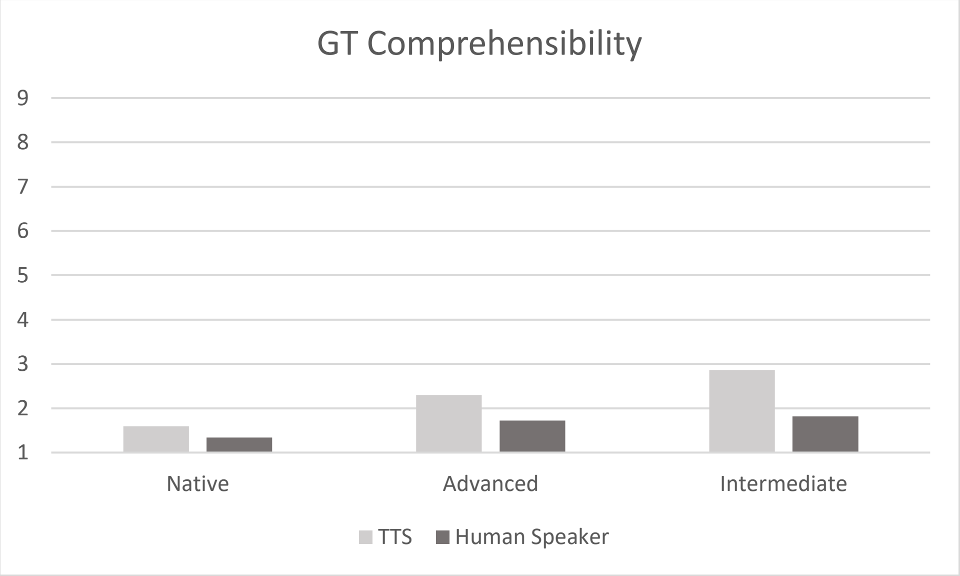
In summary, there were no significant effects for oral input, indicating that GT's TTS and the native Mandarin speaker were equally intelligible across all three language levels. The only significant difference was between the intermediate and native speaker groups indicating that the intermediate group had overall lower recognition scores than the native speakers.

Comprehensibility

Comprehensibility was measured using a 9-point scale (1 = high comprehensibility, 9 = low comprehensibility). Mixed model ANOVA results showed a significant difference with a large effect size for input type, *F*(1, 64) = 36.17, *p* = <.001, η^2 = .37, and a significant effect with a large effect size for the interaction between input and proficiency level, $F(2, 64) = 4.78$, $p = .012$, $\eta^2 = .14$. This indicates that comprehensibility changed across proficiency levels, depending on the oral input to which the participants were exposed. These results are presented visually in Figure 6.

Figure 6

Comprehensibility Results



Note: Lower is more comprehensible

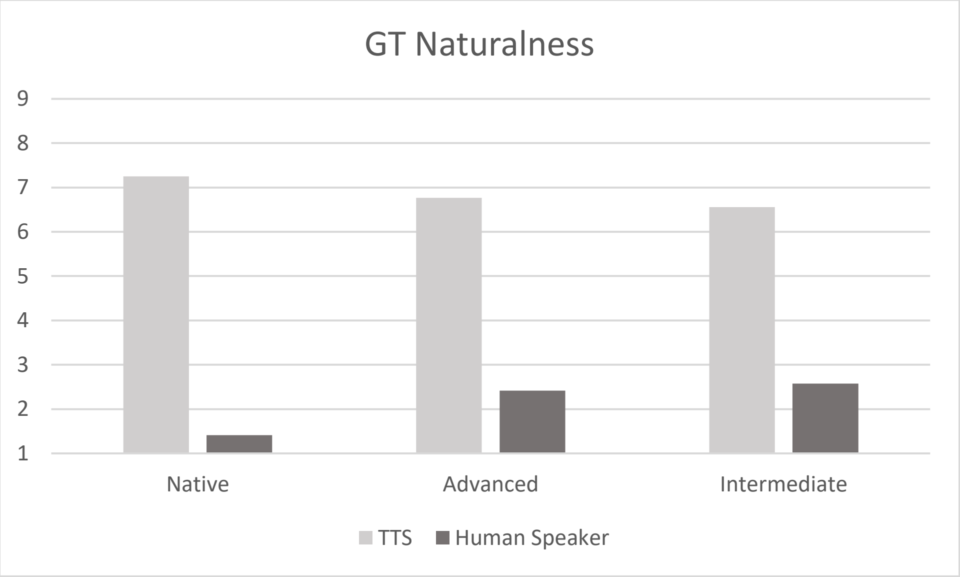
Bonferroni adjusted post-hoc tests indicate a significant difference between the native speakers and intermediate speakers depending on input type, although with only a small effect size ($p = .003$, $d = .1$). That is, intermediate speakers found GT significantly less comprehensible than a human speaker, while native speakers heard no difference in terms of comprehensibility.

Naturalness

For Naturalness, there was a significant difference for type of input, $F(1, 64) = 193.38$, $p < .001$, $\eta^2 = .76$, but there was no significant interaction between input and proficiency level, $F(2, 64) = 2.81$, $p = .068$, $\eta^2 = .08$. These results indicate that GT was significantly less natural sounding than the native speaker, regardless of level. These results are presented visually below in Figure 7.

Figure 7

Naturalness Results (lower is more natural)



Note: Lower is more natural sounding

Repeated Sentences (a methodological limitation)

The paired-sample *t*-test for intelligibility indicated no significant differences between the first and second time participants heard the target sentences (*t* = -1.53, *p* = .13). However, there was a significant difference in the comprehensibility scores (*t* = 4.59, *p* <.001) with a small effect size (*d* = .02) indicating some improvement when participants were presented with the second sentence. Naturalness also showed a difference (*t* = 5.49, *p* <.001) with a large effect size (*d* = .06) indicating a large improvement the second time the sentences were presented. In summary, these results indicate that repeated sentences did not affect intelligibility, but had a small effect on comprehensibility, and a large effect on naturalness. These results can be seen in Table 3.

Table 3

Intelligibility, Comprehensibility, and Naturalness: Means and Standard Deviations

Rater	*Time 1 M(/100)*	*Time 1 SD*	*Time 2 M(/100)*	*Time 2 SD*
Intelligibility	97	4.6	98	7.3
Comprehensibility	2.1	.97	1.8	.85
Naturalness	4.9	1.5	4	1.3

Summary of Results

To summarize, there was a significant difference in intelligibility based on participant proficiency level with a large effect size. Concerning comprehensibility, there was a significant difference for proficiency level and a significant interaction between input type (GT's TTS vs. Human) and proficiency level, both with large effect sizes. Post-hoc analysis indicated that native speakers found both inputs equally comprehensible, while there was a significant difference for intermediate speakers, who found GT significantly less comprehensible than the native speakers. Last, there was a significant difference with a large effect size for naturalness indicating that, regardless of participant proficiency level, GT's synthesized voice sounded significantly more artificial than the human speaker.

Discussion

The goal of this study was to determine whether GT's TTS system could produce Mandarin speech at a level of accuracy and quality that could facilitate language learning. We assessed GT's TTS by comparing its synthesized output to a native Mandarin speaker in terms of intelligibility using a dictation task (transcription), and comprehensibility and naturalness using the participant's holistic ratings. We then compared their results to determine whether there were any differences. These assessments are based on Munro and Derwing's (1995) work regarding the complex relationship between intelligibility, comprehensibility, and accentedness ("naturalness" in the context of a synthesized voice), as well as Bione and Cardoso's (2020) adaptations of these three constructs for use when evaluating TTS systems. Should the TTS' and native speaker's ratings be similar in all three constructs, we argue that learners could feasibly and reliably use GT within an interactionist framework for listening practice, and potentially receive accurate Mandarin language input in theoretically infinite quantities.

There were no significant differences in terms of intelligibility (measured through sentence transcriptions) regardless of language ability, thus indicating that GT can be understood as reliably as an L1 speaker. Comprehensibility (how challenging it is for the listener to understand a sentence) showed a significant effect for input type (TTS vs. human) and an interaction between input type and proficiency level. This indicates that there was a difference between the native and intermediate speakers based on the input they received, although with a small effect size. That is, intermediate speakers found GT less comprehensible than equivalent sentences produced by a native speaker, but native and advanced speaker raters found each equally comprehensible. Lastly, there was a significant difference in terms of naturalness, determined by ratings on how different from a native speaker a sentence sounds, but in this case the difference was confirmed regardless of rater ability: GT was significantly less natural sounding than a native Mandarin Chinese speaker. These results demonstrate that GT's TTS can accurately produce Mandarin speech in a way that is similar to a native speaker, albeit with some challenges, due to its artificial nature and unnaturalness. They also indicate that the target technology can be used as an intelligible conversation partner within the interactionist framework adopted.

However, as indicated earlier, an important caveat to the above results is that both GT's TTS and the native speaker produced the same sentences for assessing the quality of TTS-based speech in comparison with that of a human. Due to the study design, whether a participant heard a given sentence first from TTS or the native speaker was entirely random, but each sentence was listened to twice. While this decision allowed us to fully and accurately compare both TTS and native speaker samples, it risked a testing effect due to the repeated input. Consequently, although there was no effect for repeated input for the intelligibility ratings, there was a significant albeit small effect in terms of comprehensibility, and a significant large effect for naturalness. Therefore, the above results are influenced by our testing methodology. These results are discussed in detail below through that lens.

Pronunciation Quality: Intelligibility, Comprehensibility, and Naturalness

Intelligibility: Dictation Task

The combination of GT's TTS and translation functionalities allows learners to type in their L1 and produce target languages accurately. This study has found that regardless of GT's artificiality, it can produce language as intelligible as a native speaker at the sentence level. Intelligibility is different from the other two constructs because it is a more objective measurement of a listener's actual understanding of a text measured through transcriptions (Derwing & Munro, 2009). Comprehension and accentedness (or naturalness in the case of this study) are, instead, assessed primarily through Likert scales representing learners' subjective judgements. Consequently, comprehension and accentedness are considered to be partially related and overlapping constructs (Trofimovich & Isaacs, 2012), while intelligibility is more distinct from the other two constructs (Derwing & Munro, 1995, 2009).

The high intelligibility of GT's TTS is a strong take-away from this study, as it indicates that the input that learners would receive is highly accurate in Mandarin Chinese. These results are in line with Bione and Cardoso's (2020) results in English, despite some earlier works finding TTS less accurate than a native speaker (e.g., Cardoso et al., 2015). This may indicate that, as Bione and Cardoso posited, TTS software have advanced to the point where it can now reliably produce phonetically accurate language. Considering that their 2020 study and this one use different TTS software and target different languages, the bar may already be that high for most TTS software, at least with regard to intelligibility. In addition, this analysis also demonstrated that native and non-native speakers performed similarly, regardless of the learner's language level, as will be discussed in detail below.

Comprehensibility and Naturalness: Listener Judgements

Comprehensibility and naturalness, though distinct constructs, are known to be related (Derwing & Munro, 2008). Comprehensibility ratings showed an interaction between human vs. TTS input and the speaker's proficiency level (i.e., the TTS was less comprehensible for intermediate speakers than native speakers), and naturalness had a strong effect for input type (i.e., the TTS was less natural sounding than the native speaker), regardless of the proficiency of the listener. In terms of comprehensibility, these results differ from Bione and Cardoso's (2020), who found that their TTS software and human speakers were judged as equally comprehensible. This may indicate that GT's TTS struggles to produce comprehensible Mandarin Chinese in general. However, considering much of the previous TTS research in English and French (Bione & Cardoso, 2020; Liakin et al., 2017), it is possible that there were other factors at play.

For example, one factor may be that TTS is more developed in English and/or French than in Mandarin, and specifically, GT's TTS artificial productions may not be able to produce tones in such a manner that learners can understand them as effortlessly as a native speaker. In general, research into TTS has found that listening to synthetic voices requires listeners to pay more attention and, consequently, its use is associated with a greater workload (measured in length of time needed to recognize a word; Delogu et al., 1998). Listening to Mandarin TTS productions requires slightly different processes than English and French, as mapping the correct pitch change to the correct meaning adds a new layer for the TTS software, and may complicate Mandarin L2 speakers' ability to understand TTS productions. For example, the functional load of Mandarin tones has been found to be as high as that of its vowels (Surendran & Levow, 2004), and in a study on functional load in English, Munro and Derwing (2006) found that errors on features with a high functional load can affect comprehensibility and accentedness ratings. This may indicate that GT's TTS is not always able to mimic tones precisely, and when combined with its synthetic voice, comprehensibility and naturalness ratings may be negatively affected.

Another possibility for the interaction between type of input and proficiency level in the comprehensibility results, as mentioned above, is GT's TTS' low naturalness and the well researched effect of accentedness on comprehensibility (e.g., Trofimovich & Isaacs, 2012): it is possible that the low naturalness scores of the TTS affected its comprehensibility ratings. The results show a significant difference in naturalness for input type, which was consistent across levels, indicating unsurprisingly that TTS was less natural sounding than a native speaker. No research mentioned in this study has found TTS to be as natural as a native speaker at the sentence level (Bione & Cardoso, 2020; Cardoso et al., 2015; Liakin et al., 2017). However, something that may be underexplored is whether the well-known low naturalness of TTS systems is affecting their comprehensibility in a similar manner to how accentedness and comprehensibility are known to be associated (Munro & Derwing, 1995). That is, the results show a strong effect for input type with TTS's naturalness, a small effect for comprehensibility with only intermediate speakers, and no effect for input with intelligibility; interestingly, Munro and Derwing (1995) found correlations between accentedness and comprehensibility, but not between accentedness and intelligibility. So perhaps, the (un)naturalness of GT's TTS is associated with lower comprehensibility ratings in the same manner as accentedness is associated with lower comprehensibility ratings.

However, despite what may seem like a straightforward connection between the TTS analog of accentedness (i.e., naturalness) and comprehensibility, research has not always shown a clear connection between these two constructs. For example, Bione and Cardoso (2020) found no significant difference for intelligibility or comprehensibility but a significant difference for naturalness with their target TTS voice (Neospeech's Julie), perhaps suggesting that naturalness had no effect on comprehensibility. However, Cardoso et al. (2015) did find significant differences in intelligibility, comprehensibility, and naturalness between a human and a synthesized voice. Consequently, considering the mixed results found in the literature on TTS, it is possible the results in this study may actually indicate that GT's TTS elicits ratings more similar to what research has found when rating *human* speech. That is, GT's TTS (un)naturalness is affecting its comprehensibility but not its intelligibility, similar to how a human's accentedness affects their comprehensibility but not their intelligibility (Munro & Derwing, 1999; Trofimovich & Isaacs, 2012).

To summarize the above comparison between TTS and L1 Mandarin speaker input, the results of this study indicate that GT is intelligible, somewhat comprehensible, and assuredly unnatural sounding. In the next section, I discuss how input repetition may have had some effect on the results above.

Input Repetition

This study was designed so that the input from GT's TTS would be compared with the input provided by a native speaker. To do this, we had both TTS and the native speaker use the same sentences to create the target input; this decision, however, introduced possible testing effects from the speakers listening to the same sentence twice. Consequently, despite randomizing and counterbalancing the input, there was a small improvement in comprehensibility and a large improvement in naturalness the second time the participants heard the same sentence regardless of the input (GT or native speaker), but there was no change in intelligibility, considered by some as the ultimate goal in oral communication (Levis, 2018). These results thus indicate that naturalness and comprehensibility scores were affected. Specifically, it is clear that the participants' comprehensibility scores improved the second time a sentence was presented, and their naturalness scores improved a lot upon the second listening. Consequently, it is important that these results be validated in a future study by comparing groups who are not presented with the same input twice such as seen in Ruivivar and Collin's work (2019). That is, the sentences presented should not be repeated even by different speakers (GT and the native speaker).

Nonetheless, despite the improvements observed in comprehensibility and naturalness the second time a sentence was presented (see Table 3), the results of the study still indicate that the TTS sounds very unnatural, and that testing effects, although present, were minimal for intelligibility (none) and comprehensibility. In the next section, I discuss how the proficiency of the participants may have had a larger impact on the results.

Speaker's Proficiency Level

This study's design was loosely based on Bione and Cardoso's (2020), which tested human- vs. TTS-based input with *intermediate-level* Brazilian Portuguese learners of English. However, the current research differs considerably from its predecessor by targeting a different (and understudied) L2 (Mandarin), and by adding an additional proficiency level: advanced speakers. Previous studies into intelligibility, comprehensibility, and accentedness (e.g., Bione & Cardoso, 2020; Cardoso et al., 2015) focused mostly on native and either intermediate or advanced speakers of the target language, but not both. As shown here, GT's advantages are arguably suited for both less proficient and more proficient learners, who may all want more access to Mandarin input outside the classroom where there may not be sufficient time or resources (Collins & Munoz, 2016; Foote et al., 2016).

In terms of intelligibility, there were no significant differences for input type across all three levels, indicating that TTS-based Mandarin speech can be understood as well as a native speaker's, regardless of the listener's ability in their L2. These results align with some recent research into TTS' use in English and French (Bione & Cardoso, 2020; Liakin et al., 2017b), which have also found TTS to be highly intelligible. These results indicate that intermediate and advanced Mandarin learners can practice with GT's TTS with the confidence that they are receiving intelligible (and pedagogically appropriate) input.

In terms of comprehensibility, the results show that both input type and speaker proficiency level had an effect, indicating that GT's TTS was less comprehensible for intermediate speakers than native Mandarin speakers; however, there was no difference between native and intermediate speakers' comprehensibility ratings of the Mandarin native speaker. Naturalness, on the other hand, was flatly different based on input type, with no effect for language proficiency. These results may indicate that GT's TTS comprehensibility ratings follow similar trends to that of ratings of human speakers. That is, comprehensibility and naturalness are likely associated (as outlined above), and the artificiality of the TTS productions may be reducing comprehensibility, which may affect lower proficiency speakers more than higher proficiency speakers. However, regardless of the TTS' artificiality, these results indicate the importance of including less proficient speakers in future TTS analyses. This study, as well as others that focus on using TTS for language practice (e.g., Liakin et al., 2015; Cardoso et al., 2012), argue that the benefits of TTS are the provision of intelligible input, something that learners of all levels in need of listening practice might benefit from (Munro & Derwing, 2006), as this would allow them to interact with an interlocutor (in this case, GT) outside of the classroom. Therefore, it is essential that we understand its pedagogical appropriateness at various language levels.

Interaction

The results obtained in this study indicate that, although GT's Mandarin production is not particularly human sounding, it is very accurate in its productions and learners can understand it without too much difficulty. Pica (1994) outlines three learner-oriented aspects of language acquisition that take place during negotiation for meaning: comprehensible input (Krashen, 1982), comprehensible output (Swain, 1995), and noticing (Schmidt, 1990). With GT's TTS, learners are able to receive Mandarin language input, and in cases where there is a miscommunication, modify the input to improve comprehensibility (see Krashen, 1982), and regardless of the changes made, they can be sure that the input they are receiving resemble that of a human. Further, although beyond the scope of this paper, learners would continue the interaction by using GT's speech recognition capability to create opportunities for comprehensible output practice (Swain, 1995). Then, with the combined suite of translation, TTS, and ASR, there are also theoretically ample opportunities for noticing new or unique Mandarin language structures (Schmidt, 1990), and practicing them in listening and speaking interactions with the tool.

