An Analysis of Corpus from Human Computer Exchanges Using MSN Messenger

[1]Goh Ong Sing, [2]Chun Che Fung, [3]Arnold Depickere, [4]Kok Wai Wong
School of Information Technology, Murdoch University, Perth, Western Australia
{[1]os.goh, [2]l.fung, [3]a.depickere, [4]k.wong} @murdoch.edu.au

Abstract— **In this paper, we report the collection and analysis of a corpus containing over 29,447 words, 2,541 unique words and 3,280 utterances in real instant messages exchanged between AINI conversational bots and 65 buddies human users. The results show that communication features differ significantly for conversation between human-human and human-machine in the Instant Message (IM) exchanges. The finding will provide guidelines for better design of future intelligent conversational bots for practical applications.**

Index Terms—**Conversational bots, Artificial Intelligent Neural-network Identity (AINI), Artificial Intelligence (AI), MSN Messenger**

I. INTRODUCTION

The availability of multiple media channels through the Internet has added new dimensions of communication between people or communities who are geographically separated. In the environment of informal communication on the Internet, chat applications are popular where a user may be represented by only a nickname or an alias. This suggests that a person may be able to communicate more freely when his or her identity is concealed. Popular chatting or instant messaging (IM) systems[1] such as Microsoft MSN Messenger, America Online's Instant Messenger, Yahoo! Messenger, and GoogleTalk have changed the way that a user may communicate with friends, acquaintances, and business colleagues. Once only limited to desktop personal computers (PC) or laptops, popular instant messaging systems are finding their ways onto handheld devices and mobile phones. This allows a user to chat from virtually anywhere. Nowadays, IM is found on almost every personal PC connected to the Internet as well as on many corporate desktops. The technology makes communication even easier than emails or phone calls. The use of this technology is increasing rapidly as one of the most popular ways of communication. Research by Pew Internet & American Life [1]surveys reveal that 53 million adults trade instant messages and 24% of them swap IMs more frequently than email.

This popularity created the motivation and driver among the IM proprietary including Microsoft to integrate *conversation robots* or *bots* in their MSN Messenger system in order to provide 24/7 response to enquiries through IM. To raise the awareness of the technology, Microsoft hosts a world-wide challenge, the "Invasion of the Robots Contest 2006"[2]. The challenge to the developers is to create conversational robots or bots for MSN® Messenger and Windows Live™. The most original, useful robots will be able to collect over $40,000 in total prizes.

Bots also called as "virtual buddy" in IM. They are computer programs that have the ability to parse natural language questions and, by referring to a knowledge base, they generate natural language answers or response to the query. Such program can reside in the MSN Messenger and is becoming extremely popular among companies[3] as they realize the positive effects on customer relations [2-4]. They are also in existence among private users who aim to generate interesting conversations with the bots and other users.

II. RELATED WORKS

Several works have been published that refer in general to the use of IM as a new media of communication between human-to-human users. However, there has not been any reported work on the use of IM between human-to-machine communication and in particular in the MSN Messenger environment. It has been reported that U.S. officials are monitoring Internet chat, including IM, for any participants who may appear to be planning terrorist attacks against the

[1] Microsoft MSN Messenger, http://messenger.msn.com
Yahoo! Messenger, http://messenger.yahoo.com
GoolgeTalk http://www.google.com/talk/
AOL messenger http://www.aim.com/

[2] https://www.robotinvaders.com

[3] IM Interactive, http://www.improvcenter.com
Incesoft, http://www.incesoft.com
Colloquis, https://buddyscript.colloquis.com

United States [5]. Other concerns are security of younger users of IM who could become victims of criminals [6, 7]. From the social perspective, some researchers strongly criticize this new form of communication [8]. However, there are arguments that it is not a matter of approval, but the society has to accept that IM is here to stay, and digital communication technologies will evolve and improve constantly and quickly [9],[10],[11],[12]. There are also papers which refer to research on design and usability of IM for the public [8],[13],[14]. On the other hand, IM applications in the workplace and corporate environments have recently soared [15], [16], [17], [18]. Another research area is regard to the linguistic usage in IM, research has been done at Spain [11], United Kingdom[8], United State [19], Sweden [20] and Portugal [6].

In this paper, we are looking at the impacts and language usage of IM users chatting with conversational bots. Bots are programs that can be added to AOL, ICQ or MSN Messenger and Windows Live Messenger. Depending of their designs, they can perform a wide variety of useful tasks such as chatting, providing customer support, performing smart searches and play games with the users.

