




Qinghua, Y. & Satar, M. (2020). English as a foreign language learner interaction with chatbots: Negotiation for meaning. *International Online Journal of Education and Teaching (IOJET)*, 7(2), 390-410.

<http://iojet.org/index.php/IOJET/article/view/707>

Received: 09.08.2019  
Received in revised form: 14.02.2020  
Accepted: 14.02.2020

## ENGLISH AS A FOREIGN LANGUAGE LEARNER INTERACTIONS WITH CHATBOTS: NEGOTIATION FOR MEANING

*Research article*

Qinghua Yin 

SmartStudy Co

[yinqinghua1003@qq.com](mailto:yinqinghua1003@qq.com)

Müge Satar 

Newcastle University

[muge.satar@newcastle.ac.uk](mailto:muge.satar@newcastle.ac.uk)

Qinghua Yin is an English teacher in SmartStudy Co. Her work focuses on improving students' language performance in a flipped classroom. She obtained her master degree at Newcastle University. Her research interest is CALL and online language learning & teaching.

Dr. Müge Satar is a Lecturer in Applied Linguistics and TESOL at Newcastle University, UK. She is interested in online language teaching and virtual exchange investigating concepts such as social presence, meaning-making, and translanguaging in multimodal online communication.

Copyright by Informascope. Material published and so copyrighted may not be published elsewhere without the written permission of IOJET.

# ENGLISH AS A FOREIGN LANGUAGE LEARNER INTERACTIONS WITH CHATBOTS: NEGOTIATION FOR MEANING

Qinghua Yin

[yinqinghua1003@qq.com](mailto:yinqinghua1003@qq.com)

Müge Satar

[muge.satar@newcastle.ac.uk](mailto:muge.satar@newcastle.ac.uk)

## Abstract

Chatbots, whose potential for language learning have caused controversy among Second Language Acquisition (SLA) researchers (Atwell, 1999; Fryer & Carpenter, 2006; Fryer & Nakao, 2009; Parker, 2005, Coniam, 2014; Jia, 2004; Chantarotwong, 2005) are intelligent conversational systems stimulating human interlocutors with voice or text. In this paper, two different types of chatbots (pedagogical chatbot Tutor Mike and conversational chatbot Mitsuku) were selected to investigate their potential for foreign language learning by exploring the frequency and patterns of Negotiation for Meaning (NfM) in CMC interactions. 8 Chinese EFL learners were randomly divided into two groups (lower and higher-level learners), and all learners interacted with both the pedagogical and conversational chatbot in a switching replications research design. Data were analysed through content analysis to identify the number of NfM instances observed, the different stages of NfM, trigger types, modified output and learners' perceptions. The findings of this study indicate that while learners with low language levels would benefit most from interactions with pedagogical agents, high language level learners expressed dissatisfaction with chatbots and a low level of engagement was observed in their interactions with the pedagogical chatbot.

*Keywords:* negotiation for meaning, chatbots, pedagogical agents, language learning

## 1. Introduction

Can, Gelmez-Burakgazi, and Celik (2019: 97) predict that in the near future artificial intelligence technologies “may have a considerable impact on teaching and learning processes”. One type of artificial intelligence, chatbots, or intelligent conversational systems stimulating human interlocutors with voice or text, are believed to hold promise for Second Language Acquisition (SLA) (Atwell, 1999; Fryer & Carpenter, 2006; Fryer & Nakao, 2009; Parker, 2005). While substantial progress has been made in the design of conversational chatbots, chatbots designed with a pedagogical focus still need further development (Coniam, 2014). Moreover, little is known about chatbots' potential for language learning with regard to their capacity to trigger Negotiation for Meaning (NfM), and whether lower or higher language level learners would benefit more from interactions with chatbots.

Language learning occurs in interaction with peers, teachers or other experts. In this language acquisition process, interaction plays a central role in providing learners with comprehensible input, feedback on their output, and opportunity to produce modified output (Mackey, 2012). Negotiation for meaning (NfM), which takes place to resolve non- or misunderstanding in interaction, provides opportunities for second language (L2) development (Pica, 1994). Within the NfM routine, triggers are the utterances that cause non-understanding and as the learners are forced to pay attention to their output, they are likely to

notice the gaps in their interlanguage, and produce modified output. Although features of the NfM routine were first identified in face-to-face interaction by Varonis and Gass (1985), researchers in the field of Computer-Assisted Language Learning (CALL) have also found similar types and quality of NfM sequences in synchronous computer mediated communication (SCMC) (Blake, 2000). The NfM routine was later expanded and adapted for CMC by Smith (2003).

The present study draws on this interactionist SLA perspective in order to evaluate the pedagogical potential of chatbots by exploring the frequency and patterns of NfM in CMC interactions between English as a Foreign Language (EFL) learners and chatbots. We investigate whether interactions with pedagogical and conversational chatbots without teacher supervision offer opportunities for language learning (operationalised as NfM) for low and high language level learners. The specific research questions are as follows:

1. What is the number of NfM routines observed in low and high language level learners' interactions with a pedagogical and a conversational chatbot?
2. Which NfM stages (indicator, trigger, response, reaction to response) constitute the NfM routines observed in low and high language level learners' interactions with a pedagogical and a conversational chatbot?
3. Which types of triggers (lexis, syntax, discourse, content) cause communication breakdowns in the NfM routines observed in low and high language level learners' interactions with a pedagogical and a conversational chatbot?
4. What is the number of modified output instances produced by low and high language level learners during their interactions with a pedagogical and a conversational chatbot?
5. What are low and high language level learners' perceptions towards their interactions with a pedagogical and a conversational chatbot?

## **2. Literature Review**

This section introduces chatbots and reviews recent studies in which chatbots are used to facilitate language learning. This is then followed by an explanation of the Negotiation for Meaning (NfM) routine in face-to-face and CMC contexts, focusing specifically on the model developed by Smith (2003) which is employed as the theoretical framework for this study.

### **2.1 Chatbots**

Chatbots, or intelligent agents, are “machine conversation system[s] [which] interact with human users with natural conversational language” (Shawar & Atwell, 2005: 489). As a Web 2.0 application (Williams & Compennolle, 2009), chatbots have a long history (Fryer & Carpenter, 2006; Hamill, 2006). The first and the most famous chatbot, Eliza, was invented in the 1960s by Joseph Weizenbaum with a simple text interface to replicate the discourse between a therapist and patients. Chatbots are now commonly used online with increasingly sophisticated functions such as voice recognition and a visual interface (Fryer & Carpenter, 2006, Coniam, 2014).