To summarize, these results indicate that GT's TTS functionality is intelligible, reasonably comprehensible, but unnatural sounding. Based on these findings, I argue that within an interactionist perspective, although not perfect, GT offers the possibility for learners to decide for themselves what they want to learn, how they want to learn, and how long they want to practice with a language "partner" that is accurate and understandable, and that never gets tired, bored, or frustrated.

Limitations and Future Research

This study has several limitations that need to be addressed in future research. One of these limitations is methodological in nature, as the study design used the same sentences for both the TTS and human input types, which may have led to testing effects. Repeating the input provided the most comparable sentences, which was ideal for this study design, but future research should consider instead using comparable but different sentences (such as sentences with the same number of words with similar tones). Another limitation is that participants self-rated their proficiencies, which although can be an accurate representation of one's proficiency (Hakuta & D'Andrea, 1992), there remains the possibility that participants were inaccurate in their assessment. Ideally, each participant would have been given more comprehensive language assessments. Nonetheless, despite these limitations, these findings point to interesting directions for future research.

The results of this study suggest that TTS software have the possibility to be used more effectively for language teaching both in and outside the classroom. GT's TTS has been found as intelligible as a native speaker and almost as comprehensible, but considerably less natural sounding. Future research may want to consider whether naturalness has any noticeable effect on acquisition or whether it is something that should be addressed pedagogically (e.g., by a teacher).

Future research should also look closely at the pedagogical implications of these results. Although the TTS software could theoretically be used to provide infinite input to L2 learners, and there has already been some research into how TTS can be incorporated into classrooms (Liakin et al., 2017), there has been no research as far as we know on how GT's Translate and TTS functions could be used together to improve the learner experience.

Conclusion

This study explored the pedagogical appropriateness of GT's TTS as Mandarin Chinese input by comparing intermediate, advanced, and L1 Mandarin speakers' ratings of both a synthesized voice and a native Mandarin speaker. It found that GT's TTS can produce intelligible Mandarin sentences with high comprehensibility, but low naturalness. These results suggest that GT can be used for practice by Mandarin learners. For instance, using GT, they can type in a sentence in their L1, translate it, and have GT read it out loud for them with some confidence that what they hear is intelligible for native and non-native Mandarin speakers. From there, they can create new sentences, modify them, and even negotiate for meaning with the tools available in GT.

Generalizing these results to the pedagogical context, teachers and students can investigate incorporating GT into their classrooms or as part of their homework, providing students with more independence in their studies and addressing some limitations of the classroom, such as finite time and space (Collins & Munoz, 2016). Further, GT is free, accurate (as this study has shown), and is available on all devices with an internet connection. Intermediate learners could choose to use it in real foreign language situations such as to practice survival Mandarin sentences while travelling. More advanced learners such as those living in China can use GT to practice their listening by preparing scripts to cover quick, challenging, or rare situations such as going to the hospital or interviewing for a job. These results show that as GT produces reliably intelligible sentences, and because learners can choose their target input from the L1, they can truly study Mandarin in anytime-anywhere situations using content that is always level and user appropriate.

Chapter 3

Evaluating Google Translate's ASR as an Effective Tool for Mandarin Language Learning

Corrective feedback has been repeatedly shown to be helpful for language acquisition (Nassaji & Kartchava, 2017). It provides useful information to learners in the form of error corrections that will make miscommunications more salient, and it helps learners attend to non-target like forms that cause communication breakdowns (Lyster & Ranta, 1996). However, opportunities for corrective feedback on pronunciation in in-class environments may be infrequent or inefficient (Foote et al., 2016). In their study on French learners in Québec, Foote et al. found that only 17% of teacher talk time consisted of language-related episodes (LREs, utterances focused on teaching or discussing language), and only 10% of teacher talk time was categorized as pronunciation instruction. Further, the majority of this time was used for recasts, the most common yet least effective form of feedback in terms of learner uptake (Lyster & Ranta, 1996). In addition, classroom time itself is also finite and often short (Collins & Muñoz, 2016), further limiting the class time devoted to spoken corrective feedback. This is particularly alarming from an interactionist perspective, which argues that modified interactions between speakers, also known as *negotiation for meaning*, provides essential opportunities for language learning (Long, 1996; Gass, 1997).

Interactionist theory posits that interlocutors often modify their speech to avoid communicative trouble or to repair discourse (Long, 1996). These modifications lead to more comprehensible input (Krashen, 1982; e.g., via lexical repetition or sentence simplification) and comprehensible output (Swain, 1995; e.g., via hypothesis testing by reading target forms out loud). When comprehensible input and output are combined, they can then increase the saliency of problematic productions and lead learners to attend to those challenging structures (Schmidt, 1990). Consequently, these modifications can often act as corrective feedback as interlocutors may use corrective feedback strategies (e.g., recasts, reformulations, explicit or implicit prompts, etc.) to make input or output more comprehensible (Pica, 1994). However, a barrier to entry in traditional interactionist paradigms is that there must be at a minimum of two interlocutors (Ellis, 1999), a challenging proposition for learners who do not have one readily available when studying from home or in a foreign language context.

Computer-assisted language learning (CALL) is rooted in interactionist approaches and the potential of the computer as an interlocutor (Chapelle, 2003). Research in the field has recently explored using technologies such as automatic speech recognition (ASR), and how a computer may better be able to provide corrective feedback (Bibauw et al., 2019). ASR listens to oral data produced by a speaker, interprets it, and then transcribes it. Research has shown that speaking to a computer using ASR or similar technologies can be beneficial for language acquisition (Liakin et al., 2015; Van Lieshout & Cardoso, 2022). However, for ASR to be useful for learning, it must reflect how people process L2 speech by recognizing L2 speakers and identifying their errors at a rate similar to an L1 listener (Derwing et al., 2000; McCrocklin, 2019). Otherwise, speakers will become frustrated if the ASR software frequently misinterprets correct L2 productions or if it does not provide accurate feedback (McCrocklin, 2016).

This study addresses these concerns by exploring whether a popular ASR system found in Google Translate (GT) can be used for the acquisition of Mandarin. To do this, a selection of pre-recorded sentences from intermediate, advanced, and native Mandarin speakers were analyzed with Google Translate (GT) through its ability to transcribe L2 speech accurately. These results will indicate whether GT's ASR can reliably transcribe both L1 and L2 Mandarin speech and whether the speaker's proficiency status affects these results. Mandarin was chosen as the target language because most previous ASR research has focused on Western languages such as English and French. Further, because its tonal system may be challenging at first for non-tonal L1 speakers, learners will likely benefit from additional opportunities for output practice and feedback (Celce-Murcia et al., 2010; Halle et al., 2004). GT was chosen because its translation function adds extra utility for learners when compared with other ASR software that require learners to know aspects of the L2 so that they can choose learning targets (Van Lieshout & Cardoso, 2022). Further, GT is free, available on all platforms with access to a web browser, and participants will almost universally have some previous experience with it.

ASR for Mandarin Pronunciation Instruction

Mandarin can be challenging for learners due to its phonemic inventory, but the body of research on Mandarin learning agrees that its complex tonal system is often the biggest hurdle in terms of language acquisition (Chen et al., 2013; Song, 2021). There are four tones in Mandarin: a high-level tone (T1), a rising tone (T2), a dipping tone (T3), and a falling tone (T4). There is also a neutral tone (T0, no pitch change), but it is only used in certain suffixes, reduced syllables, and particles. Tones are essential for lexical meaning (Yip, 2002). For example, ma(T1) means mother while ma(T3) means horse; as such, mixing them up is not recommended when visiting your in-laws. These tones complicate Mandarin acquisition for both tonal and non-tonal L1 learners due to the existing mental categorizations of what pitch change might mean in those L1 languages (Halle et al., 2004).

Tones can be particularly challenging for L1 speakers of non-tonal languages such as English or French, which use pitch change primarily for pragmatic or emphatic reasons. For example, in English, a speaker would raise their pitch at the end of the sentence “Is dinner ready?” to indicate that they are asking a question. However, if that English speaker were to raise their pitch at the end of a question in Mandarin, the question would become unintelligible. Because pitch is used differently in tonal and non-tonal languages, English and French speakers find even identifying tones challenging. Halle et al. (2004) found that French speakers were often unable to describe pitch change accurately and struggled to notice it in isolation although they were able to reliably perceive differences between two dissimilar words.

Due to their initially challenging nature and low salience for non-tonal L1 speakers, tones are often acquired separately from their accompanying syllable (Wan & Jaeger, 1999). That is, speakers first begin to produce target-like productions without tonal information, and tonal information is added later. However, it is essential that learners acquire tones at early stages in order to be intelligible. Consequently, pronunciation instruction early on is crucial if learners want to start producing intelligible speech quickly. Celce-Murcia et al. (2010) outline five general steps for pronunciation instruction: 1) initial introduction to the target structure 2) followed by discrimination tasks to raise salience, 3) controlled tasks with feedback, 4) guided tasks with feedback, and finally 5) communicative practice with more feedback. The increasing independence of these tasks, by design, fosters autonomy as the learner progresses through each stage until they are able to produce phrases that will allow them to communicate effectively and, at the same time, interact with others for effective language acquisition (Long, 1996; Gass, 1997).

Interaction Hypothesis

The Interaction Hypothesis (Long, 1996; Gass, 1997) posits that interaction, and specifically modified interaction or negotiation for meaning, can facilitate language learning by providing opportunities for comprehensible input (Krashen, 1982), output (Swain, 1995), and noticing (Schmidt, 1990), couched in the exchanges between two interlocutors. Interactional modifications such as repetition, paraphrasing, or other strategies can happen anytime two interlocutors need to repair or avoid a miscommunication even if both interlocutors are L1 language speakers. However, they are more common when there is a non-L1 speaker present (Gass, 1997).

These strategies for modifying interactions go under the umbrella of negotiation for meaning, defined as when competent speakers interpret signals about another interlocutor's level of comprehension, and then proceed to adjust or modify some aspect of their interaction including the linguistic forms, the conservational structure, or the content itself until understanding is reached (Long, 1996). Negotiation for meaning plays several roles in helping learners process and acquire language (Pica, 1994). Pica outlines three learner-oriented conditions that are affected during negotiation for meaning: comprehensible input, comprehensible output, and attention to L2 forms or noticing.

First, comprehensible input is necessary for learners to internalize new linguistic forms and structures from the target L2 (Pica, 1994). However, exposure to L2 input is not always sufficient for learners to internalize L2 forms and rules as it may be too advanced or challenging for the learner. Krashen's (1982) comprehensible input hypothesis argues that learners will acquire new language when input is just above the learner's current level (*i* + 1). Negotiation for meaning may increase the comprehensibility of the input when the interlocutor repeats key words, phrases, or sentences, slows down production, or invokes other strategies that aid the listener. Modified interactions also create opportunities for comprehensible output (Swain, 1995). Swain argues that output draws learners' attention to problematic or missing forms in their interlanguage. The interlocutor then modifies the language to remedy these issues which will lead to acquisition. Last, as outlined in Schmidt's (1990) noticing hypothesis, learners are unable to attend to new forms without first “noticing” them. Both comprehensible input and output help learners notice new forms that may be otherwise not salient in the L2 input. This is particularly relevant to the current study as tones, as mentioned before, are not particularly salient for learners with non-tonal L1s (Halle et al., 2004).

To conclude, negotiation for meaning offers multiple benefits that theoretically enhance second language acquisition. Each speaker is able to provide constant feedback as they signal miscommunications, and depending on the level of assistance required, it aids in always keeping input and output at suitable levels for understanding and language acquisition. However, the requirement of a competent interlocutor can be difficult to meet outside of the classroom for many learners. In part to address the above concerns, CALL research has examined whether a computer may make an adequate interlocutor (Bibauw et al., 2019; Chapelle, 2003).

CALL and Interaction

CALL research is motivated by addressing learners’ individual needs (including one’s ability to effectively communicate with others) with the help of computers and, as such, it has much of its roots in interactionist theories (Chapelle, 2005; Chapelle & Jamieson, 2008). ASR’s use in CALL is of specific interest to this paper as it provides speakers the opportunity for output practice and immediate feedback in the form of transcriptions. Within an interactionist model, ASR allows a human interlocutor to practice output, and attend to that written output as the software signals miscommunications (i.e., via incorrect transcriptions). This should further lead to more comprehensible output as the speaker experiments with various strategies to attempt new constructions as they struggle to improve their speech’s intelligibility.

Some of the most recent ASR research has focused on using intelligent personal assistants such as Amazon’s Alexa or Apple’s Siri (Bibauw et al., 2019; Dizon, 2020; Moussalli

& Cardoso, 2019). Known as "dialogue-based CALL" (Bibauw et al., 2019), this research has reviewed a broad range of software and shown how the interaction and negotiation for meaning afforded by these technologies can lead to noticeable learning gains in language acquisition. In a study on Japanese learners of English, for instance, Dizon (2020) found that 12-minute weekly sessions with Amazon's Alexa (a virtual personal assistant) were sufficient for significant improvement in speaking. In another study with Amazon's Alexa, Moussalli and Cardoso (2019) found that while interacting with an intelligent assistant, speakers implemented different strategies when confronted with communication errors, including repeating themselves, rephrasing their output, or sometimes abandoning the specific production and moving on to the next. However, despite its promise, Bibauw et al. (2019) argue that this area of research is still quite young and often focuses on broad topics with small sample sizes. The authors also contend that more research should continue to focus on the relative effectiveness of specific features of ASR-based technologies such as intelligent personal assistants.

Concerning ASR software, there has been some research that has shown it to be effective for language learning. For instance, Liakin et al. (2015) explored using ASR for the acquisition of the French /y/, a challenging structure for L2 learners in terms of both perception and production (a shared trait with Mandarin tones). They divided their participants into three groups: an ASR group which completed pronunciation activities with immediate feedback from the ASR software, a non-ASR group which completed the same activities but with teacher feedback, and a control group that practiced conversation skills with a teacher. Results found that only the ASR group improved at post-test. Concerning GT's ASR, the target of this study, Van Lieshout and Cardoso (2021) found that participants who used GT's ASR combined with its text-to-speech (TTS) and translation functions were able to learn 10 Dutch phrases with high comprehensibility, intelligibility, and low accentedness, attested in post- and delayed post-tests.

ASR has some possible drawbacks when used for language acquisition, however. As mentioned previously, Derwing et al. (2000) outlined two criteria for ASR to be useful for SLA: it must both understand L2 speakers and identify errors at similar rates of accuracy to L1 speakers. In the same study, they found that Dragon System's Naturally Speaking, a speech recognition program, could transcribe L1 English speakers' speech accurately 90% of the time, while it was only 70% accurate with L2 speaker productions. These same productions were found to be highly comprehensible by L1 speakers of English. Therefore, although it can used for practice, the ASR could not be relied on for quality feedback as it may cause leaners to mistrust it or grow frustrated.

However, the technology has evolved since then. More recently, McCrocklin et al. (2019) tested two ASR systems, Google Voice Typing and Windows Speech Recognition, using Derwing et al's (2000) methodology with 20 advanced L2 English speakers; they found that Google Voice Typing had up to 90% accuracy while Windows Speech Recognition had between 55-75% accuracy. Although promising, these results need to be interpreted with care as the target language was English, the most frequent language choice for all of the ASR studies mentioned in this manuscript. Further, the language levels for participants in both the above studies were advanced, with no intermediate or beginner level speakers. Questions remain regarding ASR's software effectiveness at multiple language levels and for the acquisition of languages besides English. Consequently, it is not clear whether current ASR software, which may be effective in English, is able to accurately capture L1 or L2 Mandarin speaker language productions.

The Current Study

To extend the above body of work on the value of computers as interlocutors, this study evaluates GT's ASR's ability to serve as a pedagogically appropriate tool for speaking practice and providing feedback. As such, it also explores the tool's ability to facilitate the acquisition of Mandarin, according to Derwing et al.'s (2000) criteria for determining ASR's effectiveness for SLA. Mandarin was chosen as the target language for this study as it is outside the gamut of Western languages commonly examined in ASR research, such as English and French. The research questions for this study are:

1) Does GT's ASR recognize Mandarin speech at a level commensurate with the speaker's ability? Are there proficiency effects in its speech recognition?
2) Does GT's ASR transcribe speech accurately enough to be used for signalling miscommunication and providing feedback?

Based on the research outlined above, I hypothesize that there will be significant differences between the three groups in line with their language level (e.g., intermediate-level speakers of Mandarin will have more transcription errors than advanced speakers, and advanced speakers will have more errors than native speakers). Further, assuming that GT's Mandarin ASR is as capable as Google Voice Typing is with English (McCrocklin et al., 2019), I hypothesize that there will be a non-significant number of transcription differences, specifically between advanced and native speakers; accordingly, there may be a significant number of differences between intermediate and advanced/native. This research will help determine whether GT can be used in anytime-anywhere online environments to provide theoretically unlimited practice opportunities for Mandarin language learners.

Method

Participants

Participants included 36 learners who have either studied Mandarin or have lived in mainland China. Participants were recruited through word of mouth, first targeting students and former students of Beijing Language and Culture University. Each participant interested in joining the study was initially sent an email detailing the requirements of the study; if they agreed, they were then asked to fill out a consent form. Participants who agreed to participate were then divided into three groups based on self-reported proficiency: intermediate, advanced, and native speakers. As the population being sampled is very diverse in terms of language ability and history of instruction, there is no objective measurement for language level available. Many high-level speakers have lived in China for an extended period without receiving any instruction, while some have years of Mandarin instruction but retain an intermediate communicative level. However, self-reported language level has been found to positively correlate with objective measurements ($r > .5$; Hakuta & D'Andrea, 1992; Luc & Bialystok, 2013). There is no beginner group as we believed their pronunciation is unlikely to be consistently intelligible; in addition, beginners may struggle with even the basic vocabulary required. Intermediate is likely the earliest level where they are able to produce full sentences and read Chinese characters with some ease.

Data Collection

Data collection took place in anytime-anywhere environments. Participants recorded themselves reading a list of True/False (T/F) statements adapted and translated from the list used in Derwing et al.'s (2000) study and later used again with Google Voice Typing by McCrocklin (2019). Due to cultural and linguistic differences, some of the translated statements created sentences that contained vocabulary that intermediate or even advanced users would be unlikely to know. Therefore, after initial translation, three L1 Mandarin speaking language teachers identified vocabulary they believed an intermediate Mandarin student would be unlikely to know. Interestingly, these vocabulary items were also deemed to belong to low frequency bands (less common in Mandarin Chinese) as per the ranked frequency by the University of Leeds Mandarin Corpus (University of Leeds, 2021). These rejected or infrequent items were then replaced with more frequently used synonyms (e.g., “grocery store”, 杂货店, was changed to “store”, 商店) or replaced with semantically similar words (e.g., “people play *baseball* with a piano” was changed to “people play *soccer* with a piano”; note that these examples are deliberately nonsensical, as will be discussed later). The new statements were then verified as suitable for intermediate students by the L1 Mandarin language experts (two professional language teachers with experience teaching Mandarin, and one expert in linguistics, all native Mandarin speakers), indicating that the target Mandarin learners should have little difficulty recognizing the target vocabulary.

Recording took place on a Moodle website (a learning management system or LMS), as its recording software has proven adequate for measuring comprehensibility in a related study (see Chapter 4). Participants were given instructions for using Moodle’s recording software, and were required to practice using it at least once. However, when recording the T/F statements used in the study, they were only able to record themselves once.