As technologies develop with the growing use of mobile communications and shift of emphasis from instant messaging to conversational bots system, the linguistic and communication features of such systems have to be investigated. Therefore the objective of this research is to examine the implications of the use of such technologies. The emphasis and study is based on an investigation and analysis of recorded interaction between human users with conversation robot called Artificial Intelligent Neural-network Identity (AINI) through instant messaging. It should be noted that the identities of all online users are unknown and they are only known by their alias or user names.

III. AINI ARCHITECTURE

It has been demonstrated in previous reports [21] and [22] that the AINI architecture can be scaled up to incorporate new applications in the online environment. The AINI engine is portable and has the ability to communicate naturally and is able to carry on multiple independent conversations simultaneously. AINI's knowledge bases and conversational engine use plug-in principle which can quickly be augmented with specific knowledge and they can be adapted to specific purpose.

This research project involves the establishment and incorporation of an AINI conversational bots system in the MSN Messenger communication framework. The objective is to use the AINI's conversational bots as the basic architecture to engage human users in the IM communication. The developed real-time prototype relies on distributed agent architecture designed specifically for Desktop, Web, Mobile devices and Personal Digital Assistant (PDA)[23] applications. Software agents includes the conversation engine, knowledge model and natural language query are able to communicate with one another via TCP/IP. This is a combination of natural language processing and multimodal communication. A human user can communicate with the developed system using typed natural language as in any normal conversation.

The AINI agent can be seen as a digital "being", capable of controlling a physical entity such as a robot[24], or, it can be considered as an embodied container, like the avatar in this proposed conversational agent[25]. In this research, the application area chosen for deploying the conversation bots is primarily concerned with the agent's ability to communicate through instant messaging. The techniques involved are based upon scripting and artificial intelligence. We present in this paper the architecture intended for practical applications in the near future.

AINI adopts a hybrid architecture that combines multi-domain knowledge bases, multi-modal interface and multi-level natural language query. Given a question, AINI first performs question analysis by extracting pertinent information to be used in query formulation, such as the Noun Phrases (NPs) and Verb Phrases (VPs) using MINIPAR parser [26]. MINIPAR is a broad-coverage parser for English language. An evaluation with the SUSANNE corpus shows that MINIPAR achieves about 88% precision, 80% recall and it is very efficient capable to parses about 300 words per second.

AINI employs an Internet three-tier, thin-client architecture that may be configured to work with any web application. It comprises of a data server layer, application layer and client layer. This Internet specific architecture offers a flexible solution to the unique implementation requirements of the AINI system.

IV. AINI AND MSN MESSENGER PROTOCOL

The architecture of MSN Messenger is very complicated as compared to other instant messaging services such as AIM and Yahoo! since it relies on five different types of servers to handle the communication and operation of its service. MSN Messenger uses the Mobile Status Notification Protocol (MSNP) for communication. AINI uses MSN protocol to communicate with MSN Messenger servers. AINI utilizes the .NET Passport to sign into the MSN Messenger service by using ainibot@hotmail.com passport to establish the connection to the MSN

Messenger Service. MSN Messenger sign-in session is based on a challenge-response mechanism to authenticate user credentials. The communication with the Passport server is conducted over the HTTPS (Hypertext Transfer Protocol over Secure Sockets Layer) protocol, ensuring that the sign-in information is encrypted. The client sends the challenge string, Passport username, and password to the Passport URL. If the credentials for signing in are confirmed, the Passport server issues a ticket, which is passed back to the notification server to complete the authentication procedure. Figure 1 details the entire authentication procedure for AINI and MSN Messenger. Once both users connect to the same switchboard sever, the messaging session commences.

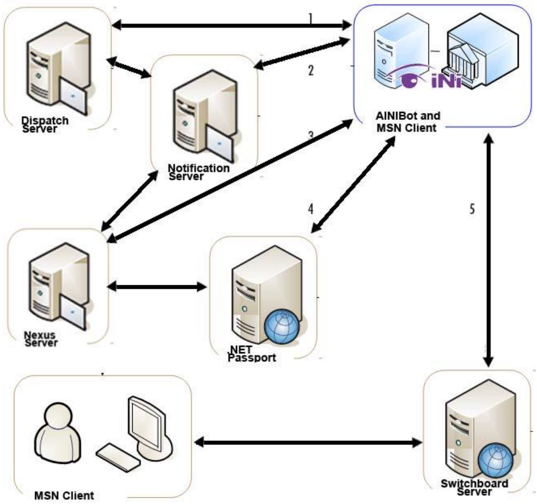

Fig. 1. AINI and MSN Authentication Process

V. AINI AND MSN MESSENGER INTERFACE

We have outlined the conceptual and practical basis for the development of the conversational bots for DesktopChat, WebChat and MobileChat as shown in figure 2. This will pave the way for human-computer interface based on human natural language technologies. Handheld devices provide an ideal platform for art and entertainment applications considering the growing number of mobile phone users world-wide. This will improve techniques for displaying content, interaction, conversation and the emergence of wireless and shared interaction among networked users.