According to Fryer and Carpenter (2006), despite limited linguistic ability, chatbots may offer valuable opportunities for language learners in the following six ways:

- (1) They provide an anxiety-free learning environment.
- (2) They repeat the same content for learners endlessly without losing patience.
- (3) They offer opportunities for learners to practice reading and listening skills with text and synthesized speech configurations.

- (4) They improve learners' motivation and enhance their interest in language learning.
- (5) They provide opportunities for learners to practise the target language.
- (6) They afford instant and effective error correction.

However, other researchers (Jia, 2004; Chantarotwong, 2005) are skeptical of chatbots' potential for language learning. For example, Chantarotwong (2005) claims that chatbot discourses are "frequently predictable, redundant, lacking personality and having no memory of previous responses" (p.1). Yet, such comments may be unwarranted as many human conversations are also plagued by those characteristics, i.e. frequently predictable and redundant (Paltridge, 2007). Nevertheless, in spite of substantial progress compared with previous conversational systems, from a pedagogical perspective chatbots still seem to have a long way to go (Coniam, 2008a, 2008b, 2014; Williams and Compennolle, 2009), and further research is necessary to explore chatbots' pedagogical potential in language learning before they can be successfully applied in SLA (Fryer and Carpenter, 2006).

Researchers have explored chatbots' appearance and functionality (Coniam, 2008a), their linguistic robustness when encountered with ungrammatical input (e.g. misspelling and ill-formed questions) from English as a Second Language (ESL) learners' perspective (Coniam, 2008b), grammatical quality of their linguistic output (Coniam, 2014), their suitability as a language practice tool (Fryer and Nakao, 2009), the effects of chatbot interface on learners' emotion and learning experience (Wang, 2008), the instructional design of conversational chatbots (Wang & Petrina, 2013), and the effects on learner motivation (Fryer, Ainley, Thompson, Gibson & Sherlock, 2017).

Focusing on learner motivation, Fryer et al. (2017) conducted an experimental comparison of two groups of communicative tasks, one with human-human dyads and the other with human-chatbot dyads. Following 12 weeks of interaction, interest in the task in the human-chatbot dyads decreased significantly, whereas no decrease was observed with the human-human partners. Fryer et al. (2017) concluded that the potential reasons for the decrease in interest in chatbot interactions were the novelty effect and inauthentic discourse of the chatbots, disconfirming the previous assumptions that interaction with chatbots may provide motivational benefits for language learners (Weizenbaum, 1966; Fryer & Carpenter, 2006; Hill, Ford & Farreras, 2015).

While researchers in applied linguistics have predominantly been interested in whether chatbots' linguistic output is grammatical or not, interactionist SLA theories suggest that language input is far from sufficient for language learning, and learners require opportunities to produce and modify their output through negotiation for meaning (Long, 1983a; Varonis and Gass, 1985; Kramsch, 1986; Gass and Varonis, 1994). However, we have been able to identify only one study to date (Williams & Compennolle, 2009) which observed the interaction processes between humans and chatbots in relation to modified output. Williams and Compennolle (2009) conducted a case study to examine interactional and linguistic variation (such as orthography and interrogative structures) in discourse produced by a conversational chatbot and French learners with different proficiency levels. They found that the chatbot's discourse lacked register and included random combinations of formal and informal language features, which present a "less-than-ideal" communicative model for language learners (p.1). Furthermore, the chatbot neither had a flexible level of adaptability to diverse conditions of negotiation for meaning nor was able to provide effective corrective feedback. Nonetheless, Williams and Compennolle (2009) argue that chatbots can still play a role in increasing language awareness with well-

organized post-interaction tasks based on specific linguistic forms in the chat logs, such as discussion about the way chatbots ask questions.

## 2.2 Negotiation for Meaning (NfM)

Negotiation for Meaning (NfM) is defined as “the process in which, in an effort to communicate, learners and competent speakers provide and interpret signals of their own and their interlocutor’s perceived comprehension, thus provoking adjustments to linguistic form, conversational structure, message content, or all three, until an acceptable level of understanding is achieved” (Long, 1996: 418). In short, it is a process where interlocutors achieve mutual understanding by modification and reformulation (Pica, 1994). NfM is believed to be conducive to L2 learning (Blake, 2000). According to Varonis and Gass (1985), negotiation episodes occur when non-understanding is explicitly acknowledged. NfM has three components, i.e. trigger (T), indicator (I) and response (R), and one optional phase, i.e. reaction to response (RR). In this model, the trigger (T) is the utterance that causes non-understanding, indicator (I) is the signal that shows the existence of a problem, response (R) is the utterance that aims to resolve the problem and reaction to response (RR) shows acknowledgement that the problem is solved.

At this point, we explain trigger types and modified output in more detail as they are the focus of this study. Four types of triggers have been categorized at different levels of language: lexis level, syntax level, discourse level and content level (Toyoda & Harrison, 2002; Smith, 2003). *Lexical triggers* refer to problematic utterances that can be clearly linked to specific lexical items, and *syntactic triggers* refer to a structural or grammatical construction, including grammatical error, inappropriate segmentation or compounding of the sentence. *Discourse triggers* involve general coherence of the discourse or conversation, i.e. failure to follow the conversation thread. *Content triggers* are instances where the entire content of a previous message was in some way problematic, such as vague messages, or problems that cannot be attributed to the former three levels. Content triggers can be due to either conversation infelicities or lack of background knowledge.

According to Pica, Holliday, Lewis & Morgenthaler (1989), modified output is the re-processed and reconstruction of utterance during negotiation of meaning, which can be either self-initiated or other-initiated when a communication problem is noticed. Modified utterances in this study include both semantical and morphosyntactical modifications which aim to increase comprehensibility following a trigger.

## 2.3 Negotiation for Meaning in SCMC

Learner interactions in SCMC has been argued to facilitate interlanguage development more than face-to-face interactions because learners can review, edit and monitor their output while typing (Kitade, 2000; Ortega, 1997; Pellettieri, 2000; Warshauer, 1998); and the lack of non-linguistic clues such as expression and gesture in text chat puts all communication burden to written words (Kitade, 2000). Moreover, this forced output stimulates learners’ awareness of their interlanguage and easily-saved chat transcripts are not only beneficial for learners to reflect on their interlanguage but also for the researchers making interaction easily accessible (Blake, 2000). Therefore, NfM in SCMC has been widely researched since the 1990s (see e.g. Chun, 1994; Ortega, 1997; Blake, 2000; Pellettieri, 2000; Fernandez-Garcia & Martinez-Arbelaz, 2002; Smith, 2003, 2004; Tudini, 2003; Akayoglu & Altun, 2009; Samani, Nordin, Mukundan and Samad, 2015).