These recordings were then played for GT in a laboratory environment. Although there may be some minor loss of fidelity by playing recordings rather than having the participants speak to GT directly, this method offers several benefits. First, GT is difficult to access in China and requires the use of an often-expensive private virtual private network (VPN). Second, as participants would be recording at their leisure, we would be unable to control for participant issues such as if they chose to read to Google Translate multiple times or fail to take appropriate screen shots (required for the analysis of intelligibility). Cases where the recordings are of poor quality (e.g., noisy background) were discarded. In total, 2088 True/False (T/F) statements were available for analysis.

The list of T/F statements and their English translations can be seen in Appendix B. Similar to their English counterparts, all Mandarin translated sentences contain high frequency words and simple syntax. The True/False nature of the sentences are important as the false sentences may reduce Google's ability to predict the participant's meaning. That is, if GT's ASR is predicting what the user is saying (e.g., guessing based on context) rather than truly listening to what the user is saying, the nonsensical sentences would negatively affect the ASR's accuracy.

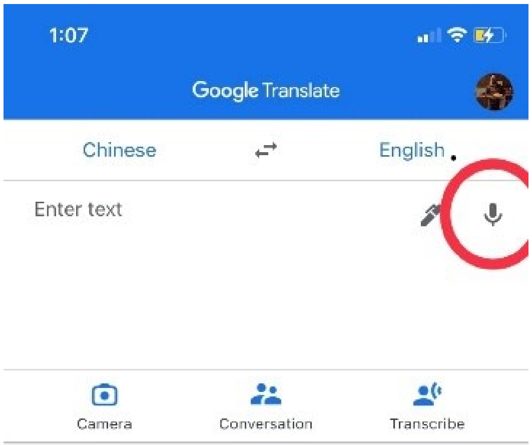
Data Analysis

Each statement produced by the participants was played out loud to an iPhone 12 using high quality speakers; after testing various available microphones, the iPhone 12 test recordings had the highest fidelity for ASR processing. The iPhone was connected to the internet via Wi-Fi, and the iOS Google Translate app was used for ASR testing. For each sample, the researcher pressed the microphone icon on the top right of Figure 8.

Following Derwing et al.'s (2000) methodology, each sentence was given a recognition score (the number of correctly recognized words as a percent of total words). As both Derwing et al. (2000) and McCrocklin (2019) found that there was near 100% accuracy with L1 English speakers, it is expected that Google will be similar with L1 Mandarin speakers, but that it will struggle more with L2 Mandarin speakers regardless of proficiency.

Figure 8

Google Translate Example Screenshot



To answer research question 1, “does Google Translate recognize Mandarin language speakers at a commensurate level to their language level and will there be an effect for proficiency?”, a one-way ANOVA was used to determine whether there are significant differences in recognition scores between groups (intermediate, advanced, and L1). To answer research question 2, “does GT’s ASR provide transcriptions that are adequate representations of what is said as to be used for signalling miscommunication and feedback?”, a random sample of 50 T/F statements at each level (intermediate, advanced, and native) was transcribed by three L1 Mandarin speakers, and then compared with GT’s transcriptions. A mixed-model ANOVA was used to determine whether there is a significant difference between GT’s recognition scores and the three raters, and whether those differences can be predicted by language level.

Results

GT’s ASR

In total, 2088 True/False statements were transcribed using GT, compiled from 36 Mandarin users drawn from three self-rated language levels: intermediate, advanced, and native speaker. Each sentence was given a recognition score by a native Mandarin speaker, and a one-way ANOVA was used to determine whether there were any significant differences between the two groups. The means and standard deviations can be seen in Table 4.

Table 4

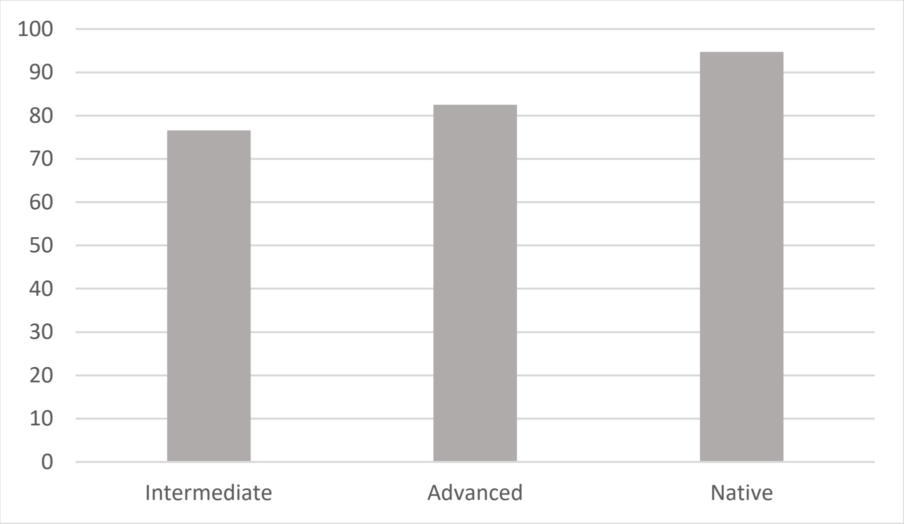
Rater Means and Standard Deviations

Rater	*M(/100)*	*SD*
Intermediate	76.61	10.50
Advanced	82.52	9.49
Native	94.67	3.94

The ANOVA showed a significant effect of speaker proficiency, *F*(2, 33) = 14.156, *p* <.001, η^2 = .46, and Bonferroni adjusted post-hoc pairwise comparisons indicate that native speakers had significantly higher recognition scores than both advanced (*p* = .004, *d* = 1.95) and intermediate speakers (*p* <.001, *d* = 2.28). The results are displayed in Figure 9 below. There was no significant difference between intermediate and advanced recognition scores.

Figure 9

Recognition Scores



Native Speaker Raters and GT's ASR

From the 2088 transcribed samples, 50 from each group (intermediate, advanced, and native speaker) were randomly selected for additional analysis. Three L1 Mandarin expert raters with language teaching experience provided recognition scores for all 150 sentences. Because there are so many possible variables including recording quality, speaker quality, and familiarity with a wide range of different accents (to name only a few), each rater was compared with both each other and GT using a mixed model ANOVA to determine both whether GT was significantly different from the human raters, but also whether there were any differences between the raters. The means and standard deviations can be seen in Table 5.

Table 5

Native Speaker Recognition Scores vs. GT Recognition Scores

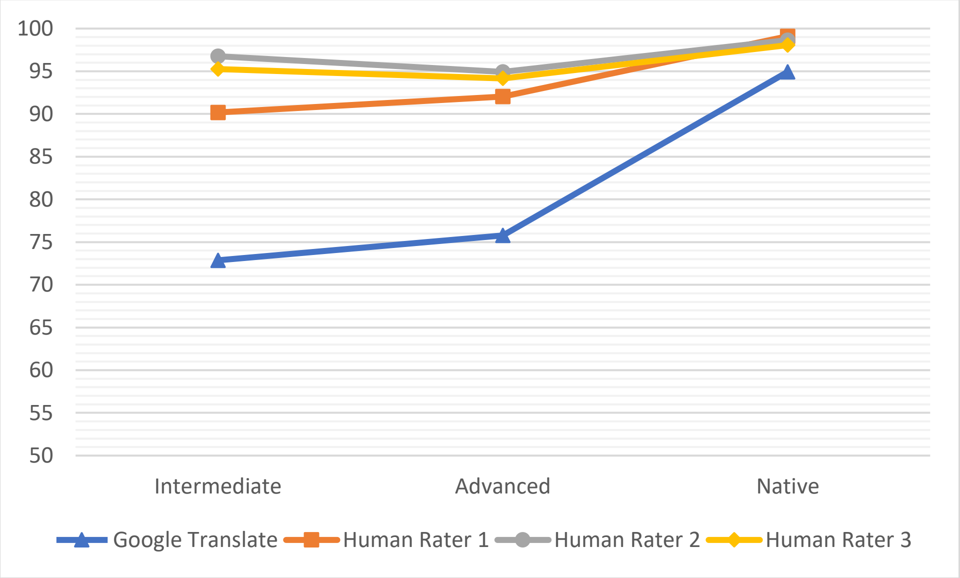
	Intermediate Group			**Advanced Group**			**Native Group**		
Rater	***n***	***M(/100)***	***SD***	***n***	***M(/100)***	***SD***	***n***	***M(/100)***	***SD***
Google Translate	50	72.87	29.30	50	75.76	29.60	50	94.94	11.24
Human Rater 1	50	90.18	15.71	50	92.05	15.48	50	99.10	3.13
Human Rater 2	50	96.75	7.93	50	94.93	12.90	50	98.65	4.76
Human Rater 3	50	95.26	9.02	50	94.16	13.29	50	98.06	4.84

The mixed-model ANOVA showed significant differences among the raters (GT, 1, 2, 3), $F(3, 441) = 43.61$, $p <.001$, $\eta^2 =.23$. Bonferroni adjusted pairwise comparisons revealed a difference between the recognition scores of GT and all three human raters ($p < .001$) with large effect sizes, but also between Rater 1 and Rater 3 ($p = .012$, $d = .17$) albeit with a small effect size.

The interaction between rater and proficiency level was also significant, *F*(6,441) = 6.70, *p* = <.001, η^2 =.08 indicating that the recognition scores varied according to the proficiency of the speaker. Bonferroni adjusted pairwise comparisons indicate that GT's recognition scores for intermediate and advanced speakers were significantly lower than that of the native speaker raters (*p* < .001), but that there was no significant difference for native speaker recognition scores (*p* = 1.0). The results are summarized in Figure 10.

Figure 10

Human Rater and GT Recognition Scores



Note: The vertical axis ranges from 50-100 to highlight rater differences

In summary, there was a consistent significant difference observed in GT's ASR's recognition scores between native and non-native speakers, regardless of their self-reported proficiency level (intermediate or advanced). GT's ASR was not significantly different than the expert L1 Mandarin raters when listening to native Mandarin speakers, but it was again significantly less accurate with non-native speakers.

Discussion

This study had two research questions: (1) Are there proficiency effects in speech recognition (i.e., does GT's ASR recognize Mandarin Chinese speech at a rate commensurate with their proficiency level), and (2) does GT's ASR provide transcriptions that are adequate representations of what is said? The results indicate that GT's ASR has difficulty transcribing non-native Mandarin speakers regardless of their proficiency level, and consequently the answer to both questions is that GT's ASR recognizes and accurately transcribes L1 Mandarin speech, while it does not recognize or accurately transcribe intermediate or advanced speakers nearly as well.

L1 Mandarin Speakers and L2 Mandarin Speakers

Regardless of actual language level, these results strongly indicate that L1 Mandarin speakers using GT's ASR can expect highly accurate results with up to 95% accuracy, and L2 Mandarin speakers can expect significantly less accurate results with around 80% accuracy, regardless of language level. These results align with the results seen more than 20 years ago in Derwing et al.'s (2000) study using Dragon NaturallySpeaking, which was able to transcribe 90% of L1 English productions accurately, but only 73% of L2 English productions. Although GT's ASR has a slightly higher accuracy rate than Dragon NaturallySpeaking (73% and 80% respectively), the difference is not large. However, in a much more recent study that tested Google Voice Typing and Windows Speech Recognition in English, McCrocklin et al. (2019) found that Google was 92% accurate for L1 speakers and 88.6% accurate for non-native speakers in a task similar to the one in this study, using the same target sentences as Derwing et al.'s (2000). Therefore, the results in this study may be affected by the choice of target language, suggesting that GT's ASR is not as capable in Mandarin as it is in English. We discuss below some of the possible reasons why GT's ASR seems to struggle with non-native Mandarin speakers.

Mandarin and English

English is the most popular language in the world and, as such, most ASR research cited in this study focus on English; interestingly, Google's head office is in a country where the principal language is English, the United States. These reasons alone may be why GT's ASR finds Mandarin L2 speakers challenging while a comparable Google program has no issues with English L2 speakers. However, I believe these results illustrate more complex concerns. There are two reasons for the lower ASR recognition scores in English: (1) tonal languages are inherently more challenging for ASR software and, as a result, it is possible that (2) the intermediate and advanced speakers do not produce tones with sufficient accuracy for the speech recognizer.

First, as mentioned previously, Mandarin is a tonal language that requires the speaker to change pitch for lexical meaning (Yip, 2002). This adds complexity for both the speaker and the listener (Halle et al., 2004). English, on the other hand, uses pitch change primarily for emphasis or to ask questions, which GT's ASR would likely not require to accurately transcribe English speech. For example, English questions often begin with a question word such as "what" and "how", which can be easily identified, regardless of pitch. Consequently, these results may indicate that GT's ASR system struggles with tones, and specifically the tones produced by non-native speakers.

Further, the ASR may have been further confounded by inaccurate tonal productions to begin with. That is, the tones produced by the intermediate and advanced speakers may not have been very accurate, and the raters may have been more forgiving than GT's ASR. Patel et al.

(2013) using fMRI (functional magnetic resonance imaging) found that native Mandarin speakers can still understand monotone (flattened pitch) sentences correctly, albeit by using additional cognitive resources. The expert raters may have understood (and therefore, correctly transcribed) the non-native speakers, despite the participants' incorrect tone productions. Other research (involving English) has also found that raters can sometimes insert missing phonological information (Strachan & Trofimovich, 2019), further suggesting that the native-speaker raters in this study might not have required the participants' correct tone production to understand them. Therefore, it is possible that the ASR software is simply not as intuitive as the native speaker raters when it comes to missing or incorrect phonological information.

To summarize the results reported here, GT's ASR is able to transcribe native speech very accurately, at a level commensurate with the expert raters, but not non-native speakers, regardless of whether they were intermediate or advanced Mandarin speakers.

Participants' Proficiency: Intermediate or Advanced?

The results indicated that GT's ASR recognition scores were significantly lower for L2 speakers, regardless of proficiency. One possible explanation for why there was no significant difference in recognition scores between L2 speaker proficiency levels may be that the intermediate and advanced learners were not accurate in their self-reported ratings (either the intermediate were too advanced, the advanced were more intermediate, or somewhere in the middle). However, I believe this is not the case. In Derwing et al.'s (2000) study, by carefully transcribing the phonemic data of the input, they were able to determine that there was no correlation between phonemic errors and the software's recognition scores, which aligns with the results of this study. In our findings, the expert raters gave both intermediate and advanced recognition scores similar to the native speakers, while GT scored the L2 speakers significantly lower. Considering that the number of phonemic errors likely negatively correlates with proficiency (the less proficient speakers will likely have more errors), our results likely indicate that there is no correlation between the number of tonal or phonemic errors and the ASR's results, just as Derwing et al. (2000) found in their research, *mutatis mutandis*.

Another explanation may be that the expert raters themselves were too lenient. Raters are known to be biased when they are familiar with the L2 accent in question (Carey et al., 2011), speak the test taker's L1 (Winke & Gass, 2013), or have general experience with L2 speech (Kennedy & Trofimovich, 2008; Saito et al., 2016). As expert raters, they certainly have some experience with L2 speech, and it is possible the human raters may fall into at least one of the above categories. Therefore, the possibility exists that GT's ASR was actually more objective and accurate than the expert raters. However, to determine whether there was human rater bias in these findings is beyond the scope of the current study.

Despite the significantly lower scores for intermediate and advanced speakers, GT was still able to recognize 77% of intermediate speaker productions and 83% of advanced speaker productions. This may indicate that intermediate and advanced learners can mostly use GT's ASR transcriptions as feedback, and that despite its flaws, it may still be an effective language learning tool, especially because reliable, accurate transcriptions are only part of what may make GT'S ASR helpful for language learning, as I outline below.

ASR accuracy and its effect on learners

GT's ASR was able to correctly transcribe L2 Mandarin speech around 80% of the time. As mentioned at the beginning of this work, Derwing et al. (2000) had two criteria for whether ASR software could be useful for language learning: it needs to understand language and identify errors at a rate similar to an L1 speaker. The results of this study indicate that GT is not up to this task in a Mandarin language learning context. However, this should not be taken as the definitive answer in regard to using GT or ASR software in general for Mandarin learning. For starters, Moussalli and Cardoso (2019) found that a certain amount of rejection (i.e., when the ASR does not transcribe the production accurately) could motivate learners to increase their interaction with the ASR software using interactionist strategies such as repetition and reformulation.

In her seminal text, Chapelle (2001) outlines seven criteria for adopting CALL tools, including reliability and learner fit, authenticity and generalizability, operationalization of learning conditions, interactivity, meaningful use of abilities, positive impact, and practicality. Strictly following these criteria, GT's ASR is a bit of a mixed bag. GT's ASR software is interactive, can presumably have a positive impact on language learning by providing access to corrective feedback, and is practical as it is ubiquitous on all platforms (computer, tablet, phone). However, in terms of reliability, accuracy, and positive impact, a recognition score of only 70-80% could still easily be frustrating, something that has been found in previous ASR studies (McCrocklin, 2016). In addition, most learners respond positively to their interactions with any ASR software, and in general, ASR software has been found to be an effective language learning tool across contexts (Dizon, 2020; Liakin et al., 2015; Van Lieshout & Cardoso, 2022). This may indicate that reliability and authenticity are important, but that other factors such as interactivity, meaningful use of abilities, positive impact, and practicality may play larger roles in the CALL process (e.g., Moussalli & Cardoso, 2019).

Considering Chapelle's (2001) seven criteria, GT's unique ability to function outside the classroom may increase its usefulness even more because it fully allows the learner to take control of their learning, interact meaningfully as much as they want with the software, and develop personalized strategies when using it. Although it may be frustrating for users to learn that GT's ASR only accurately transcribes 80% of an L2 speaker's speech (McCrocklin, 2016), there is some evidence that frustration impacts learning in human-computer interactions less than boredom does (Baker, D'Mello, Rodrigo, & Graesser, 2010). That is, perhaps GT's ASR is so much fun to use that it does not matter if its frustrating. For now, however, how enjoyable GT's ASR is to use is beyond the scope of this study.

This study also asked whether GT could reliably transcribe learner speech, and although 80% is respectable, it was nonetheless significantly less accurate with L2 speech than with native speaker speech when compared with the expert rater data. To summarize, although GT's ASR remains promising, the results of this study cast doubt on GT's ability to interact effectively with L2 speakers for the purposes of learning Mandarin Chinese.

Limitations

This study has several limitations that need careful consideration in future research. One clear limitation is that the learners all self-rated their language level. Despite research finding self-reported language levels correlating well with objectively measured levels ($r > .5$; Hakuta & D'Andrea, 1992; Luc & Bialystok, 2013), there remains the possibility that some learners may underestimate their own language ability, while others may overestimate it. However, it is also possible that the self-reported language levels were accurate considering that GT's results show that advanced participants were slightly (albeit not significantly) higher than intermediate, and that the native speakers' ratings showed significantly more variability for intermediate speakers than for advanced or native speakers. In a future study, the inclusion of objective proficiency measurements such as the HSK (the Hanyu Shuiping Kaoshi, a Chinese language test similar to IELTS or TOEFL) may be useful.