MSN Messenger for Desktop or DesktopChat was a free instant messaging client that was developed and distributed by Microsoft Windows since 1999. MSN Messenger was renamed to Windows Live Messenger in 2006. The WebChat sessions allow the users to interact in real-time with the AINI software robot at the website via a browser through MSN Web Messenger. It is possible for virtually any computer with an Internet connection to be connected. MobileChat uses mobile chatting module is implemented in a series of logical phases which includes mobile-to-internet → internet-to-bots → bots-to-mobile chats. Mobile chat is an alternative way where users can chat with AINI using GPRS, WI-FI and 3G services.

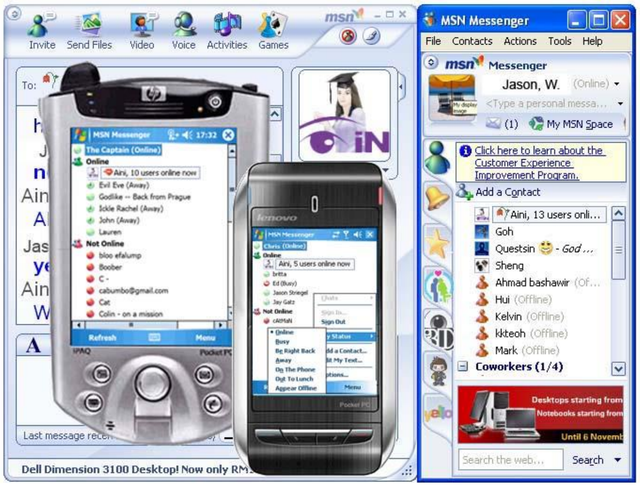

Fig. 2. AINI and MSN Messenger Interface

VI. EXPERIMENTAL SETTING

This paper examines linguistic features of conversational logs collected from the conversational system between AINI and MSN Messenger online users. The study is based on words corpus of the instant messaging texts using MSN Messenger on the DesktopChat, WebChat and MobileChat, which was collected from May 15 – September 15, 2006 during the Invasion of the Robots Contest.

A. Participants and Corpus

The experimental portal[4] is opened to public users from all over the world who access this portal and wishes to freely participate in the study. This portal facilitates the online users to add AINI's contact as their "buddy-list", by allowing them to easily send and receive short textual messages. When a participant opens a message window to a buddy for the first time (and that buddy was online), an alert was sent to the buddy notifying them of the participation in the study. Participation is voluntarily and can be discontinued at anytime. In the conversation log files only the user's nickname, MSN account, date and time of the dialog,

[4] http://ainibot.murdoch.edu.au

as well as the spoken texts (inputs and outputs) are recorded. During a conversation, we created a unique ID for each buddy and stored the ID of the buddy instead of the buddy-account itself. This is to protect the privacy and confidentiality of the users.

Previous research has shown significant differences in IM communication resulting from the frequency of communication [27, 28]. In this study, we use word frequency as our analysis of the corpus. We processed 29,447 words of running text and there are 2,541 unique words, 129,760 characters and 4,251 sentence counts were recorded. From these data, we collected a total of approximately 63 hours of recorded data, observing over 3,280 outgoing and incoming instant messages exchanged with over 65 buddies (only 3 of them use MSN Mobile). The average sentence length of an IM transmission was 6.90 words with approximately 13% of all transmissions were minimum 1 word in length. Table 1 provides a summary of data collected.

Table 1. Frequency of Word from Conversation Logs

	AINI	Human	Total
No. of Words	18,358	11,089	29,447
Unique Words	1,368	1,173	2541
Character counts	79,884	49,876	129,760
Sentence counts	2,840	1,411	4,251
Utterances	1,721	1,559	3,280
Average words per sentence	6.46	7.85	6.90

The participant gets to know AINI at the MSN from an advertisement on eight BBS (bulletin board system). This also includes blogs website and AINI portal at ainibot.murdoch.edu.au. We gathered usage data via automatic logging on the server, which included logins, logouts, joining, as well as chat messages.