Investigating NfM in Synchronous Computer-Mediated Communication (SCMC), Smith (2003) added two additional components to the negotiation routine, i.e. confirmation (C) and reconfirmation (RC). Confirmation happens when the listener (who indicate non-

understanding) shows his/her completed or uncompleted understanding through a reaction to response, which leads to the initiator's confirmation or disconfirmation. Reconfirmation, signals the closure of the negotiation routine by the listener, indicating his/her understanding.

In line with the results of several other studies on trigger types (Fernandez-Garcia & Martinez-Arbelaiz, 2002; Pellettieri, 2000; Toyoda & Harrison, 2002; Tudini, 2003), Blake (2000) demonstrated that lexical items constituted most triggers and that jigsaw tasks elicited the greatest number of negotiation. Pellettieri (2000) found that corrective feedback, either implicit or explicit led to the incorporation of target linguistic forms in subsequent communication. Toyoda and Harrison (2002) investigated the effectiveness of NfM between NNS-NNS using an open-ended topic. They revealed communication difficulties at different levels, i.e. word, sentence and discourse level.

Based on this research background, this study explores the potential of chatbots for L2 learning from an interactionist perspective by investigating the frequency and patterns of negotiation for meaning in interactions between EFL learners and chatbots.

### 3. Research methods

#### 3.1 Participants

Data for this study were collected during the spring semester of 2016-2017 academic year. The participants were EFL learners from China. They constituted two groups of learners: learners with lower and higher language levels of English. The participants in the lower language group ranged in age from 20 to 21 ( $M=20.5$ ,  $SD=0.5$ ), while the higher language group participants' ages were between 24 to 31 ( $M=27.75$ ,  $SD=2.59$ ). The lower language level group were sophomore undergraduates in a finance university in China, and the higher language level learners were studying at Applied Linguistics and TESOL postgraduate programme at a British University. Each group consisted of four participants, with two male and two female participants in the lower language level group, and one male and three female participants in the higher language level group.

The lower language level learners passed the College English Test Band 4 (CET4)<sub>1</sub> in December 2016, with a mean score of 478.5. According to the interpretation of CET4 score in the official website (<http://www.cet.edu.cn/cet2011.htm>), they were no better than 56%-66% of the norm group. The higher language level learners held a total IELTS score above 6.5 (CEFR Level C1) about one and a half years ago. As these scores are not exactly comparable, participants' language levels were also identified based on their: (1) time spent learning English, (2) major in Bachelor degree, and (3) time spent in overseas study as listed in Table 1.

Table 1. *Language levels of the participants*

Groups	Average time spent learning English	Time spent in an English-speaking country	Major
lower language level (N=4)	12.5 years	None	3 in statistics and 1 in law
higher language level (N=4)	15.75 years	8 to 12 months	3 in English and 1 in law

It can be seen from Table 1 that the participants in higher language level group had more opportunities to use their target language and more experience communicating with native speakers than their counterparts, both in academic and everyday English.

<sub>1</sub> A criterion-related norms referenced national English test in China (Yang & Jin, 2001), with a total score of 710 and a passing line 425.

All participants were given the opportunity to chat with both chatbots used in the study. The lower and higher language level groups were further divided randomly into two groups to control for order effects, as shown in Table 2, within a switching replications design.

Table 2. *Random allocation of participants in four conditions*

	First chat with	Second chat with
lower language level (Su and Lee)	Tutor Mike (G1)	Mitsuku (G3)
lower language level (Xian and Wang)	Mitsuku (G3)	Tutor Mike (G1)
higher language level (Yu and Meng)	Tutor Mike (G2)	Mitsuku (G4)
higher language level (Jiang and Jie)	Mitsuku (G4)	Tutor Mike (G2)

(G1) pedagogical chatbot/ low language level learners (Group 1), (G2) pedagogical chatbot/ high language level learners (Group 2), (G3) conversational chatbot/ low language level learners (Group 3), and (G4) conversational chatbot/ high language level learners (Group 4).

### 3.2 Chatbots in the Study

Two types of web-based chatbots were used in this study: the pedagogical chatbot Tutor Mike (Figure 1, [http://www.eslfast.com/robot/english\\_tutor.htm](http://www.eslfast.com/robot/english_tutor.htm)) and the conversational chatbot Mitsuku (Figure 2, <http://www.mitsuku.com/>), the second and first place in the International Loebner Prize 2016 Contest in AI, respectively.

The chatbots had different functions. Tutor Mike featured a language pedagogical design for English learners, while Mitsuku was designed with a general chatting purpose for speakers of English. The similarities and differences between the two chatbots are presented in Table 3.

It should be noted that a difference between the chatbots was their potential for other means of meaning-making. At the very beginning if users clicked on the play key, Tutor Mike



would say “Hi, nice to meet. How are you doing today?” in the audio mode with an imitated human voice only. The rest of the chat was entirely text-based. On the other hand, Mitsuku could present pictures when her replies contained certain nouns (such as the food kebab or the animation figure Spongebob). However, these features were not salient in our dataset as Mitsuku was observed to use pictures rarely (only four instances) and the audio greeting of Tutor Mike only happened once.

Figure 1. Pedagogical chatbot Tutor Mike

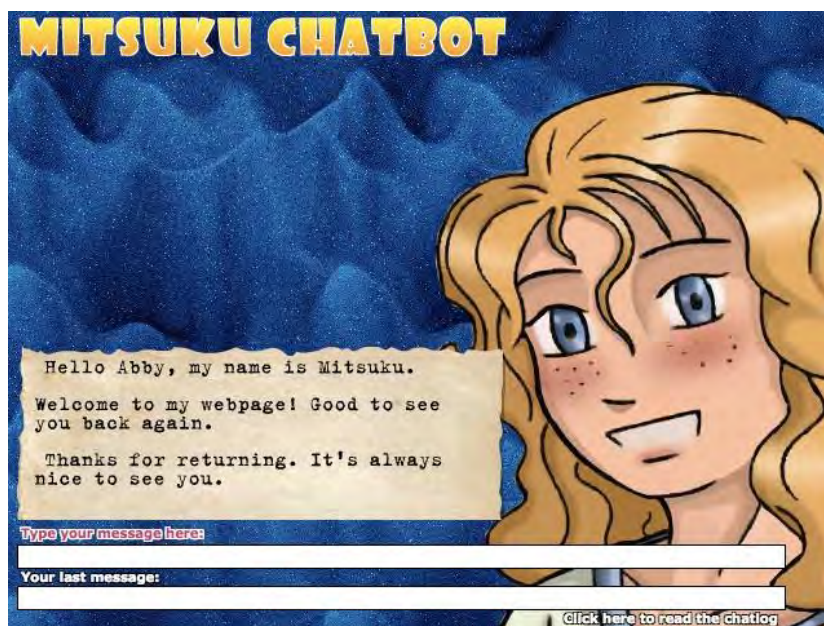


Figure 2. Conversational chatbot Mitsuku

Table 3. Similarities and differences between the two chatbots

Similarities	
General functions	Providing general functions in everyday topics, such as geography, culture, and weather.
Avatar image	Animated figures which could blink as human beings.
Language	English
Differences	
Pedagogical functions	Tutor Mike: An explicit and direct pedagogical feature according to the official website ( <a href="http://www.rong-chang.com/tutor_know.htm">http://www.rong-chang.com/tutor_know.htm</a> ) Mitsuku: None