Another limitation is that every participant used their own microphone and computer to audio-record their sentences. Participants were living all over the world, including Canada, China, Italy, France, Japan, and Brazil, to name only a few locations. There was no possible way to control for hardware with such a diverse sample. Even if it were possible to invite the participants to a laboratory, data were collected in 2021 in the middle of the COVID-19 pandemic, and meeting participants in-person would have been inappropriate. However, I argued in this study that one of GT's strengths is that it can be used anytime-anywhere on any platform. It would be disingenuous of me as a researcher to then collect data only from the highest performing microphone and computer combination when that is a relatively rare use case. I believe this limitation allows for more realistic data because users were allowed to use whichever device in whichever way felt comfortable – it constitutes a "pedagogical reality" (Erlam & Tolosam, 2022) that clearly reflects the anytime-anywhere SRL learning scenario adopted in this study.

Last, one possible limitation not explored in this study is that the ASR is limited in its ability to provide feedback: it can only provide what Lyster (2002) calls *negotiation of form*, that is, it provides text that signals learners when, where, and how to self-repair, but it does not provide options for recasts or rephrasing, and thus cannot fully negotiate for meaning. However, I nonetheless still argue that the tool makes an effective interlocutor within an interactionist approach, and that although limited, with help from the user and GT's other functionalities (translation, not explored in this dissertation, and TTS, which was explored in Chapter 2), it can create a negotiation for meaning experience with the participant. For instance, when the ASR's feedback is insufficient, the TTS can still provide what is effectively a recast if the target sentence is presented, or when combined with the translation function, a learner can manipulate the target L2 sentence by choosing to translate a “reformulated” phrase from their L1. However, using all of GT’s functionalities in this manner is beyond the scope of this study.

Future Research

This study unlocks multiple avenues for important future research in using ASR for language acquisition. By adapting Derwing et al’s (2000) methodology, future studies should consider using the same T/F statements but translated into new languages to create a more robust picture of ASR across languages and applications. GT and its ASR specifically are also interesting targets for future research. After all, GT is free and available on theoretically every platform, both online and offline. Research might consider comparing the CALL and MALL (its mobile counterpart) experiences as well as both the online and offline experiences in a variety of locations and languages.

Considering GT, our results show that native speakers are able to transcribe all three speaker levels (intermediate, advanced, and native) very accurately, while GT struggles with L2 speakers. Perhaps this may indicate that GT’s ASR’s transcription abilities are more comparable to an L2 language speaker than a native speaker (i.e., comparable to the typical interlocutor that language learners interact with in classrooms). Future research may want to consider comparing GT’s recognition scores with intermediate and advanced Mandarin speakers providing the ratings instead of only native speakers. They may discover, for example, that GT’s abilities are analogous to a classroom L2 learner, who are regularly used as effective language partners within an interactive framework (Long, 1996; McDonough & Mackey, 2008). If its ratings are similar to an L2 speaker’s, this may also offer an explanation as to why learners enjoy using ASR software despite some frustration.

Future research may also directly address the limitations of this study. For example, if objective language assessments were used to measure proficiency, differences between the intermediate, advanced, and native speaker groups may have been more obvious. Further, verifying whether a high-end microphone and computer in a laboratory environment produces different results, or whether different use-cases such as smartphones vs. computers shows any noticeable differences, would also be interesting.

One last direction for future research would be to determine what aspects of L2 speech GT's ASR struggles with. In the case of Mandarin, there may be issues with tones, but considering Derwing et al.'s (2000) original results with an English language ASR, the answer is likely more complex. It would be of value to the field to determine what aspects of L2 speech give ASR software so much difficulty.

Conclusion

This study's goal was to address how GT's ASR recognizes Mandarin speech at different proficiency levels, and whether the transcriptions provided by the speech recognizer were adequate representations of what was said. These results indicate that GT can recognize native speaker speech at a level similar to an L1 speaker, but that it struggles to accurately transcribe L2 speech, regardless of the L2 speaker's proficiency level. This has important ramifications in CALL and specifically research around using ASR for language learning. Although we might assume the software will improve over time, for now, specifically for Mandarin language learning, we can assume that the ASR will be accurate most but not all of the time.

Learners and their language instructors who rely on ASR should attempt to address possible frustration in advance, and ideally continue to provide opportunities to practice with other L2 or native speakers to shore up the learner confidence when frustration peaks. However, learners can still use GT’s ASR as an interaction partner and receive valuable (but not always accurate) corrective feedback. It may not be as accurate as a native speaker, but unlike a native speaker, the technology can be used as many times as necessary, and will never stop paying attention or become frustrated with the learner. This study does not invalidate ASR’s usefulness as a pedagogical tool, but instead, it should temper users’ expectations of how effective ASR might be in true anytime-anywhere conditions.

Chapter 4

Learning pronunciation in an online, self-regulated environment with Google Translate: Focus on Mandarin tones

Research has shown that practice is crucial for second/foreign language (L2) learning and that it has a positive effect on L2 acquisition (e.g., Ellis, 2002; Swain, 1995). However, classrooms have spatial and temporal limitations that restrict practice opportunities (Collins & Munoz, 2016). Technologies such as automatic-speech recognition (ASR) and text-to-speech synthesizers (TTS) allow for both speaking and listening practice (e.g., Van Lieshout & Cardoso, 2022). ASR found in software such as Google Translate (GT) listens to oral productions, automatically interprets them, and transcribes them. Another GT feature, TTS, does the reverse: it converts textual input into audio. These technologies allow for anytime-anywhere speaking and listening practice that could be a boon to learners, especially those attempting to learn in an autonomous, self-regulated manner.

However, language learning online can be challenging (De Paepe et al., 2018). The online environment can be distracting, and even with some direction from instructors, students must learn to work autonomously and engage in self-regulated learning (SRL; e.g., Andrade & Bunker, 2009). SRL requires that learners develop strategies for completing their work and self-monitoring, which itself requires targeted scaffolding such as instruction and supervision for success (Winne, 2018). Without this scaffolding, students are less likely to succeed in their learning and will be unable to take full advantage of this new learning environment.

In an attempt to address the needs of language learners and the challenges associated with in class language learning, this study adopted GT's translation and its built-in ASR and TTS features to examine the acquisition of an L2 phonological system in an online environment (using Moodle as the platform) designed to foster both language learning and SRL. Our target for instruction is Mandarin Chinese tones (and associated vocabulary), defined as changes in pitch that affect lexical meaning (Yip, 2002). For example, consider the words for *horse* (/ma/, pronounced with a falling and rising pitch) and *mother* (/ma/, pronounced with a high pitch), which have identical segmental content (i.e., /ma/), but are differentiated by their tones. Mandarin was chosen as the target L2 because, although it is the most widely spoken tonal language in the world, it has not received the same level of attention in comparison with languages such as English and Spanish (Yang, 2021).

Mandarin has four tones, which are usually described using a 5-point pitch scale (Chao, 1968; where 5 is high and 1 is low; tones will be labeled using this system. For example, the 2nd Tone is T35 and the 3rd Tone is T215).

1st Tone: 5 → 5 (a high, even pitch)

2nd Tone: 3 → 5 (a rising pitch)

3rd Tone: 2 → 1 → 5 (a falling and then rising pitch)

4th Tone: 5 → 1 (a falling pitch)

Tones are difficult for learners even from tonal L1 backgrounds to perceive and produce (e.g., Halle et al., 2004; Saito & Wu, 2014). For example, L1 Cantonese learners of Mandarin often mistake both the T55 and T51 tones for the Cantonese T55 (Saito & Wu, 2014). Those from non-tonal L1 backgrounds, such as English and French speakers, struggle because they do not process tones for lexical information, and so although they can perceive that two tones sound different, they are not able to identify those differences (Halle et al., 2004).

This study pilots an online, SRL environment designed for the instruction of Mandarin tones and associated language features. As such, it intends to address the feasibility of key components of the proposed environment, particularly whether: (1) the proposed technology-based SRL approach would be perceived positively by participants, (2) the participants would develop their own SRL strategies, and (3) a challenging phonological structure such as Mandarin tones can be acquired using the proposed GT-based technologies (i.e., translation, ASR and TTS) for input and output practice. The results of this pilot will inform a larger research project and provide online language instructors with crucial data for interpreting how their students learn in online, self-regulated environments.

Background

Scaffolding in SRL and Online Language Environments

Autonomous and self-regulated learning do not necessarily require the learner to be alone (Godwin-Jones, 2011). Teachers and CALL tools can scaffold independence and autonomy and foster language learning by providing tools and strategies that students can take advantage of in their learning (Godwin-Jones, 2011). For this study, we define autonomous learning as learners being responsible for their own learning (Andrade & Bunker, 2008) through the development and use of SRL strategies (Zimmerman, 1998). Zimmerman's (1998) cyclical model for understanding how self-regulated learning includes: goal setting and strategic planning, strategy implementation and monitoring, strategic outcome and monitoring, and self-evaluation and monitoring. Zimmerman argues that, whenever a learner participates in their own learning, they are activating these processes. Most SRL research agrees that SRL learning happens most effectively when scaffolded to reduce the challenge and accompanying frustration when learning by oneself (Winne, 2018).

SRL researchers often apply a sociocultural framework using Vygotsky's Zone of Proximal Development (ZPD) (Vygotsky, 1978; Winne, 2018). The implication is that when a learner within the ZPD receives assistance from outside sources, they are able to learn beyond what they could on their own. An online learner with scaffolding and assistance (within the ZPD) should be more motivated and successful than those without, who are consequently more likely to become demotivated (Gibbons, 2002; Winne, 2001). Although theoretically some learners could operate entirely autonomously (with or without assistance), most research agrees that this is unlikely. For example, in a study of 40 Thai learners of English, Vandijee (2003) found that even the most autonomous learners required some degree of structure or scaffolding to be successful.

SRL research emphasizes the importance of help from outside sources including assistance for developing learning strategies and providing effective feedback (Dabbagh & Kitsantas, 2005). Dabbagh and Kitsantas (2005) specifically argue for the role of web-based pedagogical tools such as learning management systems (LMSs) to scaffold for SRL and provide this outside help. Looking at WebCT, a web-based course management system similar to Moodle, they found that multiple online tools (such as those for content creation, administration, communication, and assessment) all contribute to the development of SRL strategies in students. Studies in second language acquisition have also found that SRL can be fostered in these online environments. Dembo et al. (2006), for instance, concluded that a learner’s access to scaffolding and use of self-regulation strategies correlate with their language success, and that this scaffolding can be provided through an online experience. Thus, online scaffolding design should address both SRL strategies and language learning, and it should take place during any interaction with the language (Ellis, 2002), whether with teachers or other learners (Swain, 1995), or even with computers (Chapelle, 2005).

Computer-assisted language learning and Mandarin tones

Chapelle (2005) argues that, in a usage-based theoretical framework, interaction with computers is an effective replacement when authentic interaction is unavailable, particularly because computers allow for more opportunities for language input and output practice, which serve to increase the frequency and salience of the target forms - strong predictors of language acquisition (Ellis, 2002). For pronunciation instruction, technologies such as ASR and TTS allow for both input and output practice, in and/or outside of the classroom.

Research into the benefits of ASR for language learning has shown mixed results. It can provide learners with unlimited output practice (e.g., Liakin et al., 2015; Neri et al., 2008; Van Doremalen et al., 2016), but it is not always able to understand and interpret learner speech correctly (see Chapter 3 and Derwing, Munro, & Carbonara, 2000). There is an element of untrustworthiness, and this could lead to frustration and reduced motivation for learners. Like ASR, TTS also has been found to enhance language learning by increasing opportunities for oral input (e.g., Liakin et al., 2017). There is some concern as the oral output has an artificial quality to it (Bione & Cardoso, 2020; Cardoso et al., 2015), but it has nonetheless been found effective for pronunciation instruction (e.g., Cardoso, 2018), particularly when TTS is combined with ASR, as is the case with GT.

Google Translate and Mandarin Tones

GT is a freely available, web based and downloadable software for translating up to 103 languages (Alphabet Inc., 2020). In addition to its primary function as a translator, GT has both TTS and ASR functionality for approximately 50% of its available languages including English and Mandarin. In a recent study, Van Lieshout and Cardoso (2022) found that 30 participants using GT for practice were able to learn (i.e., recall and orally produce in posttests and delayed posttests) 10 Dutch phrases with high intelligibility, high comprehensibility, and low accentedness. The authors further evaluated user perceptions of GT using four criteria: learnability (ability to promote learning), usability (practicality, ease of use, convenience), motivation, and willingness to use. They found that the participants had overall positive ratings in terms of these perception markers. This landmark study is the first to use GT's translation, TTS, and ASR functionalities in a self-regulated learning environment. Mroz (2020) also found similar results using ASR in a GT-enhanced learning context, showing that ASR users significantly outperformed non-ASR users on measures of intelligibility. The authors encourage further research to validate GT's (or any similar applications') pedagogical potential in different contexts and with other languages.

The target language for this study, Mandarin Chinese, is a difficult language to acquire (Yip, 2002), as previously mentioned. Accordingly, learning Mandarin and its four tones is complicated by the learners' L1 as they often attempt to map tones to existing mental categories leading to inaccurate perception and production (e.g., Halle et al., 2004; Saito & Wu, 2014). For successful acquisition of such a challenging structure, practice and instruction are essential. Celce-Murcia et al. (2010) outline four key steps for pronunciation instruction starting with developing the ability to hear and perceive (e.g., aurally distinguish) the target sounds through listening discrimination tasks, and slowly progressing to controlled practice with feedback, guided practice with feedback, and last, communicative or unguided production with further feedback. This approach is very similar to that outlined by Zimmerman (1998): practice, receive feedback, and repeat while reducing instruction or scaffolding each iteration. In this study, we hypothesize that GT (or any similar translation tool) can provide learners with some of the tools necessary to support an anytime-anywhere, autonomous, and technology-enhanced learning environment to optimize the acquisition of Mandarin tones.

The Current Study

This study examines the use of a GT-based online environment for autonomous pronunciation learning focusing on a challenging linguistic structure: Mandarin tones. Specifically, its main goal is to assess the pedagogical feasibility of using GT and its built-in features (translation, ASR and TTS) in a self-regulated learning environment for a future larger study. For our study, we conceptualize SRL according to Zimmerman's (1998) cyclical model; as such, we assume that SRL is activated whenever goal setting (to learn Mandarin tones), strategy implementation (e.g., students listen to TTS and practice with ASR), and self-monitoring (e.g., deciding when and how to learn) are implemented. The following research questions (RQs) guided this study:

1) Does the proposed technology-enhanced SRL pedagogy lead to the learning (operationalized as achievements in comprehensibility) of Mandarin tones?
2) How do participants self-regulate their learning?
3) How do participants perceive the proposed GT-enhanced pedagogical environment (via translation, TTS, and ASR) for learning aspects of a foreign language phonology (tones)?

Method

A custom Moodle site designed to foster SRL and the acquisition of Mandarin was used for instruction and data collection. Moodle and other learning management systems have been shown to support online learning and, more importantly, SRL (e.g., Dabbagh & Kitsantas 2005; Darasawang & Reinders, 2011). We chose Moodle because its object-oriented design allows for easy instructional building, and the platform includes essential elements required for this study, including built-in voice recording capabilities. Further, Moodle is open-source, widely popular, and already used by multiple institutions in Canada (where the study took place) and in other countries, thus leading us to believe that the participants were likely to have some previous experience with the platform.

The tasks for this study were designed to incorporate self-regulated pronunciation instruction, following Chapelle's (2001) and Chapelle and Jamieson's (2008) criteria for CALL tasks. Specifically, tasks were designed to provide opportunities for students to engage with language, personalize the experience, attempt to provide interactions reflecting real-world experiences, and to be practical (Chapelle, 2001; Chapelle & Jamieson, 2008). First, each participant was prompted to complete three "Study Practice" sections before moving on to five targeted assessments, four of which were used in this study (the fifth, a listening discrimination task, will be analyzed in a later study). Each assessment was designed to test a different aspect of instruction: translating, listening to the TTS, speaking using the ASR, creating complex spoken sentences, and then communicative-like practice by introducing themselves with multiple spoken complex sentences. The five participants included in this study were the first to finish the entire learning process (see forthcoming discussion for details).

To establish tone acquisition (RQ1), we adopted a combination of interview analysis (see below) and comprehensibility ratings, one of the quantitative measures for L2 pronunciation (see Yang, 2016 for the use of a similar measure to assess tone development in Mandarin). Comprehensibility is defined as a listener's perceptions of understanding a speech sample, using scalar ratings of how easily they understand speech (Munro & Derwing, 1995). To answer the remaining research questions, participants were interviewed via nine open-ended questions that probed how they self-regulated their learning (RQ2), and their perceptions of GT (i.e., how they view the technology in terms of usability, learnability, motivation and willingness to continue to use it in their L2 learning endeavors, as will be discussed later), their overall learning experiences, and what they learned about Mandarin and Mandarin tones (see Appendix C for a list of the interview questions). These interviews were then coded thematically using a phenomenological approach following Saldaña's (2009) methodological recommendations for qualitative coding.

Participants

There are currently 39 students enrolled in the online Mandarin class (the custom Moodle website). The first five to finish instruction volunteered to be further interviewed and were remunerated 20 Canadian dollars for their time. Their demographic information is summarized in Table 1. All self-rated their English as native or native-like, and none reported any proficiency in Mandarin or in a tonal language – a condition to participate in the study. This explains the absence of a pretest, whose implementation would require some knowledge of Mandarin or pinyin (Romanization of Chinese characters based on pronunciation) from the participants, in addition to pitch change for tone production. Last, all had some previous experience with GT as a translation tool. Table 6 illustrates the demographics of the five participants (their names are fictitious).

Table 6

Demographic and Time-on-Task information

	Hunter	Lola	Bruce	Phoebe	Gary
Age	26-35	18-25	26-35	45+	26-35
Country of Birth	Canada	Canada	Canada	Brazil	Canada
First Language	English	French	French	Portuguese	English/French
Time on Task	13 hrs	4.5 hrs	4 hrs	8.5 hrs	8 hrs

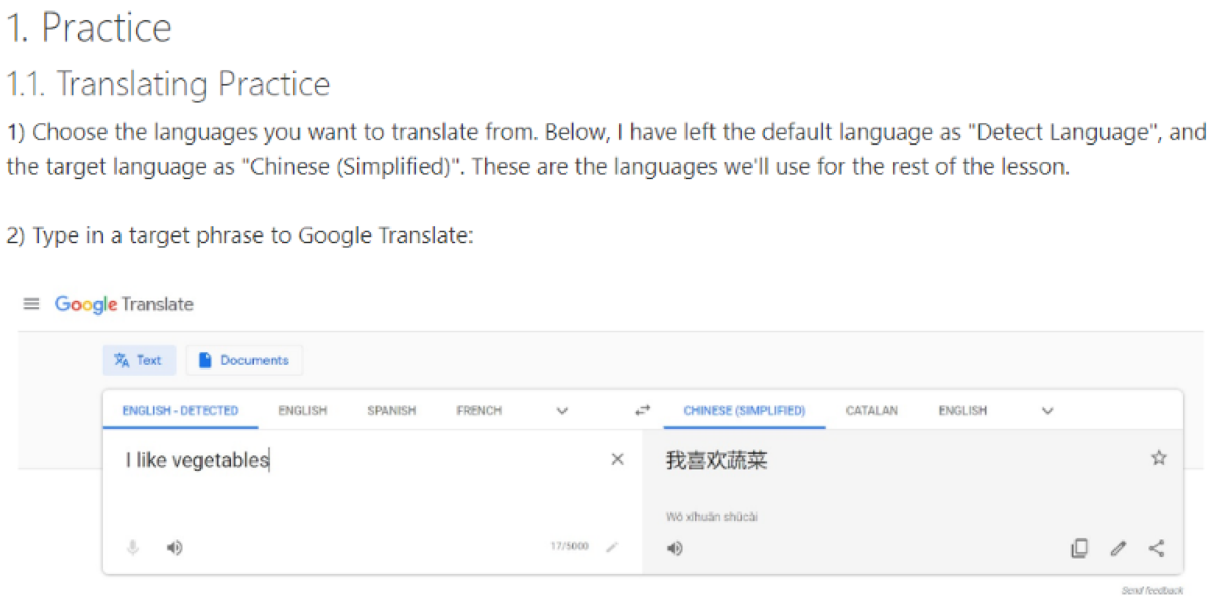
Procedure

Instruction took place on a Moodle website, which was used as a platform for instruction and data collection. First, participants were asked to sign an online consent form, followed by a demographic questionnaire. As the website was possibly a new experience for some learners, learners were then asked to practice using the submission system followed by a practice quiz where learners would practice recording themselves and uploading their recordings. These activities were strictly to practice using the system itself, and no data were collected for analysis.