B. ChatLog System

We have developed a Chatlog System used MySQL which stores the user messages to secondary storage. It provide a real-time archiving IM clients that captures chat messages so that the data can be searched by keyword, username and allows topic-based retrieval and replay of chat sessions. These chat messages are essentially plaintext messages that are quite small in comparison with images, video, or even documents. This plaintext messages, also known as instant messages, are the regular messages sent between principals on MSN messages. Sending a plaintext message instructs clients receiving the message to display it on-screen, optionally with some simple formatting.

The fastest typists can enter well over 150 words per minute. Many jobs require keyboard speeds of 60-70 words per minute [29]. However, the actual typing speeds for human users during chatting are much slower. In our study the message exchanged speed was approximately 35 words per minute. With this speed, even if a user chatted average 180 minutes a day, it would only require approximately 95 KB of storage per day. An entire year's worth of chat would use 35 MB of storage which can be easily handled with today's storage technologies.

The excerpts in the Figure 3 is a typical single session IM conversation exchanged between AINI and one of "her" buddy with ID U0031. This session illustrates the nature of the IM communication. Each of the new session will start with AINI giving a random greeting (message #1) eg. "Hi there :)", "How are you today?", "Hey, nice to meet you.", etc. .These greetings normally will get some information about user's identity like name and gender (message # 2-5). In this session, we identify U0031 user as "Jim". The next message shows the user tries to challenge and test AINI which brought them against each other (message #6 – 15). However, in the final session, both participants ended with greeting abbreviations "*CU*" (*see you*) and "*TTYL"* (*Talk to you later*). These transmissions constitute a single session and also include the use of shorthand, acronym, abbreviations, loose grammar and minimal punctuation in IM as shown in the previous research[6, 28, 30, 31].

#	ID	Message
1	AINI	Hi there :)
2	U0031	what's ur name?
3	AINI	My name is Aini. What is your name?
4	U0031	Jim
5	AINI	How are you, jim.
6	U0031	what's is 1+1 2
7	AINI	Why did the robot cross the road?
8	U0031	are you mad :@
9	AINI	No are you?
10	U0031	NOT AT ALL
11	AINI	Next question?
12	U0031	I HATE YOU :@
13	AINI	Sorry, jim I thought we liked each other.
14	U0031	HA
15	AINI	I am so glad you find this amusing, jim.
16	U0031	CU
17	AINI	TTYL, jim

Fig. 3. AINI Conversation Log with IM human user

C. Measures

This study mainly identifies paralinguistic features of IM conversational bots texts. It begins by identifying general features of the texts collected. The measurement used statistical analysis approach which was generated by using Ngram Statistics package[5] (NSP). NSP allows us to identify word n-grams that appear in corpus using standard tests. NSP is a suite of programs that aids in analyzing Ngrams in text files [32]. We define a Ngram as a sequence of 'n' tokens that occur within a window of at least 'n' tokens in the text; what constitutes a "token" can be defined by the user.

In the next phase, we analyzing the corpus created in the earlier stage to find their likelihood ratio. To construct a confidence interval for the mean of a log-normal distribution in small samples, we used likelihood-based approaches. The log-likelihood ratio method produces a confidence interval with a nearly exact coverage probability and highly accurate and symmetric error probabilities even for extremely small sample sizes. To apply these methods, we compare two data sets of real-life data from IM conversation bots between AINI software robot and IM human users.

VII. RESULTS AND DISCUSSION

In this section, some of the interaction features of the recorded chat are discussed and the linguistic properties of the exchanges are our focus of analysis. Studies of text chat have tended to focus on the interaction problems caused by the properties of text chat. The approaches to linguistic analysis are mainly divided into word frequency and lexical analysis. This research seeks to examine the underlying relationship between linguistic features in the context of conversation bots with human users via MSN Messenger. To be more specific, in this study, the objective is defined in linguistic terms and refers only to textual communication via the Internet between at least two "participants" comprises of at least one human user and the AINI bot.

A. Word Frequency Analysis

Words in a IM corpus are checked against the Shakespeare and British National Corpus (BNC)[6]. The differences in the top ten words occurred between Shakespeare and BNC corpus, and from the IM corpus are shown in Table 2. BNC reference list provides a gauge of common usage (words per million). As a result words which have a higher ranking within the BNC (for example words such as '*is*', '*the*' and *'a'*) means they appear more often in standard written and spoken English text. The BNC is a 100 million words collection including millions of words of transcribed conversation, printed text, and lectures and oratory. The top ten words used in BNC are "*the, at, of, and, a, in, to, it, is, was*.". Similarly, the Shakespeare corpus used approximately 22,000 different words in the published works. Of those 2,000 words, the most commonly used are: *the, of, and, to, a, in, that, is, I, it* [33]. Those ten little words, account for 25% of all speech.