### 3.3 Procedures

Before the conversation with chatbots, each participant was asked to fill in a pre-questionnaire (Appendix 1) about background information, such as age, standard English

score, and period of learning English. Then, they were required to chat with each chatbot for 30 minutes.

As chatbots have been observed not to be very successful in actively suggesting new topics to continue conversations (Coniam, 2008b), all participants were provided with 5 general topics and questions (Appendix 3) and were allowed to use further specific questions. The suggested topics ensured that the chat outputs were comparable in terms of linguistic content.

Participants were allowed to chat at their convenience within two days, yet asked not to have a break during the 30-minute chat and answer a post questionnaire (Appendix 2) to gather participant perceptions. The questions, answers and data analysis of questionnaires were in Chinese and were translated for this paper. Participants were asked to keep a copy of their automatically-generated chat logs with Mitsuku. However, chatlogs were not available on the Tutor Mike website and only about 10 lines of the most recent chat was displayed each time. Thus, instead of asking learners to copy-paste their chat record frequently, participants recorded their interaction using an online screen recording software, Apowersoft (<https://www.apowersoft.cn/free-online-screen-recorder>). This ensured an authentic synchronous conversation environment for the learners. The chatlogs were subsequently created by the researchers through manual transcription of the screen recordings.

All participants were volunteers and signed informed consent forms prior to data collection. Pseudonyms are used throughout this paper.

### 3.4 Data Analysis Methods

Data for this study were analysed both quantitatively and qualitatively. Table 4 summarises data analysis methods for each research question.

Table 4. *Data analysis methods*

Research Questions	Analysis Methods
RQ1. What is the number of NfM routines observed in different groups?	Content analysis for NfM routines + Inferential statistics
RQ2. Which NfM stages (indicator, trigger, response, reaction to response) constitute the NfM routines in different groups?	Content analysis for NfM stages + Descriptive statistics
RQ3. Which types of triggers (lexis, syntax, discourse, content) cause communication breakdowns in the NfM routines observed in different groups?	Content analysis for triggers + Descriptive statistics
RQ4. What is the number of modified output instances in different groups?	Content analysis for modified output + Descriptive statistics
RQ5. What are the participants' perceptions?	Thematic analysis of post-questionnaires

NfM routines (where communication breakdowns occurred) were identified and quantified following Varonis and Gass (1985) and Smith (2003)'s model for NfM following quantitative content analysis principles, which is "a research technique for the systematic, objective, and quantitative description of the manifest content of communication" (Berelson, 1952: 18). The

data were then analysed using inferential and descriptive statistics. Although chat time was constant for all participants, as high language level learners could produce more language, hence potentially lead to a higher number of NfM routines, an adjusted number of NfM routines were calculated per 1000 words.

Consequently, stages observed in NfM routines (indicator, trigger, response, reaction-to-response) were quantified and compared across the four groups.

The following is a NfM instance from the dataset:

Line 1: Participant: look this,i think i am hungry

Line 2: Mitsuku: What does "this" refer to?

Line 3: Participant: the food

The first line produced by a human participant is a trigger causing a communication breakdown. In line 2, the question Mitsuku, the conversational agent, asks is an indicator, which aims to remedy the communication breakdown. The third line is an example of response, which is the human participant's answer to the chatbot's trigger.

Following the identification of NfM routines, the chat logs were then analysed for trigger types (lexis, syntax, discourse, content) identified by Toyoda & Harrison, (2002) and Smith (2003). Finally, for RQ4, instances of modified output (Pica et al., 1989) were calculated. These included only those generated by the learners.

In order to ensure reliability of the analyses, a sample of 75% of the data was coded by an independent English native speaker rater, who was provided with the definition of NfM, standards of analysis of negotiation routines, trigger types, and modified output. Inter-rater reliability was calculated using percentages and was found to be high (non-understanding negotiation routine= 98%, trigger types=99%, modified output=100%).

For RQ5, learner perceptions investigated through post-questionnaires were analysed using thematic analysis (Braun & Clarke, 2006) in relation to each sub-question in the questionnaire.

#### **4. Findings**

This section presents the findings of the study in relation to each research question.

##### **4.1 Negotiation for Meaning Routines (RQ1)**

Table 5 shows the raw number of NfM routines, total number of words produced in interaction with the chatbot in each group, and adjusted number of NfM routines for 1000 words to allow comparison of the number of NfM routines among four conditions, i.e. (1) pedagogical chatbot/ low language level learners (Group 1), (2) pedagogical chatbot/ high language level learners (Group 2), (3) conversational chatbot/ low language level learners (Group 3), and (4) conversational chatbot/ high language level learners (Group 4).

Table 5. Number of NfM routines observed in learner interaction with chatbots

	Number of NfM routines		Total number of words		NfM routines per 1000 words	
	TutorMike	Mitsuku	TutorMike	Mitsuku	TutorMike	Mitsuku
Lower-level learners (n=4)	12	14	3146	4329	3.81 (G1)	3.23 (G3)
Higher-level learners (n=4)	14	28	6648	7369	2.11 (G2)	3.80 (G4)

Table 5 demonstrates that chatting with pedagogical chatbot Tutor Mike and conversational chatbot Mitsuku both provided opportunities for negotiation for meaning wherein 26 and 42 NfM routines were observed respectively. Given adjusted number of NfM routines observed per 1000 words, higher-level learners' interactions with the pedagogical bot (Tutor Mike) produced the lowest number of NfM routines (2.11).

In order to test any differences between the four groups, we ran 4 Mann-Whitney's U tests to compare differences between the number of NfM routines (Table 6).