Instruction consisted of three Study Practice Moodle “books”. Each book had several short chapters designed to foster self-regulated learning strategies and GT experience while teaching basic Mandarin vocabulary and their associated tones. Each chapter required some knowledge from previous chapters, and participants had to build on that knowledge while practicing various learning strategies. The first Study Practice provided a tutorial for using GT for translating, listening (TTS), and speaking practice (ASR); it also asked learners to translate some basic sentences into Mandarin. Here, they learned to use the translate function to learn Mandarin vocabulary (e.g., by translating a provided phrase such as “I like vegetables”), the TTS function to practice listening, and the ASR function to practice speaking and receive feedback (e.g., by comparing the orthographic ASR output with their intention). Figure 11 shows an example of the translation exercise, where participants were taught first how to translate a word or phrase, and then how to control the input language (either English or Mandarin).

Figure 11

Translation exercise



1. Practice

1.1. Translating Practice

1) Choose the languages you want to translate from. Below, I have left the default language as "Detect Language", and the target language as "Chinese (Simplified)". These are the languages we'll use for the rest of the lesson.

2) Type in a target phrase to Google Translate:

I like vegetables

我喜欢蔬菜

Wǒ xǐhuān shūcài

The second Study Practice instructed participants in how to read and interpret pinyin, the Mandarin orthographic system taught to beginners that includes tonal information (e.g., mā, má, mă, mà) including some short examples (e.g., wǒ xǐhuān shūcài [I like vegetables]). Practice consisted of single and multi-character words. The final Study Practice instructed participants to form complete sentences by first asking them to read aloud some simple sentences (e.g. [I like vegetables]), and then asking them to create their own. Instruction was designed to take approximately one hour, but as the study was entirely self-regulated, participants were able to finish at their leisure, some completing instruction over several days.

The Study Practice books were followed by five “quizzes”: Translation quiz, Listening quiz, Speaking quiz, Sentence quiz, and Introducing yourself quiz. These quizzes were designed to assess learning, provide opportunities for practice and feedback, and prompt participants to move from more to less structured practice. The Translation quiz required that the participant translate simple vocabulary and then produce each out loud for assessment. The Listening quiz required participants to listen to recorded basic phrases and write them out. The speaking quiz assessed production by having participants say previously encountered Mandarin phrases in the Study Practice books. The Sentence quiz asked participants to create their own sentences by using the translate function. Last, the Introduction quiz required participants to create their own sentences to introduce themselves to a stranger and then say these sentences out loud.

Data were collected only from the post instruction quizzes. As mentioned earlier, no pretest data was collected as the learners had no experience with Mandarin Chinese or speaking/learning tonal languages. Accordingly, they had no ability to perform in the tasks (quizzes), an assumption that was confirmed by the low level of achievement (approaching zero) in their first assignment submissions. According to log data collected on Moodle, participants spent an average of 7.6 hours on instruction and practice divided over two to three days.

Comprehensibility Ratings

The website collected spoken data consisting of short phrases before and after instruction, as well as two quizzes that prompted full, participant-created sentences. Participants had to begin each recording themselves; consequently, practice attempts were not recorded. Productions not directly prompted from instruction (such as self-talking) were discarded. Two expert raters or judges, both L1 Mandarin speakers with language teaching experience, rated each production from 1 to 9 in terms of comprehensibility, as previously defined (1 is high comprehensibility, 9 is low). First, each rater rated three practice items so that they could understand the task and the construct being evaluated, and then rated 141 randomized phrases from all five participants. Cronbach's alpha indicated ratings reached an acceptable level of interrater reliability ($a = .74$), $r = .61$ ($p < .001$).

Interviews

The interviews were conducted after instruction (see Appendix C for the list of questions). As participants were allowed to complete instruction in an anytime-anywhere manner, there was no way to control for how long after instruction the interviews took place, and so they were administered at the participant's earliest convenience. They were asked nine questions adapted from Van Lieshout and Cardoso's work (2022) and were designed to elicit information regarding the three research questions. The interviews were then transcribed and coded thematically following a phenomenological approach (Saldaña, 2009), chosen to capture the perceived experience of participants in a recurrent and patterned system designed to foster SRL and language learning. Participants' perceptions refer to how they view the proposed autonomous learning environment and its affordances as a medium for language learning, including their awareness of its pedagogical value in terms of strategy, instruction and self-monitoring (for SRL), and usability, learnability, motivation and willingness to use the technology (for GT). First, the participants' phrases that broadly correspond with the three main themes addressed by this perception study (i.e., language/tone learning, SRL use, and GT use) were identified and coded. Next, initial codes were coded again for themes allowing the researcher to draw upon both the participants experiences and the literature to capture these patterns in their variety across their individual realities. The language learning themes developed from the qualitative coding were "tones" and "language" (a cover term for any language-related statement not related to tones). The SRL themes were "strategy", "instruction", and "self-monitoring". Finally, the GT themes were usability, learnability, motivation, and willingness to continue to use the tool.

Results

Mandarin Learning: Comprehensibility Ratings

Due to the number of participants, we were unable to perform sophisticated statistical analyses. However, the descriptive statistics for the comprehensibility ratings (out of 9) across the four tasks can be seen in Table 7. As mentioned previously, we do not report any pretest scores as the participants had no experience with Mandarin Chinese or tonal languages, requirements to participate in the study. As discussed earlier, the Speaking and Translation quizzes focused on short and "easy" phrases while the Sentence and Introduction quizzes prompted participants to produce more complex phrases and their own constructions.

Table 7

Comprehensibility Ratings by Participant: Posttest (/9)

	Hunter		Lola		Bruce		Phoebe		Gary	
	M	SD	M	SD	M	SD	M	SD	M	SD
Translation Quiz	3.21	1.72	1.69	1.08	2.10	2.51	4.00	2.10	5.60	3.27
Speaking Quiz	2.75	1.60	4.25	2.96	2.50	1.57	6.75	3.42	3.33	2.31
Sentence Quiz	4.17	2.78	2.83	2.14	5.50	3.27	2.33	1.03	6.50	2.17
Introduction Quiz	3.30	2.50	2.33	1.21	6.67	3.01	4.70	2.58	5.33	1.51
Average Score	3.36	2.15	2.78	1.85	4.19	2.59	4.45	2.28	5.19	2.32

Note: 1 indicates high comprehensibility (easy to understand), 9 indicates low comprehensibility (difficult to understand)

The average comprehensibility scores indicate that all participants became comparatively comprehensible after instruction, but their performance decreased as the quizzes required more complex constructions. Hunter, Bruce, and Gary specifically were more comprehensible in the Translation and Speaking quizzes than in the more complex constructions (Sentence and Introduction quizzes). Lola and Phoebe conversely were more comprehensible with more complex constructions, but in Phoebe's case, these results were not consistent. These ratings indicate that all participants achieved a certain level of comprehensibility after instruction, and that more complex constructions led to more variability in comprehensibility compared with more simple productions.

Mandarin Learning: Learners' Perceptions

As discussed earlier, the interviews were first coded into three central themes (i.e., Mandarin learning, self-regulated learning, and Google Translate), and then further divided into subthemes according to the research question and participants' own responses. Mandarin learning was coded into two broad thematic categories: tones and language. Tones and Language were separated as almost half of all comments about language learning concerned tones specifically (recall that the Language theme refers to any language-related comments that do not refer to tones). Comments were coded as language if the participants discussed language outside of tones, including general pronunciation (e.g., references to syllables, fluency), words, grammar, and orthography. The themes, subthemes, and examples can be seen in Table 8.

Tones

All participants reported that tones were difficult to perceive, defined here as one's ability to discriminate tones aurally (e.g., after listening to the Chinese word for *mother* /maT55/, did they incorrectly hear its tonemic minimal pair /maT215/ - *horse*, or the intended word?). Specifically, perception was possible though difficult, and production was more challenging than perception. Hunter, Lola, and Phoebe reported difficulty in initial perception. Hunter and Lola reported guessing when presented with two words such as mā (mom) and mǎ (horse) with different tones, while Phoebe found the abstractness of the tones challenging and, accordingly, was unable to form a mental picture of pitch moving up or down.

Table 8

Themes, subthemes, and examples

Themes	Subthemes	Example
Mandarin learning	Tones	I can recognize them a little bit better. Right? There are four tones. I know two pretty well. One is okay, one is ehh. (Gary)
	Language	So with the text to speech thing, I remember focusing very intently on trying to figure out where the sound was different from what I expected it to be given how it was written down. (Phoebe)
Self-regulated learning	Strategy	…my personal style with pronunciation is just unending practice. That's how I lost most of my accent in English. (Bruce)
	Instruction	Translation practice. Choose the languages. I think the instructions are very clear, very easy, very straight to the point. (Phoebe)
	Self-monitoring	It had instant gratification. If you said a word, and then Google Translate was like yes, that is the word you said. It felt like a victory. (Hunter)
Google Translate	Usability	So, I think it's one of those tools that, especially, like the more experience you had, the better it went. (Hunter)
	Learnability	I realized what the focus was and then what the purpose was and how Google Translate could help me achieve that focus. (Lola)
	Motivation	When you get constant feedback it's really motivating, right? But at the same time, you don't repeat it, right? (Phoebe)
	Willingness	I could see myself using it just for fun. Let's say I'm watching something, and I like how it's said, I can use that to try and learn it. I could use it for small things. (Hunter)

Overall, difficulty identifying and producing tones was compounded by multisyllabic units, as both Lola and Phoebe reported identifying word boundaries as difficult: “I cannot hear the endings of the words in Mandarin. I don’t know where the words end” (Phoebe).

In terms of production, both Lola and Gary found it “overwhelming”. That is, at first producing tones with their accompany sounds was challenging: “When I started producing tones, then I was like over, cognitive overload of tones” (Lola). However, as practice continued, Lola reported finding production easier (also evidenced by her higher comprehensibility ratings, reported in Table 6). Gary also found tone production overwhelming but did not show the same level of success. However, he did mention that some tones were easier than others, a phenomenon also mentioned by Bruce. However, neither could identify which tones were easier at the interview.

As all participants felt that production was challenging, they discovered individualized strategies to address these challenges. Some showed evidence of attempting to map the lexical pitch change to existing categories in their L1s. For example, Hunter attempted to write out each Mandarin phrase phonetically using English and not relying on the Mandarin based pinyin. In another example, Lola and Bruce used their musical experience: Lola pictured a musical bar to “see” pitch change, while Bruce (a singer) related learning tones to learning how to “growl” and produce other challenging musical sounds. Phoebe related her challenge with the abstract nature of pitch change in tones to the same challenges she faced with English pronunciation. Last, Gary believed that tones were particularly challenging because he was unable to compare them to either French or English in any meaningful way, and so relied on hand gestures to mimic pitch changes.

Language

Language coding offered insight into what participants found challenging outside of tone use. Phoebe and Gary reported that phrases and multisyllabic units were difficult, and their comprehensibility ratings in Table 6 seem to support their analysis. Phoebe felt that the biggest challenge was identifying word boundaries. Gary reported that fluency was a challenge and felt that GT provided poor input whenever it joined multiple words together.

Participants further felt that additional language instruction could have improved their language learning. Comments included recommending more explicit orthographic instruction (Gary), more examples provided of each tone (Gary), explicit grammar instruction (Phoebe), and explicit pronunciation instruction outside of tones (Hunter).

Self-regulated Learning

As indicated earlier, based on participant interviews and SRL literature, the SRL aspect of the research was divided into three subthemes: strategy, instruction, and self-monitoring.

Strategy

The most common chosen strategy across the participants was repetition. Bruce said that his language learning experience was centered around "unending" repetition, and Phoebe and Gary reported their focus was on listening and repeating as much as possible. Within the repetition, they would often use listening discrimination strategies to make the pitch changes clearer, followed by using GT's ASR to self-monitor. The exception was Gary, who although having completed instruction, did not use ASR to self-monitor unless specifically asked to. Instead, Gary listened to GT's TTS, and relied on his own ear to self-monitor production.

Instruction

When participants made particular reference to the website's instructional materials and its effects on learning, they were coded as "Instruction". Participants reported that instruction was necessary and felt that the short nature of each activity kept motivation high. Hunter and Lola felt having consistent, simple instruction was itself motivating. However, an issue with having provided instruction as such was that there were cases when the participants would have complications and not adapt unless instructed to. For example, Gary believed that searching outside the course for Mandarin learning strategies would have been cheating. Phoebe also reported that they did not want to deviate in anyway from the provided instruction, even when they had issues with tone production they felt were not addressed within the course. However, when instructed to, they were willing to exercise independence.

Coincidentally, participants reported higher motivation when instructed to create their own sentences. Both Hunter and Lola began creating their own examples early on without instruction because they felt they learned more that way. The two others reported the more control they had, the more they enjoyed the process. The self-paced instruction also put participants at ease as they adapted their session lengths to fit their personal learning styles. Hunter and Gary both took multiple sessions to finish instruction, while Lola, Bruce, and Phoebe finished quicker.

Self-monitoring

The feedback received from Google (via the ASR output, orthographically) seems to have impacted motivation and repetition. Except for Gary, who self-monitored using listening discrimination exercises and his own output, the remaining participants used GT's ASR to good effect. Hunter, Bruce, and Phoebe all felt its unbiased, immediate feedback motivated them to continue to try until they received a satisfactory (orthographic) response. However, they felt the ASR had a clear limitation: it lacked nuance, as it could only provide feedback in terms of whether its transcription matches their target output. It was unable to identify where pronunciation failed in any detail, and it could not provide explicit feedback on where to improve. Consequently, both the success and failure of the ASR would eventually lead to ceasing repetition. Lola and Phoebe reported they would cease repetition if the ASR transcript matched their goal. On the other hand, Phoebe, Hunter, and Bruce said that if they did not achieve satisfactory results quick enough, they would also cease repetition and either move on or end the session.

Google Translate

As discussed earlier, GT was coded and analyzed considering four subthemes: usability, learnability, motivation, and willingness of the participants to continue to use the application in their learning experiences.

Usability

Two factors that impacted GT's usability the most seemed to be its overall ease of use and previous experience with the platform. All participants reported that GT was easy to use, likely related to the fact that all participants were also very familiar with the technology. There were usability concerns with ASR specifically. Phoebe, who self-identified as having a strong accent, was constantly frustrated with what she felt were unfair responses: "C'mon Google. I mean, you want Google to understand you. It's almost like you have a person in front of you. Seriously? I'm saying it. Just get it." She believed that the technology in general was challenged by her accent, and that it did not matter whether she was speaking English or Mandarin. This translated to a lack of trust in the ASR. Even when the ASR would confirm Phoebe's output as correct, Phoebe would think it was coincidence. Lola also reported a lack of trust in the ASR, but for a different reason. Lola found that the ASR always understood her even when her productions were inaccurate. This led Lola to practice some phrases even after the ASR had reported their results as correct.

There was some criticism from Gary concerning the TTS. He felt that having more control over the speed of the voice would have helped, as he often found it "very, very quick." He also felt his learning would be improved by additional voices: "…hearing different people speak the different tones… it would have helped me learn them faster because, you know, the way that she (Google) might speak would be different from you because you're a male with a deeper voice."

Learnability

Overall, the participants felt that GT's ASR and TTS functions helped them learn. Specifically, they lauded the instantaneous feedback that ASR provided. It allowed for quick repetition which increased motivation to keep trying. The system also allowed for unlimited examples and practice, which Hunter and Phoebe both found particularly useful. All participants also reported that they found the pinyin provided underneath the Mandarin translations as very useful for learning. Interestingly, Lola reported not using the translation function for much of the instruction. She believed that her primary focus should be on pronunciation: "I just copy pasted the Mandarin part in the GT, but I didn't go to find the English translation". This focus on pronunciation also meant that Lola felt she did not learn much vocabulary.

The largest hinderance to learnability was likely ASR's limitations as a form of feedback. All participants felt they could have used more nuanced feedback, especially as they improved. This meant that learning would cease if ASR reported persistently that a participant's production was incorrect, as without explicit feedback, participants felt they had nowhere to turn. Eventually, learning would cease if only because frustration had eclipsed motivation, as articulated by Hunter: "If you can hear where you're going wrong, it's great. But if you can't, you have nowhere to go".

Motivation

Participants reported varying levels of motivation to finish instruction. Although the ASR's perceived untrustworthiness impacted Lola's motivation to continue, the other participants reported it as fun: "It felt a little weird, initially. But as I was getting better, it kind of became, almost game-like" (Bruce). Hunter also reported feeling validation whenever the ASR confirmed his successful target-like output. Furthermore, the ASR's inhuman nature was a boon for Bruce, Phoebe, and Gary, who all felt they could practice without bias or rush compared with a living person who might expect them to be faster or more accurate. Lastly, as GT was a familiar application, it was further motivating to use it in a novel way because they imagined continuing to use it in a real-world context: "If I had to travel to China tomorrow… I feel like I could be on the street with GT, and I could figure it out" (Lola).

However, the motivation had a clear end point in many cases as frustration built up. For Lola and Phoebe, the untrustworthiness of the ASR was a constant source of ire. For the other participants, the instant feedback was at first motivating, but as the challenge increased and the feedback became less useful, it led to frustration as the ASR failed to validate their output: "But when I was close but wrong every time, it was extremely frustrating. I couldn't figure out because Google was saying 'oh you're close but you're just wrong'" (Hunter).

Willingness to use GT

Participants reported a strong willingness to use GT for language learning outside of this study, but with specific limitations. As the participants already had experience using GT for translation, they all reported that they would continue to use it for translation but would add ASR and TTS for practice. For example, Phoebe reported that she was willing to use GT more than other technologies because it was familiar, quick, and easy to use, and Lola reported she believed GT was now essential for learning or communicating in unfamiliar languages. Hunter felt that it would be fun to use while watching television in other languages at home. Bruce reported that he would like to use it for practice at home as well. However, all reported that they imagined using GT mostly for words and short phrases and not more complex structures.