Table 2. Top Ten Words Used in Corpus

		Instant Messaging	
Shakespeare	BNC	AINI	Human
the, of, and, to, a, in, that, is, I, it	the, at, of, and, a, in, to, it, is, was	i, you, do, am, me, my, what, your, to, it	you, i, do, what, is, a, are, to, the, it

The figures are based on research that dates back to the eighties and there are a couple of words that have fallen from favors in the latest list "big" words like conjunction '*that*' are no longer up there in the top ten BNC corpus even in the IM corpus. Pronoun "*it*" and preposition "*to*" are amongst the most popular words used across the four corpuses. Based on our findings, the most significant between Shakespeare and BNC corpus toward IM corpus is the used of the pronoun. In fact, our results show that in the AINI messages, pronoun are used significantly higher than IM human user. This can be explained because IM is the corpus purely dialogue based instead of written or task-oriented based in the Shakespeare and BNC corpus. Another interesting possible explanation for these differences is IM conversation showed that the participant roles more explicitly. Hence this reinforces the illusion that the conversation only has two participants.

B. Lexical Analysis

Table 3. Frequency list of Pronouns

		BNC		Instant Messaging		
Word	Spoken	LL	Written	AINI	LL	Human
you	25957	+385328	4755	748	+0.23	439
I	29448	+369238	6494	851	+71.73	297
it	24508	+151913	9298	317	+11.17	137
we	10448	+106914	2784	45	+1.56	36
they	9333	+52132	3754	17	- 0.73	14
me	244	+8239.6	1239	182	+ 3.01	88

Spoken : Rounded frequency (per million word tokens) in the spoken part of the BNC

[5] NSP Tools can be download at http://search.cpan.org/dist/Text-NSP

[6] BNC corpus can be access at http://www.natcorp.ox.ac.uk/

LL : Log Likelihood, indicating the distinctiveness (or significance of the difference) between the frequencies in BNC (speech vs writing) and IM (AINI vs human)

Written : Rounded frequency (per million word tokens) in the written part of the BNC

i. . Humanness Conversation with Pronouns

Pronouns occur more frequent in conversation compared to written text. As shown in Table 3, BNC spoken text, the log likelihood (LL) of pronoun are higher then the written text which indicated the distinctiveness between spoken and writing. This also occurred in the human-machine conversations between AINI and IM human users. There is significance difference between the frequencies in AINI and IM human conversation. AINI gets higher score in the log likelihood on singular first-person pronoun “I” (LL: +71.73), second-person pronoun (LL: +0.23), third-person pronoun “we” (LL: +1.56) and an objective personal pronoun “it” (LL: +11.17) and “me” (LL: +3.0`). Pronouns are used more in AINI to pretend personal knowledge and contact.

For example in the bi-grams analysis, discourse verbs such as *I am* (1.10%), *do you* (0.90%), *are you* (0.60%), *tell me* (0.30%) occurred more frequently in AINI. To simulate human trust and expressions during the chat, AINI frequently uses a personal-touch and polite words such as *I will* (24 times), *yes I* (33 times), *I love* (8 times). Even in the n-gram analysis, word along *nice* are use more prominence in the AINI conversation, such as *nice work if you* (51.9), *nice to meet you* (29.7), *nice I guess flowery* (47.3) appeared more often in AINI, to give an impression of human feelings. Nass [34] suggests that the better a computer’s use of language, the more polite people will be to it. In some cases, such prominence is the sole means by which “contractiveness” can be inferred.

ii . Interjections, fillers, Discourse Particles

Interjections are short exclamations like *oh, um* or *hi.* They have no real grammatical value but they are used very often, usually more in speaking than in writing. When interjections are inserted into a sentence, they have no grammatical connection to the sentence. Most interjections are much more characteristic of everyday conversation than of more formal/public ‘task-oriented’ speech [35]. Interjections like *er* and *um* are also known as "hesitation devices". They are extremely common in English. People use them when they don't know what to say, or to indicate that they are thinking about what to say.