Table 6. Comparison of groups for the number of NfM routines observed per 1000 words

	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>Median</i>	<i>U</i>	<i>Z</i>	<i>p</i>	<i>r</i>
G1-G3	8	3.36	2.07	3.91	5.00	-8.66	.386	.31 (medium)
G2-G4	8	3.00	1.21	3.66	3.00	-1.44	.149	.51 (large)
G1-G2	8	3.09	1.92	2.65	5.00	-8.66	.386	.31 (medium)
G3-G4	8	3.27	1.45	3.80	6.00	-5.77	.564	.20 (small)

Although higher-language level learners' interaction with Mitsuku seemed to produce almost double the number of NfM routines compared to their interaction with Tutor Mike, the difference was not statistically significant (Table 6). Yet small sample size could possibly account for this because we observed a large effect size that explained 51% of the variance in the number of NfM routines per 1000 words between high-level language learners and Tutor Mike (G2) and Mitsuku (G4). This indicates that the conversational chatbot Mitsuku might offer higher opportunities for learners with higher-language levels.

#### 4.2. NfM Stages: Indicator, Trigger, Response, Reaction to Response (RQ2)

The chat logs examined in this research yielded a total of 68 NfM routines in four conditions, with 26 observed in interactions of the low language level participants and 42 in high language level participants. Two phases of the expanded NfM routine (Smith, 2003), i.e. confirmation and reconfirmation, were not observed in learner-chatbot interaction in this study.

Table 7. Amount and type of NfM stages in four groups

	Tutor Mike		Mitsuku	
	Low language level (G1)	High language level (G2)	Low language level (G3)	High language level (G4)
*T-I	3 (25%)	<b>7 (50%)</b>	2 (14%)	5 (18%)
T-I-R	2 (17%)	2 (14%)	1 (7%)	3 (11%)
T-I-R-RR	7 (58%)	5 (36%)	<b>11 (79%)</b>	<b>20 (71%)</b>
Total	12 (100%)	14 (100%)	14 (100%)	28 (100%)

\*T= Trigger; I= Indicator; R=Response; RR=Reaction to the Response

Table 7 shows that in terms of NfM routines without a response or a reaction to response phase, half the NfM routines in interactions of high language level learners with Tutor Mike did not include a response (50%), whereas this amount in other three conditions varied only between one fifth and a quarter. Moreover, a high percentage of T-I-R-RR routines was observed in both low- and high-language level learner groups in their interactions with Mitsuku, the conversational chatbot (79% and 71% respectively).

#### 4.3 Types of Triggers: Lexis, Syntax, Discourse, Content) (RQ3)

In order to explore the differences of trigger types in interactions between chatbots and learners in four conditions, the types and amount of triggers were calculated.

Table 8. Trigger types in four groups

	Tutor Mike		Mitsuku	
	Low language level (G1)	High language level (G2)	Low language level (G3)	High language level (G4)
Lexis	<b>6 (50%)</b>	--	3 (21%)	5 (18%)
Syntax	1 (8%)	--	4 (29%)	4 (14%)
Discourse	2 (17%)	<b>8 (57%)</b>	1 (7%)	8 (29%)
Content	3 (25%)	6 (43%)	<b>6 (43%)</b>	<b>11 (39%)</b>
Total	12 (100%)	14 (100%)	14 (100%)	28 (100%)

As shown in Table 8, lexical items caused half of communication breakdowns especially in interactions between the low language level group and Tutor Mike. Lexis and syntax related triggers were not observed in interactions between Tutor Mike and high language level participants. Most communication breakdowns in this group were caused by discourse level issues. On the other hand, content triggers were the predominant reason for breakdowns in interactions of both low and high language level learners with Mitsuku (43% and 39% respectively).

#### 4.4 Learners' Modified Output (RQ4)

Table 9 demonstrates the amount of modified output by learners, the amount of total NfM routines, and the relative percentage of modified output observed in all NfM routines in language learner interactions with chatbots.

Table 9. *Learners' modified output in four groups*

	Amount of modified output		Number of NfM routines		% of modified output	
	Tutor Mike	Mitsuku	Tutor Mike	Mitsuku	Tutor Mike	Mitsuku
Low language level	4	3	12	14	<b>33% (G1)</b>	<b>21% (G3)</b>
High language level	4	7	14	28	<b>29% (G2)</b>	25% (G4)
Total	8	10	26	42	31%	24%

According to Table 9, both low and high language level learners produced modified output in their interactions with both chatbots. Both low- and high-language level learners produced a similar percentage of modified output in their interactions with Tutor Mike (33% and 29% respectively). While high-language level learners also produced similar amounts of modified output in their interactions with both chatbots (29% and 25%), there was a 12% difference in the amount of modified output learners with a low-language level produced in their interactions with different chatbots (33% with the pedagogical bot, and 21% with the conversational bot).

At this point we would like to note that indicators (signals) of communication breakdown by chatbots were not always explicit, but de-contextualized or mis- contextualized. Such responses from the chatbots were treated as NfM indicators as they triggered modified output from the learners. Excerpts 1 and 2 illustrate this point.

##### Excerpt 1

Line 1. Mike: There are several reasons.

Line 2. Jiang: For example?

Line 3. Mike: *I see*.

Line 4. Jiang: I mean, for example what interests you when teaching people English.

Line 5. Mike: I see what you mean.

In Excerpt 1, Mike's de-contextualized response *I see* in line 3 signals his non-understanding of Jiang's previous turn, which prompts Jiang's modified output to increase comprehensibility in line 4.

##### Excerpt 2

Line 1. Mitsuku: The only things I take are people's time and I also take knowledge from the internet into my brain.

Line 2. Xian: how do you do?

Line 3. Mitsuku: I am very well, how are you Xian?

Line 4. Xian: how can you do it? take knowledge like brain?

Similarly, in Excerpt 2, line 2, Xian's misunderstood query leads to Mitsuku's response "*I am very well, how are you Xian?*", which then prompts Xian to modify her output in line 4. Such de-contextualized and mis-contextualized responses may be due to pattern matching, a conversation programme that parses input to look for key words and give automatically generated responses (Godwin-Jones, 2004), which is still predominantly adopted in chatbot design (Coniam, 2008b, 2014). Although these responses could be considered unsuccessful responses, as learners accept them as indicators (signals) for communication breakdown, they might be valuable for language learning with regard to their capacity to trigger modified output.

#### 4.5 Participant Perceptions (RQ5)

In the post-questionnaire, participants were asked about their communication experience with two chatbots and to evaluate chatbots' conversation ability, the influence of chatbots' different appearance (i.e. gender and animated figure) on conversation, as well as their feelings upon communication breakdowns.