Overall, the participants reported that they were willing and able to acquire short Mandarin phrases and produce them, but only reporting the results for those who completed the study does not tell the entire story. Five additional participants were contacted to determine why they may not have been willing to complete instruction. Their reasons included a lack of motivation to either study Mandarin or complete the study, difficulty with the instructional materials (e.g., faulty automatic recognition), and an aversion to being recorded.

Discussion

The goal of this study was to address three research questions: (1) Does the proposed technology-enhanced SRL pedagogy lead to the learning of Mandarin tones? (2) How do the participants self-regulate their learning? (3) How do the participants perceive the proposed GT-enhanced environment for input and output practice? To address the first question, data were collected and analyzed from comprehensibility ratings to judge whether their use of tones improved overtime (holistically), and from participant interviews to analyze the participants' experiences learning tones in an SRL environment. The second and third research questions were addressed using data from the interviews.

Mandarin Learning

Does the proposed technology-enhanced SRL pedagogy lead to the learning of Mandarin tones? The short answer is yes. Their interviews showed that all participants developed some knowledge of what tones are, and they used a variety of methods to form mental representations that led to more successful productions. The comprehensibility ratings also showed that they acquired some ability to produce tones, although most of the participants showed higher comprehensibility with single words and phrases. Similar to Halle et al.'s (2004) findings, our participants had an easier time with short listening discrimination tasks as they had ample practice opportunity to listen to all four tones in short words or phrases. Interestingly, listening discrimination (e.g., of word pairs) constitutes the initial stage of Celce-Murcia et al.'s (2010) framework for pronunciation instruction. These findings also align with the participants' perceptions that GT is best used with words and short phrases: that is where they were most successful.

The Sentence and Introduction assessments urged the participants to create their own, longer sentences. The goal was for more communicative, less scaffolded productions, but still within the boundaries of the instruction (i.e., for guided and communicative practice). Participants showed an average drop in comprehensibility from Translation/Speaking to Sentence/Introduction. This is in line with many of their comments expressing frustration as complexity of the tasks increased; that is, they reported difficulty understanding where the tones lie in creating the longer sentences. It logically follows that, with larger structures, participants would find identifying each tone even more challenging (see Halle et al., 2004 for similar claims). Hearing pitch change would become more difficult as the environments became more complex, and so as perception became more difficult without scaffolding or instructional moderation, production similarly suffered as the participants struggled to bring production in line with perception (see Flege, 1999).

Interestingly, Bruce and Gary reported that some tones were easier than others. This perceived difficulty might align with research on L2 tone acquisition (e.g., Hendry, 2017; Maddieson, 1977), which argues for a developmental sequence of tone acquisition. Although Bruce and Gary were unable to identify which tones were easier at the interview, we believe that, based on the literature, they were likely finding T55 and T51 easier and T35, T215 more difficult. Maddieson's (1977) implicational tonal hierarchy predicted this sequence. In tonal languages, level tones (T55) are acquired first, followed by falling tones (T51), rising tones (T35), and last are dipping tones (T215).

SRL and Online Autonomous Learning

This study provided some evidence as to the effectiveness of the proposed learning environment and accompanying technologies at fostering the development of pronunciation learning in an SRL setting, but further consideration must be made for the large number of participants who have yet to finish. Although data were collected from the five participants who completed the study, there are still questions regarding why 34 participants did not. In a study on app attrition, Tuncay (2020) found that attrition was predicted by the participants' feelings of isolation, overall motivation, poor instructional quality, inauthentic content, and a lack of learner control. Participants in our study also felt that instructional quality (e.g., the ASR's inability to provide easy-to-understand feedback) and motivation to learn was an issue, but they further reported an aversion to being recorded. Considering this study's learning environment, isolation and inauthentic content may have also contributed to attrition. These issues are likely endemic to the self-regulated online learning environment and highlight the importance of scaffolding for increasing motivation and instructional quality.

Concerning participants who did finish, however, based on evidence collected from the interviews from participants who did finish instruction, we believe the participants followed the SRL cyclical model outlined by Zimmerman (1998) and the pronunciation instruction approach proposed by Celce-Murcia et al. (2010). Participants would choose structures to learn, listen to the TTS for input practice, then repeat until the ASR validated their success. Their success would then motivate them to attempt more complex structures. Learning continued this way until the ASR was no longer capable of providing satisfactory feedback. At this point, self-monitoring would cease, breaking the SRL cycle.

Google Translate

Perceptions of GT use was coded following the four themes outlined in Van Lieshout and Cardoso's (2022) study on the pedagogical use of GT in a self-regulated setting: usability, learnability, motivation, and willingness. Regarding usability, participants reported GT was easy to use, partially because they had previous experience with it, but they also found the ASR and TTS functionalities easy to use even though none had indicated using them prior to the study. The largest issue was that the ASR was inherently untrustworthy, in line with Derwing et al.'s (2001) and Van Lieshout and Cardoso's (2022) findings. However, different participants found the ASR unreliable in different ways, varying from too easy on them (e.g., Lola) to too hard (e.g., Phoebe). However, regardless of trustworthiness, the participants found the ASR easy and practical given the learning environment.

In terms of learnability, the participants were able to learn basic Mandarin words and phrases and their associated tones with some success. The most lauded aspect of GT was the instant feedback afforded by the ASR, allowing for quick repetitions and more learning. Furthermore, as it had no judgement or bias, both Bruce and Phoebe reported feeling more comfortable using the ASR for learning when compared with a traditional language classroom, perhaps because of the form of feedback their teachers might provide (e.g., Lyster & Ranta, 1997). Last, we argue that the tones were likely acquired in some kind of sequence as Bruce and Gary reported that some tones were easier than others. This argues that the participants' instruction was similarly successful as in traditional classrooms or naturalistic settings where tones are acquired in a developmental sequence (Hendry, 2017).

Participants also seemed motivated to not only finish instruction, but also to learn. They showed an inclination to keep trying until the ASR validated their productions, sometimes repeating themselves until overwhelmed with frustration. Although not ideal, we argue that when participants were motivated to repeat themselves, that was a positive outcome notwithstanding the frustration observed. Previous research into ASR has pointed out that the technology would only be effective for language learning if it provided appropriate feedback (e.g., Derwing et al., 2000). That the ASR does not appear to automatically correct its transcriptions (as a cellphone might autocorrect a text message) leads participants to keep trying until their production is accurate. This repetition and self-monitoring can be interpreted as signs of SRL and, consequently, motivation to continue learning (Zimmerman, 1998). The immediacy of the feedback provided by the ASR was also particularly motivating. Bruce reported it as game like with the immediate feedback allowing for multiple quick attempts. Although not a game in itself, motivational affordances (actions that are taken to satisfy the actor's needs) such as feedback, challenge, and clear goals are gamification elements that have been shown to have positive effect on learning (Hamari et al., 2014).

Finally, participants showed willingness to use GT for language learning in the future, similar to Van Lieshout and Cardoso's (2022) findings. However, they also showed individual differences in that willingness. First, they were all willing to use GT in the future for their language learning but showed a preference for shorter words and phrases. We believe that this is likely linked to their difficulty with more complex structures, which can also be observed in the comprehensibility ratings. An additional finding concerning willingness to use GT was the ASR's unbiased nature. Gary, Phoebe, and Bruce all reported feeling that the unbiased or neutral feedback was appreciated, with Bruce and Phoebe specifically comparing it to the anxiety-ridden classroom environment. These preferences for ASR over human partners and for shorter, less complex sentences has also been found in other research (e.g., Forsyth et al., 2019). Consequently, the willingness to use ASR is a boon for language learners, but a larger study needs to address participants' willingness to use GT's ASR for larger, more complex structures with further scaffolding (or perhaps, with no scaffolding at all).

Conclusion

This study examined the feasibility of learning Mandarin tones autonomously in an online environment using Google Translate's ASR and TTS features. Specifically, it addressed three research questions: (1) Does the proposed technology-enhanced SRL pedagogy lead to the learning of Mandarin tones? (2) How do the participants self-regulate their learning? (3) How do the participants perceive the proposed GT-enhanced environment for input and output practice? Our results indicate that the participants learned how to perceive and produce tones to varying degrees, probably due to the challenging nature of the target structure. The participants were also able to adapt to the autonomous nature of the proposed environment to develop their own strategies for learning: they took advantage of the instruction provided to adjust their learning strategies and session length to accommodate their personal styles and level of motivation, and found methods of self-monitoring, either using the proposed ASR feature or their own personalized strategies. Overall, GT was perceived by participants as highly usable, easy to learn with, motivating, and something the participants were willing to use for learning other features of Mandarin and other languages. Consequently, we also believe the methodology employed in this study could easily be used for the study of the learning of other foreign languages.

This pilot study provides initial evidence that pronunciation instruction and practice can be effective in an SRL online environment, particularly for learning Mandarin. Based on our results, we can move forward with a larger study that will address some of the limitations presented here, such as the small sample-size, the challenges associated with an autonomous anytime-anywhere instructional method (e.g., motivating more participants to finish instruction), and an investigation of the individual differences observed. Further research should also address the not-always-effective quality of the feedback provided by ASR (e.g., by complementing it with other types of feedback), and the lack of additional materials to accommodate the individualized needs of the students.

Our results indicate that learning Mandarin tones is possible in an SRL, online environment. Consequently, TTS and ASR can be used to augment traditional classrooms to provide online, anytime-anywhere language input and output opportunities. Considering the many opportunities for practice that GT affords and the SRL skills the participants developed, we might find that the students sent home due to the COVID-19 health crisis might return to their classrooms more capable as independent language learners.

Chapter 5

Using Translation Tools for Online Learning in a Self-Regulated Environment

Near the beginning of the 1982 movie *Blade Runner*, a detective asks a woman a series of questions. The questions seem random, but the intention is fairly straightforward: the detective is performing a "Voight-Kampff" test to determine whether the woman, Rachael, is an android simulating a human experience. I think this scene is fascinating in the context of this dissertation. Unlike Rachael, who is an android and able to simulate human interactions, Google Translate (GT) cannot act on its own volition, and its responses must be generated by the interlocutor themselves either by entering text (TTS) or speaking (ASR), and in that way, they are perfectly human responses. However, this is also why GT can be simultaneously an excellent conversation partner and a poor one. Within this dissertation's interactionist approach, GT exists in the liminal space between a computer simulating a person and a real person, between Rachael and the detective. As an interlocutor, GT's goal is to provide the opportunity for language learners to negotiate for meaning, but it can only accomplish that with the help of its *human* conversation partner. It can never pass the Voight-Kampff Test, but I nonetheless argue with this dissertation that with a self-regulating language learner, GT can be an effective language partner for many of the same reasons a human can, providing unlimited opportunity for learners to interact with it and negotiate for meaning.

In this chapter, I will discuss the results of each manuscript and how they inform the broader picture of whether GT is appropriate as a language partner within an interactionist approach. I then discuss how self-regulated learning (SRL) strategies and the learners themselves may be impacted by using GT, and the specific benefits and concerns associated with using GT based on the results of Chapters 2, 3, and 4. I conclude this chapter with a discussion of future directions for research, followed by a personal explanation of my motivations for writing this dissertation. I reflect on what the findings reported and their implications mean to the bodies of literature surrounding CALL, SRL, TTS, and ASR, and to future language learners who may want to try learning a new language with GT.

Summary of Goals and Results

This dissertation's goal was to address the pedagogical appropriateness of GT and its affordances for the learning of certain aspects of Mandarin Chinese (e.g., vocabulary, tones) by answering whether GT can provide the necessary interaction, including input, output, and feedback to promote L2 learning, and whether learners are willing and able to use GT in an online, SRL environment for learning Mandarin. To answer these questions, the following three studies were designed:

1) the first examined whether GT's TTS in Mandarin Chinese is intelligible, comprehensible, and natural sounding;
2) the second assessed GT's ASR for its pedagogical effectiveness and suitability for recognizing native and non-native speech in Mandarin Chinese; and
3) the third analyzed whether the proposed GT-based online SRL environment would lead to language learning and learner satisfaction.

Based on the literature and my experiences with GT, I originally hypothesized that: 1) TTS would be highly intelligible, comprehensible, but unnatural sounding; 2) ASR would more accurately understand (i.e., transcribe) the speech of native speakers than advanced speakers, and less accurately transcribe the speech of intermediate speakers when compared with both native and advanced; and 3) learners would enjoy and benefit from using GT in an online, self-regulated environment, and learning in this environment would lead to the acquisition of certain features of the target language.

The results in Chapter 2 (the first manuscript) found that GT's TTS is highly intelligible to native, advanced, and intermediate speakers, but less comprehensible depending on the proficiency level of the user, and not natural sounding at all. However, it should be noted that due to the design of the study, participants were presented the same sentence twice and there was small improvement in comprehensibility and a large improvement in naturalness the second time participants heard a sentence. Consequently, these results should be validated in a future study where participants are not presented the same sentence more than once to confirm whether the results seen here are accurate, specifically for naturalness. The results in Chapter 3 (the second manuscript) indicated that GT's ASR struggles with L2 speech, even though native Mandarin speaking raters had no problems transcribing it. Specifically, GT was accurate about 95% of the time with native Mandarin Chinese speakers, but only 80% of the time with L2 speakers regardless of their proficiency level (intermediate or advanced). The final study described in Chapter 4 (the third manuscript), examined whether participants can/would learn aspects of Mandarin Chinese in the proposed SRL setting, how they self-regulate their learning, and how they perceive the complex online GT learning environment itself. The study's results indicated that the participants were willing and motivated to use GT to learn an L2 despite its flaws (e.g., lack of consistency in accuracy when transcribing the participants' speech), and that the environment led to effective language learning.

Combined, these studies provide evidence of GT's ability to be used as an interlocutor, as it can provide input, output, and feedback sufficient for L2 learning, and that learners are capable and willing to learn with the technology. The pedagogical experience is not perfect, but it can be effective within the interactionist approach adopted in this dissertation.

Interactionist Nature of GT

This dissertation was conceived as an exploration of what negotiation for meaning would really look like in a modern online SRL environment. As mentioned in Chapter 1, the classroom has strict limitations of time and space (Collins & Muñoz, 2016), especially when it comes to pronunciation practice (Foote et al., 2016). Further, during the COVID-19 pandemic (when the studies reported in this dissertation took place), many families had to work and learn from home, where the classroom had become an abstract concept. Consequently, the ability to have access to a virtual interlocutor became a much more appealing prospect, particularly because learning is already being mediated through the computer (Chapelle & Jamieson, 2008). Specifically, computers in general have the capacity to fill the role of interlocutor, providing students greater control and more opportunities to negotiate for meaning (Chapelle, 2005).

GT seemed like an obvious choice to test these hypotheses. There are numerous TTS and ASR programs such as Natural Reader (Liakin et al., 2017) or Dragon NaturallySpeaking (Derwing et al., 2000) which can be used within this autonomous learning environment, and they seem to be pedagogically appropriate, as confirmed in previous research (discussed elsewhere in this dissertation). However, GT provides translation with both TTS and ASR software *combined*; in addition, it is so popular that most people with internet access in the world have become familiar with it. Further, recent research has shown that the combination of TTS + ASR + Translation can be used for language learning, albeit in a limited fashion (e.g., Van Lieshout & Cardoso, 2022). Specifically, Van Lieshout and Cardoso found that participants were able to use GT's Translation capability to learn 10 Dutch phrases in a laboratory environment by practicing their listening (via

TTS) and speaking skills (via ASR) with the software. I found this study inspiring because, although small in scope, it was the beginning of truly interacting with a computer in a manner that I had only really seen within a classroom or immersion context—environments where learners can easily interact with other speakers of the target language. Consequently, GT seemed like the most obvious choice for this dissertation.

The Interaction Hypothesis states that interaction, and specifically modified interaction and negotiation for meaning, can facilitate language learning (Long, 1996). Pica (1994) outlines three learner-oriented conditions that are affected during negotiation for meaning: comprehensible input (1982), comprehensible output (1995), and attending to language structures that may otherwise go unnoticed (Schmidt, 1990). Negotiation strategies such as comprehension checks, repetition, and segmenting portions of the target speech (Gass, 1997) all target at least one of the above learner-oriented conditions (Pica, 1994). This dissertation is motivated partially by the fact that free software such as GT can fulfil these tasks and the role of interlocutor by providing opportunities for comprehensible input and output, and in translating language in a manner possibly similar to a classmate or even a teacher in a foreign language learning context.

Despite its simplicity, GT can be considered a powerful pedagogical tool for L2 learning. Using any available device with a web browser, a learner can choose a word or phrase in their L1 to translate. Next, they can listen to it in their target language with GT's TTS, modifying input as necessary to focus on specific structures by having GT repeat them, or by segmenting portions of the input. They can then practice speaking with GT's ASR. When GT signals a miscommunication (e.g., by providing a transcription that does not match the intended output), the learner can enact comprehensible output strategies such as repetition, slowing down speech, or segmenting the target speech until GT's ASR transcription is accurate. In this way, GT provides an analogous experience to a real-life interlocutor, theoretically motivating the learner to negotiate for meaning.

As GT is online and works on almost every platform, it can operate in true anytime-anywhere environments. I planned my dissertation to address whether interaction would be possible in a such a situation, and to test whether the three capabilities of GT (i.e., Translation, TTS, ASR) would be sufficient for language learning in this environment. To properly assess the pedagogical appropriateness and value of GT, I used Cardoso's (2022) chronological framework, which describes how technologies have been assessed in the CALL literature for pedagogical value: first, the conceptualization and development of the technology (not applicable for this study), next is the assessment of its suitability (including usability and learner attitudes towards it), and then last is testing its overall effectiveness with a pre-post test research design. Following this framework, my foremost goal needed to verify if GT (an existing technology) could truly be used for language learning by determining whether its basic TTS and ASR software met the assumptions required for learning. For instance, is TTS intelligible and comprehensible to serve as appropriate L2 input (see Chapter 2)? Is ASR able to understand L2 speech at approximately the same level a native or fluent speaker understands L2 speech (Chapter 3)? Following this, I wanted to test the next rung in Cardoso's (2022) framework and assess GT's suitability and usability by analyzing whether participants would or could use GT in truly anytime-anywhere settings, when a real-life interlocutor is unavailable (Chapter 4).

Self-Regulation and the Online Environment

This study from the ground up was focused on a true anytime-anywhere online learning environment. Participants used their own computers and phones to record audio from all over the world, including Brazil, Canada, China, Italy, Japan, and several other locales. This was an important aspect of the study. Although more controlled data collection like those seen in most of the studies cited in this dissertation is the norm, I felt it would be disingenuous to argue for GT's usefulness as an interlocutor in locations besides the classroom, and then only test it in the quietest environments with the best computer hardware. Some of the issues that came up in this environment will be discussed later in this chapter.

An interesting aspect of this dissertation design is that I argue that GT can act as an interlocutor, meaning that a learner can interact with GT while they are alone, at their own time and pace, wherever and whenever they want to learn. However, this scenario constitutes a contradiction: how can a language learner have an interlocutor and at the same time be alone? Consequently, I needed to mesh interaction theory with self-regulation theory. That is, I needed to think about how people would learn a language alone with a machine that was acting as a person. This has been explored in several other studies with intelligent personal assistants such as Google Assistant or Amazon's Alexa (Dizon, 2020; Moussalli & Cardoso, 2016, 2019). For example, Dizon found that learners enjoyed using Amazon's Alexa and found it useful with only a little practice, while Moussalli and Cardoso showed that Alexa provided stress-free exposure (2016) and that learners were able to develop their own learning strategies to accommodate any communication breakdowns with the device (2019). As the intelligent assistants showed promise when participants interacted with them, I knew I wanted to explore how people would self-regulate their learning in a GT-based online anytime-anywhere environment (Chapter 4) in addition to whether its TTS and ASR functionalities were effective (Chapters 2 and 3 respectively).