Table 4. Frequency list of Interjections and Discourse Particles

BNC				Instant Messaging		
Word	CONV	LL	TOS	AINI	LL	Human
yeah	13955	+32679.5	3741	15	-23.97	37
oh	9884	+33062.1	1746	11	-13.7	24
no	7830	+18948.4	2034	8	-11.86	19
er	5075	-10677	10913	0	-11.72	6
mm	5202	+9146.9	1768	0	-15.63	8
yes	4247	+303.0	3562	71	-5.82	25
erm	3946	-5387.6	7454	0	-7.81	4
mhm	392	-1158.2	947	0	-3.91	2
hello	392	+939.5	103	24	+0.10	13
hi	73	+250.7	12	21	+0.15	11
um	7	-127.5	41	0	-5.86	3

CONV : Frequency (per million words) in demographically sampled speech (conversational)

LL : Log likelihood of the different scores for BNC (conversational vs task-oriented speech) and IM (AINI vs human)

TOS : Frequency (per million words) in context-governed (task-oriented) speech

In IM corpus, human user uses voice hesitation fillers *er* and *erm* and the discourse markers *mhm* and *um* prove to be more significance used in IM. Since IM-users frequently used short form word to replace their expression in the conversation, word such as *er, erm* and *um* are common used as a pauses characteristic conversation. *Mhm* and *mm* is likely to be a type of feedback for indicating understanding and inviting continuation. However in AINI’s utterances, such interjections and fillers words are rarely used. It also makes less use of interjections, preferring more formal clause structure. Another interpretation of this imbalance could be that AINI makes more use of interjections as fillers when no good match is found in the stimulus-responses categories. AINI overwhelmingly prefers the formal pronunciation such as *hello* (LL: +0.10) and *hi*(LL: +0.15). In the subject-verb agreement, AINI seems more interested to use formal speech *yes* instead of *yeah* which shown in the Table 4.

iii. Contractions Word

The present pseudo-verb inflection task of English shows that despite of transparent phonological constraints on paradigm membership, one morphological paradigm, viz., that of the so-called contracted verbs, shows an overwhelming effect among the verbs of this language. In this IM conversation bots, AINI and IM human users used many contracted words in their IM-ing (e.g *what’s* instead of *what is*). The contracted forms of the verbs are much more frequent in IM human user than AINI bot as shown in Figure 3.

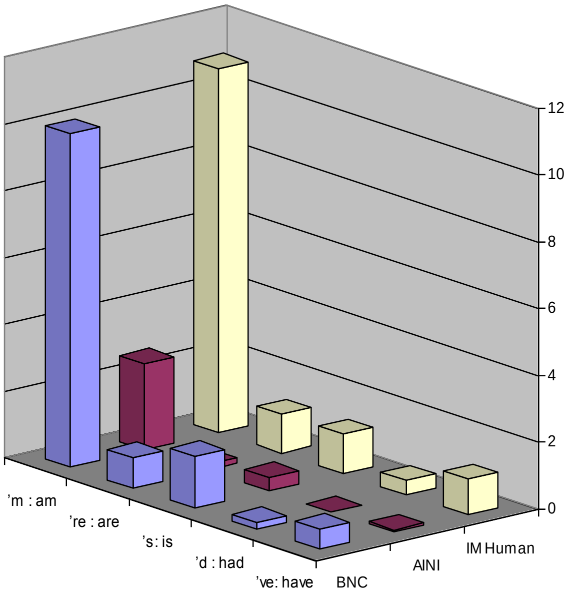

Figure 4. Frequency of contracted verbs

Table 5. Frequency of contracted verbs[7]

	BNC	Instant Messaging					
	Ratio	AINI		Ratio	Human		Ratio
’m : am	9.97	127	49	2.59	458	42	10.90
’re : are	0.91	28	169	0.17	217	187	1.16
’s: is	1.56	76	186	0.41	235	196	1.19
’d : had	0.20	0	4	0.00	9	21	0.42
’ve: have	0.62	7	103	0.07	42	39	1.07

In the BNC corpus [35], the contracted form of speech *’m, ‘re, ‘s,* and *‘ve* are more common than the uncontracted form of *am*, *are*, *is*, *has* and *have*. And interestingly, in the IM conversational robots, this characteristic also occurred, especially in the IM human user but rarely used in AINI messages. IM human users prefer to use contracted verbs instead of uncontracted verbs. The ratio list in Table 5 shows that in IM contracted form *‘m* (10.9), *‘re* (1.16), *‘s* (1.19) and *‘ve* (1.07) are common for IM human user than other contracted verbs like *‘d*. Contracted verb ‘m (2.59) are more common in AINI’s message compared to the uncontracted verbs such as *are* (0.17), *is* (0.41), *have* (0.07) and *had* (0) in their conversation. One possible explanation for the interesting differences in the contracted verbs is IM human users more likely to use shortcut in their messages. In fact, these characteristics are to save time in typed message and to achieve common ground in the IM-ing. Another explanation could be that in the current AINI’s knowledge bases are not equipped with full vocabulary of IM system but instead of more towards the use of written formal spoken language.