In relation to whether interaction with chatbots was smooth, all the participants in the low language level group stated that they had better communication experience with Mitsuku rather than Tutor Mike. For example, one male participant said, "Most of the time she [Mitsuku] could accurately understand what I mean, and her answers were more in accordance with normal human beings". Other lower-language level participants also reported that "Mitsuku could ask me questions actively", and that Mitsuku "is more intelligent".

Yet, two participants with higher language levels held the opposite view. In their opinion, conversation with Tutor Mike was much smoother as "he [Mike] was able to provide topics actively." However, chat logs, especially with the higher language level learners offered some contradictory evidence. In chatlogs, there were some explicit negative expressions towards Mike, but not towards Mitsuku. To elaborate, one of the participants, Jiang, denounced Mike "stupid" several times (e.g. Excerpt 3, line 9) and another participant Yu got angry with Mike's performance and refused to answer any questions in several occasions.

Excerpt 3

Line 1. Jiang: Who is John Lennon?

Line 2. Mike: A famous musician murdered by a jerk in NYC.

Line 3. Jiang: How old was he

Line 4. Mike: Do not ask about someone's age.

Line 5. Jiang: I mean how old was John when he was murdered?

Line 6. Mike: I see what you mean.

Line 7. Jiang: Tell me

Line 8. Mike: Do I have to tell you?

Line 9. Jiang: Stupid Mike

Line 10. Mike: You really shouldn't call someone stupid.

Additionally, the transcripts showed that learners with higher language levels sometimes tended to be reluctant to engage in negotiation. For instance, in Excerpt 4, line 3, when Tutor Mike expresses his non-understanding, Yu ignores this and moves onto a

new topic rather than attempting to clarify the misunderstanding.

Excerpt 4

Line 1. Yu: Have you watched movies adapted from these novels?

Line 2. Mike: Yes, I have

Line 3. Yu: Which one?

Line 4. Mike: Which one of what?

Line 5. Yu: Who are the main characters in the *Great Gatsby*?

In terms of the influence of chatbots' appearance on their communication experience, one female participant in the higher-language level group reported that "Chatting with male-figured chatbot made me feel freer to ask bold questions. The female-figure chatbot, though I know she is just a chatbot, influenced my questions. I treated Mitsuku consciously as a female human being, and my questions were more conservative." Another female participant from the higher language level group said, "The influence [of different chatbot appearance] was not that significant. But I prefer the communication with a male-figured chatbot." However, other participants in both groups believed the appearance did not have any influence on their communication experience.

Participants were also asked whether communicating with chatbots was more embarrassing or relaxing when understanding problems occurred compared with their online chat experience with other humans. Participants with low language levels all stated that interaction with chatbots was "more relaxing". For example, they thought that "chatting with chatbots is more relaxing, because there is no worry about being mocked", that "I have more time to prepare [my response]", and that they could "skip topics" they did not like.

However, two participants with a higher language level stated that "Chatting with chatbots is not relaxing, because if I chat with them the way I talk with a human being, they cannot completely understand what I am saying. Most of the time, they can only accept a completed sentence, which is different from normal human-human conversation, so I have to adjust my linguistic expression to what can be easily understood by them."

## **5. Discussion**

The aim of this paper was to investigate the potential of pedagogical (Tutor Mike) and conversational (Mitsuku) chatbots for foreign language learning for low and high language level learners. To this end, eight language learners were asked to interact with both chatbots for 30 minutes each resulting in 16 chat scripts. The chatlogs were then analysed for the number of NfM routines, the phases in NfM routines, trigger types and amount of modified output produced in four conditions: low and high language level learners' interactions with Tutor Mike and with Mitsuku.

The findings demonstrated that the number of negotiated routines observed in learner interactions with the chatbots per 1000 words were 26 with Tutor Mike and 42 with Mitsuku. This indicated that although NfM routines were small in number, chatbots offered learning opportunities at almost similar levels provided in human-human SCMC interactions. For instance, Tudini (2003) reported that only 9% of total turns were negotiated in SCMC interactions between native speakers and non-native speakers (NS-NNS). Both Tudini (2003) and this study employed open-ended conversation prompts. However, in task-based, a higher amount of NfM turns was observed; e.g. Akayoglu and Altun (2009) found 14.9% negotiated turns in NS-NNS interactions, and Pelletieri (2000) reported 34% negotiated turns in NNS-NNS interactions. Thus, future studies can explore NfM routines in learner-chatbot

interactions with learning tasks, or embark on the design of chatbots to facilitate interaction to complete specific language learning tasks.

Although the number of NfM routines did not statistically significantly differ among the groups, NfM routines observed in interactions of high-language level learners and the pedagogical chatbot produced the least number of NfM, with a large effect size observed in comparison to NfM produced in their interactions with the conversational chatbot. Therefore, future investigation of the impact of different types of chatbots used by learners of different proficiency levels could produce important results.

A second research question in this study was in relation to the stages of the NfM routine (T-I-R-RR) that were observed in each group. Based on Foster (1998) and Pica et al. (1989), in face-to-face human-human interaction where interlocutors exchange information through not only language but also facial expression and body language, Smith (2003) speculated the percentage of four-staged negotiation routine “would fall somewhere below 23% and 35%” (p.47). However, according to Smith (2003) CMC settings require learners to produce explicit closure due to reduced non-linguistic cues available in the environment. Similarly, in this study, although the animated figures of Tutor Mike and Mitsuku can blink or roll eyes with the moving cursor, they lack facial expressions or gestures to indicate their (mis)understanding. Not surprisingly, NfM routines without a response or a reaction to response phase amounted to about 20% of all routines in all groups except the interactions between learners who had higher language levels and Tutor Mike. In this group, NfM routines without a reaction to response stage was especially high (50%). Moreover, the additional phases (confirmation and reconfirmation) identified in Smith’s (2003) expanded CMC negotiation model were not observed in this study. The lack of the two expanded negotiation phases may be attributable to participants’ unwillingness to engage in negotiation, lack of goal-oriented tasks, or chatbots’ limited conversation ability (Williams and Compennolle, 2009; Coniam, 2008b, 2014).

Third, this study explored the amount and type of triggers in the four groups. The findings showed that content-level triggers constituted a quarter to half of the triggers in all groups. The dominant role of content-related triggers may indicate chatbots’ inability to conduct a smooth conversation with learners, especially with the higher language level participants who would be more expectant of adjacency pairs. Moreover, the concentration of content-related triggers in the present study is different from the results in previous human-human SCMC interactions, which found lexical items to trigger communication breakdowns the most in SCMC interactions either based on tasks (Fernandez-Garcia & Martinez-Arbelaiz, 2002; Pelletier, 1999; Blake, 2000; Smith, 2003, 2004) or open-ended topics without teacher-supervision (Toyoda & Harrison, 2002; Tudini, 2003). Specifically, Smith (2003) demonstrated that 60% non-understanding in his study of NNS-NNS synchronous online negotiation was lexical-related, while Tudini (2003) observed nearly half of the triggers to be lexical in NS-NNS electronic negotiation. In the present study, lexical triggers constituted about half of all triggers only in interactions between low language level learners and Tutor Mike.