Self-regulated learning is the application of strategies to be an effective autonomous learner (Andrade & Bunker, 2008). SRL research often draws from socio-cultural approaches, and specifically from research on scaffolding in Vygotsky's Zone of Proximal Development, a metaphorical location where a person is able to perform tasks with the help of a peer or tutor at a level beyond what they can accomplish on their own (Vygotsky, 1978). On a more applied level, the body of research around SRL argues that to be an effective learner in an autonomous environment, some level of assistance is necessary. This assistance can come in a variety of forms, ranging from textbooks to open online encyclopedias to private tutors. What this means for my study was that, within an SRL framework, interaction with GT may create something akin to a ZPD in which language learning, beyond what one is capable alone, would be possible. However, for that to happen, learners would first need help in developing strategies for how to use GT for language learning. SRL requires that the learner set goals, plan their learning strategically, and self-monitor (Zimmerman, 1998).

To scaffold and organize learning, several SRL and CALL studies recommended the use of learning management systems such as Moodle, Blackboard, or D2L (Dabbagh & Kitsantas, 2005; Godwin-Jones, 2015). Moodle seemed appropriate because it is highly customizable with multiple audio and text recording apps available, it is free (except for the purchase of server space – required for its installation and management), and it allows participants to log in individually and securely. Following insights from interactionist, SRL, and CALL theories, I was able to create a real anytime-anywhere scaffolded learning environment which, at the same time, also allowed for data collection (i.e., surveys, pretests, and posttests) and the compilation of information about the participants' use of the system (e.g., via their daily logs). Theoretically, learners could now use my Moodle website to develop strategies for learning and interacting with GT even though they are in effect learning by themselves.

Listening to TTS and Speaking with ASR: Interacting with GT

The combined results from this dissertation indicate that GT would make a valuable interlocutor within the proposed online autonomous self-regulated environment. Learners had access to intelligible input that they were able to alter to make more comprehensible (see Chapters 2 and 4), and they were able to practice the newly acquired forms using ASR.

GT's TTS software was very intelligible and mostly comprehensible, but sounded unnatural in the view of raters and language users. However, the tool was effective within this learning context and learners were able to develop self-regulated learning strategies when using it. The participants in Chapter 4 were able to take advantage of the TTS' functionality and appreciated that it could be used in an unlimited number of times, without bias and judgement. They developed individual self-regulated learning strategies to take advantage of its strengths (e.g., the ability to repeat words, phrases, or longer stretches of text), and even began to discriminate between the Mandarin tones, despite their low salience. Overall, the results from Chapters 2 and 4 indicate that the TTS functions work reasonably well for producing the appropriate input for language practice.

The ASR functionality in this context is more questionable, but I argue that this technology is still pedagogically useful. The ASR had recognition scores of approximately 80% for L2 speakers, regardless of level, and when compared with the results in Chapter 4, we can make further assumptions about what may have happened. For example, although all participants in Chapter 4 mentioned issues with Google's inconsistencies in transcribing their speech, it was the participant who self-identified as having a strong accent who became the most frustrated. This aligns with early research from Derwing et al. (2000), who found ASR software struggled with accent, and accent and speech variability in general seem to be known and consistent problems within the field of ASR development (Benzeghiba et al., 2007; Vergyri, Lamel, & Gauvain, 2010). In essence, although there are studies that indicate ASR may be improving, particularly for English

(Bione & Cardoso, 2020; McCrocklin, 2019), the preponderance of research on ASR seems to indicate that it will struggle with L2 accents (see also findings reported in Chapter 4, wherein participants found it frustrating when Google did not understand them).

Simply put, the fact that the ASR software seems to struggle 20% of the time with L2 Mandarin speech but only 5% or less with native speaker speech indicates that there is still room for the software to improve. However, as touched upon in Chapter 3, the quality of the ASR's transcriptions may not impede interaction. To a degree, the findings reported in Chapter 4 seem to indicate that, as all the participants were able to finish the tasks including making their own sentences and talking about themselves with GT, they were able to use the ASR for speaking practice. The participant who self-reported as having a strong accent when talking about her struggle with GT said that she just wanted Google to understand her. However, later she also said that GT was like having a person in front of her. She also felt more comfortable with GT than a real person, and that its unbiased and immediate feedback was intrinsically motivating. That is, despite its challenges, the participants were able use the ASR for self-regulated learning even if it could be frustrating.

Each participant's experience with GT was unique with some reporting higher trust and lower frustration than others, but all the participants in Chapter 4 seemed to agree that, regardless of GT's deficiencies, it was an effective tool for language learning, particularly when used in a self-regulated manner; as I argue, it is also capable of acting as an interlocutor for L2 pedagogical purposes. Still, because GT could never pass the Voight-Kampff and trick someone into believing it was human, there exists the possibility that participants will become frustrated with it and quit because it will always rely on the human partner to do most of the work, and it will never learn to respond more effectively or in a less frustrating way.

Motivation and Frustration when Interacting with GT

As mentioned in Chapters 3 and 4, GT can sometimes be a frustrating experience when it does not recognize your speech. The motivation of the learner is one of the key factors for academic success in self-regulated learning (Zimmerman, 1998), and if the learner becomes too frustrated, learning may fail as they become demotivated (Winne, 2001). In essence, one of the goals of this dissertation was to test whether stimulating learners through interaction opportunities and CALL designs would motivate them to continue to learn, since without sufficient motivation, self-regulation would never work in this SRL-oriented setting (Zimmerman, 1998). However, as touched on in Chapters 3 and 4, motivation in this space is more complex than simply a willingness to learn, as it can be affected by factors such as the speed and accuracy of the feedback, how frustrating the experience is, and possibly, whether people find using GT "fun."

One interesting facet of GT that the participants in Chapter 4 noted was its ability to provide instantaneous feedback via a game-like experience, something that motivated them to repeatedly attempt challenging their oral productions with GT's ASR beyond the minimum expectations of the study. This may partially explain why even the most frustrated of the five participants completed their learning task. This aligns with Baker et al.'s (2010) findings that boredom is more of an issue than frustration, meaning that as long as the participants are not bored, frustration should only minimally deter learning. These results may also touch upon gamification, the use of game elements to better motivate learners to adopt a specific behavior (Deterding et al., 2011). Even though GT has no specific game-like design, Hamari (2015) argues that any activity can provide intrinsic motivation by incorporating elements such as uncertain outcomes (such as what an imperfect ASR may provide) and immediate constructive feedback. In this way, based on my findings, it is possible that GT's TTS and ASR functionalities may have kept the learners motivated, despite any frustration, because it really is "gamelike."

Regardless of GT's gamelike nature, the results of Chapter 4 show that, although frustration was likely present, it did not ultimately impact learning outcomes. This also aligns with the wider body of research into the pedagogical use of ASR (Liakin et al., 2017) and intelligent assistants for learning, where participants often enjoyed the learning environments that these technologies created (Dizon, 2020; Moussalli & Cardoso, 2019). Consequently, regardless of the frustration that may come from GT not being as capable a listener as a native Mandarin speaker (as seen in Chapter 3), the participants in Chapter 4 did not quit. This may indicate that frustration does not actually play a large role in participant motivation within this space, perhaps due to the gamelike nature of instantaneous corrective feedback, or perhaps there is more at work.

Future Directions for Research

Considering some of the limitations of this dissertation, as indicated earlier and in previous chapters, future research could focus on the last stage of Cardoso's (2022) chronological framework for describing research in CALL: to analyze GT's pedagogical effectiveness using pre-post test designs, and examine whether the proposed learning environment would be effective and generalizable to other similar contexts. A larger pre-posttest study itself can take many avenues to build on this dissertation. For example, switching to another target language may further indicate whether the issues GT has with L2 learner accent, seen in Derwing et al. (2000) and in this dissertation, may generalize to other linguistic contexts. Other tonal languages would also be interesting to study; if the results are consistent in other tonal languages, it may indicate that ASR in general struggles with tones. Research into other tonal languages may also indicate that newer studies such as McCrocklin et al. (2019), which found Google Voice Typing's ASR effective in recognizing L2 speech, are representative of GT's ability with English only, not tonal or non-Western languages.

Another area recommended for future research is examining GT with a larger number of participants in a classroom setting. If one of GT's primary benefits is its ability to be used in anytime-anywhere settings, there must be interesting benefits in a classroom setting. For example, one of the motivations of this dissertation was that there was insufficient time in the classroom for interaction (Collins & Munoz, 2006; Foote et al., 2016). A large study on using GT as an interlocutor within the classroom for creating additional opportunities for interaction would be interesting. Students could then continue to use GT at home, which would also motivate a delayed post-test to verify whether they retain gains from interacting with GT during class and after they were given the opportunity to practice with GT on their own. It would also be interesting to explore what types of self-regulated learning strategies learners may develop in this scenario. For example, they may find interesting ways of interacting with both GT and another human interlocutor.

Another interesting avenue for research that was touched on in Chapter 4 is GT's translation functionality. It is this functionality that truly allows learners to self-regulate their learning, giving them the ability to choose what they want to learn. However, this functionality has not been explored in depth. Questions remain regarding how accurate it is in Mandarin and other languages, and whether it can be used for more advanced learning beyond the scope of this dissertation.

Conclusion

In this dissertation, I set out to answer whether GT would be an effective partner using an interaction-based language learning approach. To wit, I asked three overarching questions:

1) Is GT's TTS pedagogically suitable for input practice in Mandarin?

2) Is GT's ASR pedagogically suitable for output practice and feedback in Mandarin?
3) Do learners find the GT-enhanced environment pedagogically appropriate for learning Mandarin? Can learning take place in such an environment?

In Chapter 2, I found that there was no significant difference in intelligibility between TTS and a native speaker, but the synthesized speech was less comprehensible than a native Mandarin speaker and unnatural sounding (but recall that naturalness was affected by sentence presentation, and these results require further validation). Nonetheless, these results suggested that GT's TTS could provide pedagogically appropriate input for L2 input practice. In Chapter 3, I found that the ASR's recognition scores were significantly less for L2 learners of Mandarin (approximately 80%) when compared with recognition scores from native speakers (approximately 95%). Although not ideal, I nonetheless argue that the ASR is sufficiently accurate for L2 learners of Mandarin to use it for output practice and receive adequate feedback. Therefore, I believe these results indicate that GT's ASR can provide pedagogically appropriate output practice and feedback opportunities for Mandarin learners. Last, in Chapter 4, participants used a Moodle-powered website to scaffold GT-enhanced Mandarin language learning. They all reported being willing and motivated to use GT for language learning, and that it was easy to use and effective for the purpose. In this study, participants developed their own strategies for learning, reported being able to perceive and produce tones and their accompanying vocabulary, and said they would use it again in the future. These results therefore indicate that learners find the proposed environment pedagogically appropriate for learning Mandarin, and when combined, the results of Chapters 2, 3, and 4 strongly suggest that GT would make an effective interlocutor and language learning partner within an interaction-based approach.

This dissertation is the culmination of many years of work, but it was motivated by more than 30 years of language learning and a decade of teaching experience. Consequently, I chose this direction for my research for a variety of reasons. From 2010-2015, I lived in Beijing, China, and one of my initial problems when learning Mandarin was that nobody could understand me or had the patience to listen to me slowly pronounce the simplest words. Overtime, I was lucky enough to develop a small network of acquaintances who were patient enough to interact with me in Mandarin. However, many of my friends were never able to find the opportunities to practice that I was able to find, and although we took the same Mandarin Language classes, they were unable to reach the level of fluency I did. I believe the lack of opportunities for meaningful interaction was one of the reasons that led to such poor results.

I realize I am very lucky to have had such an immersive experience while learning a language. As a language teacher and researcher, I have always wondered how L2 learners who have limited access to language practice outside the classroom could have those interaction experiences in any meaningful quantity, and it was a revelation when I realized GT could be an effective language partner, regardless of whether it could pass the Voight-Kampff test. In fact, GT could be an amazing language partner, capable of speaking and understanding numerous languages whenever, wherever I want. Although it is flawed, I realize I can have a conversation with this software in my second language, and it can listen to me and try its best to understand (i.e., transcribe) my speech, which is as much as I would ask of any interlocutor.

For many people, GT may not be the ideal language partner. It can be frustrating to use, it sounds deeply unnatural, and it has to be “puppeted” to be an effective interlocutor, which for some learners might be demotivating. However, it is free, it is available in a number of languages, and for many people, it can provide the opportunity to practice speaking, listening, and negotiating for meaning in their target language outside the classroom. Considering all of these affordances, language learners everywhere who are struggling to find human language partners could try using GT for language practice. They may be surprised to find out just how good of a language partner it may be, considering the findings reported in this dissertation.

References

Abdullah, F. & Ward, R. (2016). Developing a general extended technology acceptance model or e-learning (GETAMEL) by analysing commonly used external factors. *Computers in Human Behavior*, *56*, 238–256.

Alphabet Inc. (2020). *Google Translate*. Retrieved April 2nd, 2020.

Al-Adwan, A. S. (2020). Investigating the drivers and barriers to MOOCs adoption: The perspective of TAM. *Education and Information Technologies*, *25*, 5771–5795.

Andrade, M. & Bunker E. (2009). A model for self-regulated distance language learning. *Distance Education*, *30*(1), 47–61.

Baker, R., D'Mello, S., Rodrigo, M. & Graesser, A. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241.

Barrett, A., Pack, A., Guo, Y., & Wang, N. (2020). Technology acceptance model and multi-user virtual reality learning environments for Chinese language education. *Interactive Learning Environments*, 1–18.

Bibauw, S., François, T., & Desmet, P. (2015). Dialogue-based CALL: An overview of existing research. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds.), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 57–64). Dublin: Research-publishing.net. https://doi.org/10.14705/rpnet.2015.000310

Bione, T., & Cardoso, W. (2020). Synthetic voices in the foreign language context. *Language Learning & Technology, 24*(1), 169–186.

Cardoso, W. (2018). Learning L2 pronunciation with a text-to-speech synthesizer. In P. Taalas, L. Bradley, and S. Thouësny (Eds.), *Language Learning as Exploration and Encounters*; *Selected papers from EuroCALL* (pp. 16–21). Research-publishing.net.

Cardoso, W. (2022). Technology for Speaking Development. In T. Derwing, M. Munro, & R. Thomson (Eds), *Routledge Handbook on Second Language Acquisition and Speaking* (p. 299-313). Routledge, Taylor & Francis Group.

Cardoso, W., Collins, L., & White, J. (2012). *Phonological input enhancement via text-to-speech synthesizers: The L2 acquisition of English simple past allomorphy* [conference presentation]. AAAL 2012, Boston, MA, United States.

Cardoso, W., Smith, G., & Garcia Fuentes, C. (2015). Evaluating text-to-speech synthesizers. In F. Helm, L. Bradley, M. Guarda, & S. Thouseny (Eds.), *Proceedings of Eurocall* (pp. 108–113). Research-publishing.net.

Celce-Murcia, M., Brinton, D., & Goodwin, J. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge University Press.

Chao, Y. (1968). *A grammar of spoken Chinese.* Los Angeles: University of California Press.

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research.* Cambridge University Press.

Chapelle, C. (2003). *English language learning and technology*. John Benjamins.

Chapelle, C. (2005). Interactionist SLA theory in CALL research. In J. Egbert and G. Petrie (Eds.), *Research perspectives on CALL* (pp. 53–64). Laurence Erlbaum Associates.

Chapelle, C. A. & Jamieson, J. (2008). What is Call? In C. A. Chapelle & J. Jamieson (Eds.), *Tips for teaching with CALL: Practical approaches to computer-assisted language learning* (pp. 1–9). Pearson Education.

Chen, N. F., Shivakumar, V., Harikumar, M., Ma, B. & Li, H. (2013). Large-scale characterization of mandarin pronunciation errors made by native speakers of European languages. *Proceedings of Interspeech 2013*, *France*, 2370–2374.

Collins, L. & Muñoz, C. (2016). The foreign language classroom: Current perspectives and future considerations. *The Modern Language Journal, 100*, 133–147.

Dabbagh, N. & Kitsantas, A. (2005). Using web-based pedagogical tools as scaffolds for self-regulated learning. *Instructional Science*, *33*, 513–540.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340.

De Paepe, L., Zhu, C., & Depryck, K. (2018). Online language teaching: Teacher perceptions of effective communication tools, required skills and challenges of online teaching. *Journal of Interactive Learning Research*, *29*(1), 129–142.

Delogu, C., Conte, S., & Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication*, *24*(2), 153–168.

Dembo, M., Junge, L., & Lynch, R. (2006). Becoming a self-regulated learner: Implications for web-based education. In H. F. O'Neil & R. S. Perez (Eds.), *Web-based learning: Theory, research, and practice* (pp. 185–202). Lawrence Erlbaum.

Derwing, T. & Munro, M. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(4), 476–490.

Derwing, T., Munro, M., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly*, *34*(3), 592–603.

Deterding, S., Dixon, D., Khaled, R., & Nacke L. (2011). From game design elements to gamefulness: Defining "Gamification". *MindTrek '11: Proceedings of the 15th*

International Academic MindTrek Conference: Envisioning Future Media Environments, 9–15.

Dizon, G. (2020). Evaluating intelligent personal assistants for L2 listening and speaking development. *Language Learning & Technology*, *24*(1), 16–26.

Egbert, J. & Shahrokni, S. A. (2018). CALL principles and practices. *OER*. Retrieved from https://opentext. wsu. edu/call/chapter/principles-of-call.

Ellis, N. C. (1999). Cognitive approaches to SLA. *Annual Review of Applied Linguistics*, *19*, 22–42.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17–44.

Erlam, R. & Tolosam, C. (2022). *Pedagogical realities of implementing task-based language teaching*. Benjamins.

Flege J. (1999) The relation between L2 production and perception. In: Ohala J, Hasegawa Y, Ohala M, Granville D, and Bailey A (Eds.), *Proceedings of the XIV International Congress of the Phonetic Sciences: Volume 2* (pp. 1273–76). University of California.

Foote, J., Trofimovich, P., Collins, L., & Soler Urzúa, F. (2013). Pronunciation teaching practices in communicative second language classes. *The Language Learning Journal*, *44*(2), 1–16.

Forsyth, C., Luce, C., Zapata-Rivera, D., Jackson, G., Evanini, K., & So, Y. (2019). Evaluating English language learners' conversations: Man vs. machine. *Computer-assisted Language Learning*, *32*(4), 398–417.

Gass, S. (1997). *Input, interaction and the second language learner*. Lawrence Erlbaum Associates.

Gibbons, P. (2002). *Scaffolding language, scaffolding learning*. Heinemann.

Godwin-Jones, R. (2015). The evolving roles of language teachers: Trained coders, local researchers, global citizens. *Language Learning & Technology*, *19*(1), 10–22.

Goldschneider, J. M. & DeKeyser, R. M. (2001). Explaining the "Natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning*, *51*(1), 1–50.