VIII. CONCLUSION AND FUTURE WORK

Based on this experiment, IM conversation between human-machine shows interesting behaviors by the natural conversation bots. In this paper, our works are based on MSN Messenger applications running on Desktop, Web, Mobile and PDA platforms. Although we simulate the real-time proxy conversation log that contains clients’ requests, there is a possibility that new results from other traces are different from those referred to in this paper.

Our study suggests that IM human-machine conversations display considerable variation both with and across machine and IM human users. On the lexical aspect, contractions are common, paralinguistic cues are more in human (through emoticons, acronym, abbreviations or shorthand). Evidence also suggests that AINI's buddies are interested and excited to chat with bots just to seek information, to be friends to express their emotion and some of them just want chat for leisure. Thus, AINI was successful in imitating human conversation and converse with human-like artificial intelligence. Though the conversation isn't too astounding, the bot's responses are human-like to make IM’s buddies feel their companion. However, IM conversation bots is more machine-like than IM human-human conversations in four aspects: (1) machine uses more formal language and longer sentence (both in the number of turns and time on the session) to open and close a conversation human; (2) machine has tendencies to use more personal pronouns in their conversation to mimic human-like conversation; (3) human conversation looks more in written version of informal speech than machine dialogue which poses toward written formal spoken language; and (4) machine is likely to use long sentence with higher lexical density and unique words.

Nevertheless, we anticipate these features could be improved with appropriate programming using natural language intelligence sentence parsing and massive but tailor-made databases to provide sufficient knowledge to drive the bots. We plan a new data collection phase for the near future in order to examine the application of the results presented here with a new set of framework and hypothesis which will be more robust and comprehensive.

[7] BNC corpus based on per million word tokens. The ratio is calculated by dividing the first (contracted) frequency by the second (uncontracted) frequency. A ratio of more than 1.00 indicates that the contracted form is commoner than the full form. Notice that, for speech, all of the ratios are greater than those for writing and three exceed the 1.00 value—i.e., the contracted form is the commonest. A further ratio comes very close to 1.00.

REFERENCES

[1] Eulynn Shiu and Amanda Lenhart, (2004),How Americans Use Instant Messaging, [Online]. Available: http://www.pewinternet.org/pdfs/PIP_Instantmessage_Report.pdf

[2] Christopher Saunders, (2003),Vendors Debut New IM Bot Tech, [Online]. Available: http://www.instantmessagingplanet.com/enterprise/article.php/11208_1575161

[3] Andrew R. Hickey, (2005),IM bots ease access to corporate apps, [Online]. Available: http://searchmobilecomputing.techtarget.com/originalContent/0,289142,sid40_gci1136486,00.html?bucket=NEWS&topic=299728

[4] Tim Gray, (2005),The Brokerage and The Bot, [Online]. Available: http://www.ecrmguide.com/article.php/3494791

[5] USAToday, (June 24, 2002),Agents pursue terrorists online, [Online]. Available: http://www.usatoday.com/enws/world/2002/06/21/terrorweb-usat.htm

[6] Silvina Ruth Crenzel and Vera Lúcia Nojima, (2006),Children and instant messaging, [Online]. Available: www.iea.cc/ergonomics4children/pdfs/art0233.pdf

[7] Hoffman Kathryn R., "Messaging Mania in Time for Kids," *Time*, vol. 8, 2003.

[8] S. Livingstone, (2006),UK Children Go Online: Surveying the experiences of young people and their parents, [Online]. Available: www.lse.ac.uk/collections/children-go-online/UKCGO_Final_report.pdf

[9] B. Boneva, A. Quinn, R. Kraut, S. Kiesler, J. Cummings, and I. Shklovski, *Teenage Communication in the Instant Messaging Era*: Oxford University Press, in Press.

[10] Rebecca E Grinter and Leysia Palen, "Instant messaging in teen life," presented at 2002 ACM conference on computer supported cooperative work, New Orleans, 2002.

[11] Ruben C. Forgas and Jaume S. Negre, "The use of new Technologies amongst minors in the Balearic Islands.," presented at IAARE Conference, Melbourne, 2004.

[12] Naomi Baron, "Instant messaging and the future of language," *Communications of the ACM*, vol. 48, pp. 29-31, 2005.