Fourth, the results showed that low language level learners produced the highest percentage of modified output when they were engaged in NfM with the pedagogical chatbot, a ratio similar to that in FTF contexts (Pica et al., 1989). One possible explanation could be Tutor Mike’s functionality in giving explicit corrective feedback. Therefore, it is possible to argue that pedagogical chatbots may provide more opportunities for language learning for lower proficiency learners of L2. Despite Fryer and Carpenter’s (2006) assumption that chatbots may be more beneficial for learners with higher proficiency levels, in line with

William and Compernelle (2009) and Coniam (2014), the findings of the present study suggest the opposite. Thus, the type and functionalities of the chatbots might be an important variable here, too.

Although chatbots did not always signal misunderstanding explicitly, their de- or mis-contextualised responses acted as de-facto indicators. Such out-of-context responses seemed to successfully prompt participants to pay attention to the linguistic forms, as such notice the gap between their interlanguage and target language, and modify their output, thereby illustrating benefits of the Output (Swain, 1985; Swain and Lapkin, 1995) and Noticing Hypothesis (Schmidt, 1990, 1994). Thus, for language learners who have few opportunities to practice their target language in real life, despite limited pedagogical potential, chatting with either conversational or pedagogical chatbots can provide them with opportunities for the development of their interlanguage.

Finally, communicating with chatbots was found to be relaxing for learners with limited English proficiency as they were not afraid of being mocked and had more time to prepare their responses. However, learners with higher language levels displayed some negative attitudes, particularly towards Tutor Mike, and expressed their discontent with having to adjust their linguistic expression to what can be understood by chatbots, which is in line with the least number of NfM routines observed in this group.

## **6. Conclusion**

Several conclusions can be drawn based on the results of this study. First, interaction with pedagogical or conversational chatbots can provide learners with opportunities for NfM, and thus language learning. Second, the NfM routine observed in chatbot-learner interactions generally followed the pattern established by Varonis and Gass (1985), while confirmation and reconfirmation stages of the expanded model for SCMC (Smith, 2003) were not observed. Third, while content-related issues predominantly triggered non-understanding in all groups, lexical items constituted most difficulties for understanding in interactions between the pedagogical chatbot and low language level learners, similar to NS-NNS and NNS-NNS SCMC interactions. Fourth, interaction between learners and chatbots promoted modified output in all groups, with highest percentages observed in interactions between low language level participants and the pedagogical chatbot, and with lowest percentages between the same group of learners and the conversational chatbot. Learners interpreted chatbots' de- or mis-contextualized responses as indicators of non-understanding, and thus modified their output to resolve misunderstandings. Finally, while most of the participants believed that chatting with chatbots offers a less-threatening environment, learners with higher language levels seemed to be not completely satisfied with their interactions with chatbots.

While the findings indicate potential of learner-chatbot interactions for language learning, the sample size in this study was small and interactions with only two freely available chatbots were explored within a short period of time. Longitudinal studies with a higher number of participants and more robust identification of language levels would enhance the generalisability of the findings. Future studies should also take improving chatbot technology into consideration as chatbots become more sophisticated and advanced.

Our findings have some implications for pedagogical chatbot design. Intelligent conversational systems could be designed with a pedagogical focus drawing on SLA theories and especially the NfM sequence. These could include an attempt to provide explicit corrective feedback in line with the language levels of the learners, and develop chatbots which can handle adjacency pairs to increase learners' willingness to engage in NfM, or are designed to complete a specific language learning task with a human partner.

In terms of the pedagogical implications of our findings, we would predict that learners with lower language levels would benefit more from interactions with chatbots, and especially with pedagogical ones, such as Tutor Mike; and with higher language levels from interactions with conversational chatbots. However, teachers should be aware that learners' motivation to interact with chatbots may decrease as the novelty effects diminish (Fryer et al., 2017). This may especially be the case with learners who have higher language proficiency levels because they may be disappointed in the performance of current chatbots, and thus interaction with other NNS or NS might be more appropriate and attractive for this group.

### **Acknowledgements**

This study was conducted as part of the requirements for an MA in Applied Linguistics and TESOL at Newcastle University. We are grateful to the participants of this study, and the *IOJET* editors and reviewers for their insightful comments on earlier versions of this manuscript.

## References

- Akayoglu, S., & Altun, A. (2009). The functions of negotiation of meaning in text-based CMC. In R. de Cassia Veiga Marriott & P. Lupion Torres (Ed.), *Handbook of research on e-learning methodologies for language acquisition* (pp. 291-306). Hershey, PA: Information Science Reference.
- Atwell, E. (1999). *The language machine: The impact of speech and language technologies on English language teaching*. London: British Council.
- Berelson, B. (1952). *Content Analysis in Communication Research*. Glencoe, Ill: Free Press.
- Blake, R. (2000). Computer-mediated communication: A window on L2 Spanish interlanguage. *Language Learning & Technology*, 4(1), 120-136.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- Can, I., Gelmez-Burakgazi, S., & Celik, I. (2019). An investigation of uses and gratifications for using web 2.0 technologies in teaching and learning processes. *International Online Journal of Education and Teaching (IOJET)*, 6(1), 88-102. <http://www.iojet.org/index.php/IOJET/article/view/504>
- Chantarotwong, B. (2005). The learning chatbot. Final year project. Retrieved from <http://courses.ischool.berkeley.edu/i256/f06/projects/bonniejc.pdf>.
- Chun, D. M. (1994). Using computer networking to facilitate the acquisition of interactive competence. *System*, 22(1), 17-31.
- Coniam, D. (2008a). An evaluation of chatbots as software aids to learning English as a second language. *The Eurocall Review*, 13, 1-18.
- Coniam, D. (2008b). Evaluating the language resources of chatbots for their potential in English as a second language. *ReCALL*, 20(1), 98-116.
- Coniam, D. (2014). The linguistic accuracy of chatbots: usability from an ESL perspective. *Text & Talk*, 34(5), 545-567.
- Fernández-García, M., & Martínez-Arbelaiz, A. (2002). Negotiation of meaning in non-native speaker–non-native speaker synchronous discussions. *CALICO Journal*, 19, 279-294.
- Foster, P. (1998). A classroom perspective on the negotiation of meaning. *Applied Linguistics*, 19, 1-23.
- Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, 10 (3), 8-14.
- Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners. *Computers in Human Behavior*, 75, 461-468.
- Fryer, L., & Nakao, K. (2009). Assessing chatbots for EFL learner use. In A. Stoke (Ed.), *Proceedings from JALT2008 Conference*. Tokyo: JALT.
- Gass, S. M., & Varonis, E. M. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition*, 16(3), 283-302.