Hadwin, A. F. & Winne, P. H. (2001). CoNoteS2: A software tool for promoting self-regulation. *Educational Research and Evaluation*, *7*(2), 313–334.

Hakuta, K. & D'Andrea, D. (1992). Some properties of bilingual maintenance and loss in Mexican background high school students. *Applied Linguistics*, *13*(1), 72–99.

Halle, P., Chang, Y., & Best, C. T. (2004). Identification and discrimination of mandarin Chinese tones by mandarin Chinese vs. French listeners. *Journal of Phonetics*, *32*(3), 395–421.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does Gamification work? A literature review of empirical studies on Gamification. In R. H. Sprague (Ed.), *Proceedings of the 47th Annual Hawaii International Conference on System Sciences* (pp. 3025–3034). IEEE.

Hatch, E. M. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 401–435). Newbury House.

Hendry, C. (2017). The effects of type of instruction on the initial stages of L2 perception and production of tones in Mandarin Chinese (Unpublished MA thesis). Concordia University, Montreal.

Hsu, L. (2016). An empirical examination of EFL learners' perceptual learning styles and acceptance of ASR-based computer-assisted pronunciation training. *Computer-assisted Language Learning*, *29*(5), 881–900.

Isaacs, T. & Thomson, R. I. (2013). Rater Experience, Rating Scale Length, and Judgments of L2 Pronunciation: Revisiting Research Conventions. *Language Assessment Quarterly*, *10*(2), 135–159.

Kang, O., Moran, M., Ahn, H. & Park, S. (2020). Proficiency as a mediating variable of intelligibility for different varieties of accents. *Studies in Second Language Acquisition*, *42*, 471–487.

Kennedy, S. & Trofimovich, P. (2008). Intelligibility, Comprehensibility, and Accentedness of L2 Speech: The Role of Listener Experience and Semantic Context. *The Canadian Modern Language Review*, *64*(3), 459–489.

Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Pergamon.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 579–589.

Levis, J. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal*, *32*(1), 1–25.

Liakin, D., Cardoso, W. & Liakina, N. (2017a). The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison. *Computer-assisted Language Learning*, *30*(3–4), 348–365.

Liakin, D., Cardoso, W. & Liakina, N. (2017b). Mobilizing instruction in a second-language context: Learners' perceptions of two speech technologies. *Languages*, *2*(11), 1–21.

Lightbown, P. M. & Spada, N. (2013). *How languages are learned*. Oxford University Press.

Li, C., & Thompson, S. (1976). The acquisition of tone in Mandarin-speaking children. *Journal of Child Language*, *4*, 185–99.

Li, R., Meng, Z., Tian, M., Zhang, Z., Ni, C., & Xiao, W. (2019). Examining EFL learners' individual antecedents on the adoption of automated writing evaluation in China. *Computer-assisted Language Learning*, *32*(7), 784–804.

Liu, Y. (2000). Focus on form as a pedagogical framework for fostering a native-like mandarin tonal identification system. *Language and Linguistics*, *12*(3), 627–667.

Long, M. H. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics*, *4*(2), 126–141.

Long, M. H. (1996). The role of linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Luc, G. & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, *25*(5), 605–621.

Lyster, R. & Ranta, L. (1996). Corrective feedback and learner uptake. *Studies in Second Language Acquisition*, *19*(1), 37–66.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*(1), 1–40.

Ma, L. & Lee, C-S. (2017). Investigating the adoption of MOOCs: A technology-user-environment perspective. *Journal of Computer-assisted Learning*, *35*, 89–98.

Mackey, A., & Goo, J. (2007). Interaction research: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). Oxford University Press.

Maddieson, I. (1978). Universals of tone. In J. H. Greenberg (Ed.), *Universals of human language (Vol. 2): Phonology* (pp. 335–366). Stanford: Stanford University Press.

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System*, *57*, 25–42.

McCrocklin, S., Humaidan, A., & Edalatishams, E. (2019). ASR dictation program accuracy: Have current programs improved? In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference* (pp. 191–200). Iowa State University.

McDonough, K. & Mackey, A. (2008). Syntactic priming and ESL question development. *SSLA*, *30*, 31–47.

Mlott, S.R., Marcotte, D.B. & Lira, F.T. (1976). The efficacy of programmed instruction in the training of paraprofessionals. *Journal of Clinical Psychology*, *32*, 419–424.

Moussalli, S., & Cardoso, W. (2016). Are commercial ‘personal robots’ ready for language learning? Focus on second language speech. In S. Papadima-Sophocleous, L. Bradley, & S. Thouësny (Eds.), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 325–329).

Moussalli, S., & Cardoso, W. (2019). Intelligent personal assistants: can they understand and be understood by accented L2 learners? *Computer-assisted Language Learning*, *33*(8), 865–890.

Mroz, A. (2020). Aiming for advanced intelligibility and proficiency using mobile ASR. *Journal of Second Language Pronunciation*, *6*(1), 12–38.

Munro, M., & Derwing, T. (1995). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *45*(1), 73–97.

Munro, M., & Derwing, T. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *SYSTEM*, *34*, 520–531.

Nassaji, H. & Kartchava, E. (2017). Corrective feedback and good language teachers. second language teaching and learning. In C. Griffiths and Z. Tajeddin (Eds.), *Lessons from Good Language Teaching* (pp. 151–164). Cambridge University Press.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer-assisted pronunciation training. *Computer-assisted Language Learning*, *15*(5), 441–467.

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer-assisted pronunciation training for foreign language learning by children. *Computer-assisted Language Learning*, *21*(5), 393–408.

Park, S-Y. (2009). An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. *Journal of Educational Technology & Society*, *12*(3), 150–162.

O'Brien, M. G. (2014). L2 Learners' Assessments of Accentedness, Fluency, and Comprehensibility of Native and Nonnative German Speech. *Language Learning*, *64*(4), 715–748.

Patel, A. D., Xu, Y. & Wang, B. (2010). The role of F0 variation in the intelligibility of Mandarin sentences. *Speech Prosody*. Chicago, IL: ISCA.

Penning de Vries, B., Cucchiarni, C., Bodnar, S., Strik, H., & van Hout, R. (2015). Spoken grammar practice and feedback in an ASR-based CALL system. *Computer-assisted Language Learning*, *28*(6), 550–576.

Pelzl, E. (2021). Foreign accent in second language Mandarin Chinese. In C. Yang (Ed.), *The acquisition of Chinese as a second language pronunciation: Segments and prosody* (pp. 257–280). Springer.

Pica, T. (1994). Research on negotiation: What does it reveal about second language learning conditions, processes, and outcomes? *Language Learning*, *44*, 493–527.

Pienemann, M. (2015). An outline of Processability Theory and its relationship to other approaches to SLA. *Language Learning*, *65*(1), 123–151. https://doi.org/10.1111/lang.12095

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and psychological measurement*, *53*, 801–813.

Ruivivar, J. & Collins, L. (2019). Nonnative accent and the perceived grammaticality of spoken grammar forms. *Journal of Second Language Pronunciation*, *5*(2), 269–293.

Saito, K., & Wu, X. (2014). Communicative focus on form and second language suprasegmental learning: Teaching Cantonese learners to perceive Mandarin tones. *Studies in Second Language Acquisition*, *36*, 647–680.

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2016). Re-examining Phonological and Lexical Correlates of Second Language Comprehensibility: The Role of Rater Experience. In T. Isaacs, & P. Trofimovich (Eds.), *Second Language Pronunciation Assessment: Interdisciplinary Perspectives* (pp. 141-156). Bristol: Multilingual Matters.

Saldaña, J. (2009). *The coding manual for qualitative researchers*. Sage.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–58.

Song, C. (2021). What is in the final stage of inter-language? Tone errors and phonological constraints in spontaneous speech in very advanced learners of Mandarin. In C. Yang (Ed.), *The acquisition of Chinese as a second language pronunciation: Segments and prosody* (pp. 21–54). Springer.

Strachan, L. & Trofimovich, P. (2019). Now you hear it, now you don't: Perception of English regular past –ed in naturalistic input. *Canadian Modern Language Journal*, *74*(1), 84–104.

Surendran, D. & Levow, G-A. (2004). *The Functional Load of Tone in Mandarin is as High as that of Vowels.* University of Chicago.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass and C. G. Madden (Eds.), *Input in second language acquisition*. Newbury House, 235–53.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook and B. Seidlhofer (Eds.), P*rinciple and practice in applied linguistics: Studies in honour of henry G. Widdowson* (pp. 125–144). Oxford University Press.

Trofimovich, P. & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905–916.

Tuncay, H. (2020). *App Attrition in Computer-Assisted Language Learning: Focus on Duolingo* (Master's thesis, McGill University, Montreal, Canada). Retrieved from https://escholarship.mcgill.ca/concern/theses/nv935746n?locale=en

Van Doremalen, J., Boves, L., Colpaert, J., Cucchiarini, C., & Strik, H. (2016). Evaluating automatic speech recognition-based language learning systems: A case study. *Computer-assisted Language Learning, 29*(4), 833–851.

Van Lieshout, C. & Cardoso, W. (2022). Google Translate as a tool for self-directed language learning. *Language Learning & Technology*.

Venkatesh, V. & Davis, F. D. (1996). A model of the antecedents of perceived ease of use: Developments and Test. *Decision Sciences*, *27*(3), 451–481.

Venkatesh, V. & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, *46*(2), 169–332.

Vygotsky, L. (1978). *Mind in society.* Harvard University Press.

Wan, I., & Jaeger, J. (1998). Speech errors and the representation of tone in mandarin Chinese. *Phonology*, *15*(3), 417–461.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 153–189). Mahwah, NJ: Erlbaum.

Winne, P. (2018). Theorizing and researching levels of processing in self-regulated learning. *British Journal of Educational Technology*, *88*, 9–20.

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Zimmerman, B. (1998). Academic studying and the development of personal skill: A self-regulatory perspective. *Educational Psychologist*, *33*(2), 73–86

Appendix A

TTS Intelligibility and Comprehensibility/Naturalness* Sentences

*Each of these sentences is evaluated for both Comprehensibility and Naturalness

(Adapted from Cardoso et al., 2015; Bione & Cardoso, 2020)

Intelligibility Sentences

1.	A four-year-old boy and his mother sat in the hospital room.	一个四岁的男孩和他的母亲坐在医生的房间里。
2.	He saw a pregnant woman on the other side of the room.	他在房间的另一侧看到一个女人。
3.	Is the baby in your stomach?	宝宝在肚子里吗？
4.	If he is such a good baby, then why did you eat him?	如果他是一个好孩子，那你为什么要吃他？
5.	Last Christmas, Zhao Jing received the best present: a video game.	去年圣诞节，赵婧收到了最好的礼物：一只电脑游戏。
6.	Teacher Wang heard Zhao Jing say some very bad words.	王老师听到赵婧说了一些很不好的话。
7.	Teacher Wang was so angry, he decided to criticize the student.	王老师感到非常生气，他决定批评那只学生。
8.	He carried the chicken into the kitchen and put it into the fridge.	他将鸡肉带进厨房，并将其放入冰箱。
9.	He did not know why the student stopped saying bad words after only a few minutes outside.	他不知道为什么学生在外面呆了几分钟后才停止说坏话。
10.	May I ask what the student did wrong?	请问这只学生做错了什么？

Comprehensibility and Naturalness

1.	He placed his glasses on his nose and looked up.	他把眼镜放在鼻子上，抬起头来。
2.	When he arrived, he saw that the door was open.	当他到达时，看到前门是开着的。

3.	She quickly opened the box and saw pictures and a letter.	她迅速打开盒子，找到了照片和信。
4.	I looked for your picture, but I can’t remember which girl you are.	我找了你的照片，但我不记得你是哪个女孩。
5.	He stood up and walked to the chair where she was sitting.	他站起来，走到她坐的桌子旁。
6.	The boy looked at the painting on the wall.	男孩看着墙上的画。
7.	He talked to his mother and said very nice things.	他和他妈妈说话时，说了很多好话。
8.	His mother and father explained that bad words were not polite.	他的父母解释说，说坏话是不礼貌的。
9.	The boy left the house and jumped over the river.	男孩离开家，跳进了河里。
10.	The girl gave him some change.	女孩给他一把零钱。
11.	The teacher talked for 20 minutes about school and being good students.	老师讲了二十分钟关于学校和做好学生的事情。
12.	Teacher told me to sit down.	老师让我坐下。

Appendix B

True/False Questions Adapted from Derwing et al. (2000)

1.	Doctors often work in hospitals.	医生经常在医院工作。
2.	Cabbages are usually highly intelligent.	白菜通常是高度聪明的。
3.	Mosquitos have soft pink fur.	蚊子的毛儿很软且是粉色的。
4.	Refrigerators keep food extremely hot.	冰箱让食物非常热。
5.	Some people have sandwiches for lunch.	有些人午餐吃三明治。
6.	It is good to eat rocks for lunch.	午餐吃石头是很好的。
7.	Most sailors keep their boats at the airport.	大多数水手将他们的船停在机场。
8.	You can use credit cards at many stores.	您可以在许多商店使用信用卡。
9.	Some people like to read poetry.	有些人喜欢读诗歌。
10.	Some people walk on their ears.	有些人在耳边行走。
11.	Some people play the guitar.	一些人明星在弹吉他。
12.	Exercise will make you fat.	运动会使你发胖。
13.	Many children’s books have pictures.	许多儿童读物都有图片。
14.	Most cars are made from milk.	大多数汽车是用牛奶制成的。
15.	Many people listen with their feet.	许多人用脚倾听。
16.	Most animals need air to breathe.	大多数动物都需要空气呼吸。
17.	Most caterpillars turn into butterflies.	大多数毛毛虫变成蝴蝶。
18.	Children often own large companies.	孩子们经常拥有大公司。
19.	Many people think that babies are cute.	许多人认为婴儿很可爱。
20.	Some people need to wear glasses.	有些人需要戴眼镜。
21.	Most grandmothers ride motorcycles.	大多数祖母骑摩托车。
22.	You can buy many things at the mall.	你可以在商场买很多东西。
23.	Cars generally need gas to run.	汽车通常需要汽油才能行驶。
24.	People play soccer with a piano.	人们用钢琴打足球。
25.	Sugar is bad for your teeth.	糖对牙齿有害。
26.	Some babies enjoy reading novels.	一些婴儿喜欢看小说。
27.	Rabbits usually have big wings.	兔子通常有大翅膀。
28.	Many people enjoy looking at paintings.	许多人喜欢看画。
29.	June is the first month of the year.	六月是一年中的第一个月。
30.	You can watch a movie on the radio.	您可以在收音机上看电影。
31.	In the winter, the snow is green.	在冬天，雪是绿色的。
32.	Most mothers think their children are ugly.	大多数母亲认为自己的孩子很丑。

33.	Most desks are made from spaghetti.	大多数书桌都由意大利面条制成。
34.	Lazy people work very hard.	懒的人非常努力。
35.	Adults are usually younger than children.	成人通常比儿童年轻。
36.	Rocks make a delicious soup.	石头能做成美味的汤。
37.	The earth is the shape of a triangle.	地球是三角形的形状。
38.	You can see animals at the zoo.	您可以在动物园看到动物。
39.	Some people eat shoes for a snack.	有人把吃鞋当零食。
40.	You can borrow a bicycle from the library.	您可以从图书馆借自行车。
41.	Trucks drive on the highway.	卡车在高速公路上行驶。
42.	Some people like to watch television.	有些人喜欢看电视。
43.	Many people collect postage stamps.	许多人收集邮票。
44.	Some cows like to read books.	有些母牛喜欢读书。
45.	Crayons come in many colours.	蜡笔有多种颜色。
46.	Most children like to eat cookies.	大多数孩子喜欢吃饼干。
47.	Some people find music relaxing.	有些人发现音乐让人放松。
48.	You can buy bread at the store.	您可以在商店买面包。
49.	Some flowers bloom in the summer.	有些花在夏天开。
50.	Some clothes are made from cotton.	有些衣服是棉花做的。
51.	You can write with a pen or pencil.	您可以用钢笔或铅笔写字。
52.	Stepping on a nail can hurt.	踩指甲可能会很疼。
53.	Some people live in big cities.	有些人住在大城市。
54.	Most babies like to drink milk.	大多数婴儿喜欢喝牛奶。
55.	Many teachers ride cows to work.	许多老师骑牛上班。
56.	Most people love to go to the dentist.	大多数人都喜欢去看牙医。
57.	Police often wear pants.	警察经常穿裤子。
58.	A raincoat makes an excellent bathing suit.	雨衣是极好的泳衣。

Appendix C

Interview Questions

1. Describe your experience using Google Translate (TTS and ASR) in this study. How did you like using it?

(Probing questions) What did you do in your own words? What process did you use for learning Mandarin? How did this process unfold? How much time did you use? Did you do everything in one go or over multiple sessions (and how many sessions?) What strategies did you use? What caused your decisions? What effects occurred based on your use of the technologies?

2. What were your perceptions about the learning experience, good or bad (Google Translate, Moodle, ASR, TTS)?

(Probing questions) Did you enjoy using any part? Which aspects were more or less enjoyable? What effects did your enjoyment have on your processes? What effects did it have on your learning?

3. What were your perceptions of the learning tools? Did the practice and quizzes help your Mandarin learning? Or would you have learned more without them?

Probing questions) How was the instruction in your opinion? Was it necessary? Would you have preferred more or less? How did the level of instruction effect your learning? Your strategies? How would increasing, decreasing, or changing the instructional methods have affected your learning?

4. Why did you learn Mandarin Chinese? What aspects of Mandarin did you find difficult (tones, consonants, vowels, syllables, words, phrases, everything)? Describe them.

(Probing questions) In your own words, explain what parts you found difficult? What parts were easy? Were there any aspects you noticed quickly? Were there any aspects you noticed but did not consider relevant? Describe in detail what aspects you found easy/difficult. How did you address them in your learning? How do you feel about your ability to learn these aspects?

5. What kind of strategies did you use for learning? Did you only follow the tutorials, or did you try your own strategies?

(Probing questions) Describe any strategy you used, whether provided or your own idea. Describe the effects these strategies had on your learning. How did their use unfold during your learning experience? How did you develop each strategy? What processes did you use? What caused you to try a new strategy? What caused you to stick with a previous strategy? How did you monitor yourself?

6. Did you feel the level of instruction was helpful? How would you change it if you were in total control?

(Probing questions) Would you have preferred more or less instruction? How did the level of instruction affect your motivation and interest? What kinds of changes would you make in your own learning in the future? Describe your process for following or disregarding provided instructions. What effects did you see on your learning and your motivation?

6. What types of feedback did you receive on your language? Were they useful in your language learning?

(Probing questions) Did you notice any feedback? Did it affect your learning? Would you prefer more or less? How did the feedback affect your learning? Did it affect motivation, usability, learnability, or your willingness to continue? What process did you use to incorporate it into your learning?

7. How challenging was it for Google Translate to understand you or you to understand it?

(Probing questions) How did you use Google Translate? Was it easy or difficult? Frustrating or fair? How did you incorporate Google Translate's abilities (good or bad) into your learning? How did it affect you concerning the four constructs?

8. Would you ever use Google Translate to learn language on your own in the future?

(Probing questions) Would you use it in the future? What languages would you like to study with it? What scenarios do you think it works best in? Looking at the four constructs, how would each apply to you learning using this system again?