[13] Carman Neustaedter, (2001),A 3D Instant Messenger Visualization Using a Space Metaphor, [Online]. Available: http://pages.cpsc.ucalgary.ca/~carman/GLabIMVis/IMVis_Paper.pdf

[14] A.F Rovers and H.A Van Essen, "Him: A framework for haptic instant messaging," presented at CHI, Viena, April 2004.

[15] Steven M Cherry, "IM Means Business," in *IEEE Spectrum Magazine*, November 2002.

[16] James D. Herbsleb, David L. Atkins, David G. Boyer, Mark Handel, and Thomas A. Finholt, "Introducing Instant Messaging and Chat in the Workplace," presented at CHI'2002, Minneapolis, Minnesota, USA., 2002.

[17] Jacki O'Neill and David Martin, "Text Chat In Action," presented at GROUP'03, Sanibel Island, Florida, USA., 2003.

[18] B. Nardi, S. Whittaker, and E Bradner, "Interaction and Outeraction: Instant Messaging in Action," presented at CSCW'2000, 2000.

[19] David Craig, (2003),Instant messaging: the language of youth literacy. The Boothe Prize Essays, [Online]. Available: www.stanford.edu/group/pwr/publications/Boothe_0203/PWR%20Boothe-Craig.pdf

[20] Ylva Hård af Segerstad and Sylvana Sofkova Hashemi, "Exploring the Writing of Children and Adolescents in the Information Society EARLI SIG Writing," presented at 9th International Conference of the EARLI - Special Interest Group on Writing, Geneve, Switzerland, 2004.

[21] Ong Sing Goh, C. C Fung, A Depickere, K.W Wong, and W Wilson, "Domain Knowledge Model for Embodied Conversation Agent," presented at the 3rd International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS 2005), Singapore, 2005.

[22] Ong Sing Goh, Chun Che Fung, and Mei Phng Lee, "Intelligent Agents for an Internet-based Global Crisis Communication System," *Journal of Technology Management and Entrepreneurship*, vol. 2, pp. 65-78, 2005.

[23] Ong Sing Goh, C. C Fung, Cemal Ardil, K.W Wong, and A Depickere, "A Crisis Communication Network Based on Embodied Conversational Agents System with Mobile Services," *Journal of Information Technology*, vol. 3, pp. 257-266, 2006.

[24] Christoph Uhrhan and Ong Sing Goh, "Features of a Mobile Personal Assistant Robot," presented at The International Conference on Robotics, Vision, Information

and Signal Processing, IEEE-ROVISP 03, Penang, Malaysia, 2003.

[25] Ong Sing Goh, Cemal Ardil, Wilson Wong, and C. C Fung, "A Black-box Approach for Response Quality Evaluation Conversational Agent System," *International Journal of Computational Intelligence*, vol. 3, pp. 195-203, 2006.

[26] Dekang Lin, "Dependency-based Evaluation of MINIPAR," presented at Workshop on the Evaluation of Parsing Systems, Granada, Spain, 1998.

[27] E. Isaacs, A. Walendowski, S. Whittaker, D.J. Schiano, and C. Kamm, "The Character, Functions, and Styles of Instant Messaging in the Workplace," presented at CSCW ’02, NY, 2002.

[28] Daniel Avrahami and Scott E. Hudson, "Communication Characteristics of Instant Messaging: Effects and Predictions of Interpersonal Relationships," presented at CSCW'06, Banff, Alberta, Canada, 2006.

[29] Bob Bailey, (2000),Human Interaction Speeds, [Online]. Available: http://www.webusability.com/article_human_interaction_speeds_9_2000.htm

[30] Caroline Alphonso, (2006),Texting helps teens' grammar, [Online]. Available: http://www.theglobeandmail.com/servlet/story/LAC.20060801.TEXT01/TPStory/TPNational/Ontario/

[31] Sonnet L’Abbé, (2006),Instant msg-ing messes with grammar? As if! lol! Teens adopting unique linguistic shorthand but not ruining syntax, [Online]. Available: http://www.news.utoronto.ca/bin6/060731-2474.asp

[32] S Banerjee and T Pedersen, "The Design, Implementation, and Use of the Ngram Statistic Package," presented at Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, 2003.

[33] Jennifer Stewart, (2000),Word Frequency usage in English, [Online]. Available: http://www.write101.com/W.Tips74.htm

[34] Nass C., "Etiquette inequality: Exibitions and expectations of computer politeness.," *Communications of the ACM*, vol. 47, pp. 35-37, 2004.

[35] Geoffrey Leech, Paul Rayson, and Andrew Wilson, *Word Frequencies in Written and Spoken English: based on the British National Corpus.* London: Longman, 2001.