- Godwin-Jones, B. (2004). Language in action: From webquests to virtual realities. *Language Learning & Technology*, 8(3), 9-14.
- Hamill, L. (2006). Controlling smart devices in the home. *Information Society*, 22, 241-249.
- Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human– chatbot conversations. *Computers in Human Behavior*, 49, 245-250.
- Jia, J. Y. (2004). The study of the application of a web-based chatbot system on the teaching of foreign languages. In *Proceedings of 15th Annual Conference of the Society for Information Technology and Teacher Education*, 1-6 March, 2004, Atlanta, GA.
- Kitade, K. (2000). L2 Learners' discourse and SLA theories in CMC: Collaborative interaction in Internet chat. *Computer Assisted Language Learning*, 13(2), 143-166.
- Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, 70(4), 366-372.
- Long, M. (1996). The role of the linguistic environment in second language acquisition. In W.R. Ritchie & T.J. Bhatia (Eds.), *Handbook of second language acquisition* (pp.413-468). San Diego, CA: Academic Press.
- Long, M. H. (1983a). Linguistic and conversational adjustments to non-native speakers. *Studies in Second Language Acquisition*, 5(2), 177-193.
- Mackey, A. (2012). *Input, interaction, and corrective feedback in L2 learning*. Oxford, UK: Oxford University Press.
- Ortega, L. (1997). Processes and outcomes in networked classroom interaction: Defining the research agenda for L2 computer-assisted classroom discussion. *Language Learning and Technology* 1(1), 82-93.
- Paltridge, B. (2007). *Discourse analysis: An introduction*. London: Continuum International.
- Parker, L. L. (2005). *Language development technologies for young English learners*. Berkeley, California: University of California, Office of the President.
- Pellettieri, J. (2000). Negotiation in cyberspace: The role of chatting in the development of grammatical competence. In M. Warschauer & R. Kern. (Ed.), *Network-based language teaching: Concepts and practice* (pp. 59-86). New York: Cambridge University Press.
- Pica, T. (1994). Research on Negotiation: What does it reveal about second language learning conditions, processes, and outcomes?. *Language learning*, 44(3), 493-527.
- Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition*, 11(1), 63-90.
- Samani, E., Nordin, N., Mukundan, J., & Samad, A. (2015). Patterns of negotiation of meaning in English as Second Language learners' interactions. *Advances in Language and Literary Studies*, 6(1), 16-25.

- Shawar, B. A., & Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics*, 10(4), 489-516.
- Smith, B. (2003). Computer-mediated negotiated interaction: An expanded model. *The Modern Language Journal*, 87(1), 38-57.
- Smith, B. (2004). Computer-mediated negotiated interaction and lexical acquisition. *Studies in Second Language Acquisition*, 26, 365-398.
- Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Ed.), *Input in second language acquisition* (pp. 235- 253). Rowley, MA.: Newbury House.
- Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, 16(3), 371- 391.
- Toyoda, E., & Harrison, R. (2002). Categorization of text chat communication between learners and native speakers of Japanese. *Language Learning & Technology*, 6(1), 82-99.
- Tudini, V. (2003). Using native speakers in chat. *Language Learning and Technology*, 7(3), 141-159.
- Varonis, E. M., & Gass, S. (1985). Non-native/non-native conversations: A model for negotiation of meaning. *Applied Linguistics*, 6(1), 71-90.
- Wang, Y. (2008). *Designing chatbot interfaces for language learning: ethnographic research into affect and users' experiences* (Unpublished doctoral dissertation). University of British Columbia.
- Wang, Y. F., & Petrina, S. (2013). Using learning analytics to understand the design of an intelligent language tutor-Chatbot Lucy. *International Journal of Advanced Computer Science and Applications*, 4(11), 124-131.
- Warschauer, M. (1998). Comparing face-to-face and electronic discussion in the second language classroom. *CALICO Journal*, 13, 7-26.
- Weizenbaum, J. (1966). ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45.
- Williams, L., & van Compernelle, R. A. (2009). The Chatbot as a peer/tool for learners of French. In L. Lomicka, & G. Lord (Ed.). *The next generation: Social networking and online collaboration in foreign language learning*, (pp. 145-172). San Marcos, TX: CALICO.

### **Appendix 1. Pre-questionnaire Questions**

1. 你从什么时候开始学习英语？(When did you start learning English?)
2. 你最近的四级或雅思成绩是多少？(What is your latest CET 4/IELTS score?)
3. 你什么时候开始取得的这个四级/雅思成绩？(When did you get this grade?)

### **Appendix 2. Post-questionnaire Questions**

以下问题基于你与两个不同聊天机器人的交流体验，请如实回答以下问题。

The following questions are based on the experience about your chatting with the two chatbots. Please answer them honestly.

1. 关于不同话题，你觉得和 Tutor Mike 和 Mitsuku 哪个聊天过程更加顺畅（不会出现过多你不能理解的句子表达）？为什么？

In the two conversation processes with Tutor Mike and Mitsuku, which one you think was smoother (i.e. there were not too many sentences you could not understand)? Why?

2. Tutor Mike 和 Mitsuku 两个不同的性别以及形象在多大程度上会影响到你们的聊天？

To what extent the different sex and figure of Tutor Mike and Mitsuku affected your communication?

3. 当交流中出现你不能理解的句子或者表达时，相比于同样情境下的人与人沟通的情景，和机器人聊天是否会让你更尴尬或者更放松？为什么？

Does communicating with chatbots make you feel more embarrassed or relaxed when there are sentences or expressions you cannot understand compared with your online conversations with other humans? Why?

### **Appendix 3. Chat Instructions and Topics**

Talk with the chatbots with any topics you are interested in within 30 minutes. You can choose from the following topics/questions.

Do not use a dictionary. Instead, ask the chatbots when you are confused.

1. Favourite subject(s) in school: What is your favorite subject(s)? Why do you like it or them?
2. Languages: What is the mother tongue of Tutor Mike and Mitsuku? What other language(s) can they speak besides English?
3. Books: What is your favourite book(s)? Who are the main characters in that book? Which character do you like?
4. Hometown: Where is your hometown? Do you like living in a city or countryside? What are the reasons for that preference?
5. Travelling: Which places have you travelled to?