

I hate you! Disinhibition with virtual partners

Antonella De Angeli^{a,*}, Sheryl Brahnam^{b,1}

^a Manchester Business School, The University of Manchester, Booth Street West, Manchester M15 6PB, UK

^b College of Business Administration, Computer Information Systems, Missouri State University, 901 S. National, Springfield, MO 65804, USA

Received 19 January 2008; accepted 4 February 2008

Available online 13 February 2008

Abstract

This paper presents a descriptive lexical analysis of spontaneous conversations between users and the 2005 Loebner prize winning chatterbot, Jabberwacky. The study was motivated in part by the suspicion that evidence in support of the Media Equation, especially in the field of conversational agents, was supported by incomplete data; too often omitted in its purview is the occurrence of unsavoury user responses. Our study shows that conversations with Jabberwacky often bring about the expression of negative verbal disinhibition. We discovered that 10% of the total stems in the corpus reflected abusive language, and approximately 11% of the sample addressed hard-core sex. Users were often rude and violated the conversation maxims of manner, quantity, and relevance. Also particularly pronounced in the conversations was a persistent need of the user to define the speakers' identities (human vs. machine). Users were also curious to understand and test the cognitive capabilities of the chatterbot. Our analysis indicates that the Media Equation may need qualifying, that users treat computers that talk, less as they do people and more as they might treat something not quite an object yet not quite human.

© 2008 Elsevier B.V. All rights reserved.

Keywords: Chatterbots; Disinhibition; Verbal abuse; Sex-talk; Media equation; Social agents

1. Introduction

The prevailing approach in designing and evaluating conversational agents draws heavily on the Media Equation, or Computers Are Social Actors (CASA) paradigm, proposed to account for social attributions in user interactions with computers (Reeves and Nass, 1996). A major claim of the Media Equation is that individuals' interactions with computers are fundamentally social in nature and correspond to the ways people naturally interact with other people. Since a major goal in the design of conversational agents is the provision of a more natural, human-like interface, the CASA paradigm is appealing. If it is true that

people unwittingly treat computers as human beings, then CASA extends a promise of eventual success.

There is ample evidence in support of CASA. Reeves and Nass (1996), for instance, have taken a number of social science studies, replaced one or more human beings with computers in the method of the study, and then performed the same experiments to determine whether the social principles still hold. In case after case, their subjects, many of whom were sophisticated computer users, applied exactly the same social rules of conduct to computers as they did to people. Furthermore, when the computer interface was endowed with specific human characteristics (language, the assumption of human roles, a voice) the CASA paradigm seemed particularly pronounced.

Apparent in these and other CASA studies, however, is a tendency to explore only one end of the emotional spectrum: the positive. Studies applying the CASA paradigm seldom investigate negative user emotional responses and behaviour. Even in cases purportedly examining negativity, the focus is more on brain activity and self-report than on

* Corresponding author. Tel.: +44 (0) 161 3063383; fax: +44 (0) 161 3063518.

E-mail addresses: Antonella.de-angeli@manchester.ac.uk (A. De Angeli), sbrahnam@missouristate.edu (S. Brahnam).

¹ Tel.: +1 417 873 9979; fax: +1 417 836 6907.

observing user behaviours. Yet it seems that users have the potential of responding more negatively. There are studies, for example, that show people negatively stereotyping computers once they are endowed with a voice (Reeves and Nass, 1996) or a face (Sproull et al., 1996).

This one-sidedness is also evident in the evaluation of conversational agents. Consistent with the major goal of making conversational interfaces that are useful as well as more natural, most studies address positive outcomes of virtual conversations related to social facilitation effects. Typical dependent variables in evaluation studies of conversational agents are measures of performance enhancement and indicators of motivational and social facilitation, such as attitudes, trust, politeness, self disclosure, and persuasion. With a few exceptions (De Angeli et al., 2005, 2006), research has ignored potential negative outcomes of the interaction, despite the fact social agency subsumes the occurrence of conflicts. Even in cases where purely descriptive research is conducted, negative user responses to virtual partners are normally ignored. Many researchers consider these types of exchanges unwelcome noise, even a nuisance, and they are usually deleted from the interaction logs without much thought or else examined with an eye mostly directed towards building deflective strategies. On the rare occasion when this phenomenon is acknowledged in the research literature, references are parenthetical asides or embarrassing warnings to the reader that excerpts from some of the example interaction logs contain strong and unpleasant language (see, for example, Fischer, 2006).

The study presented in this paper was motivated in part by the suspicion that the evidence in support of CASA, especially in the field of conversational agents, was inconclusive and supported by incomplete data – that there existed a troubling lacuna in current research, a gap that omitted from its purview disinhibited and unsavoury user responses. This paper contributes to addressing this omission by analysing and discussing verbal disinhibition in conversations with an online chatterbot. Anecdotal evidence suggesting that 20–30% of user interactions with conversational agents would be considered abusive if directed towards another person prompted a number of questions: what are the major themes in conversations with artificial partners, how much of the language is abusive, do users obey normal conversational norms when talking with machines, what do these conversations tell us about how the users characterize virtual conversational partners, is a generalized stereotype of the artificial partner evinced in user responses?

In an attempt to answer some of these questions, we performed a descriptive lexical analysis of user conversations with the online chatterbot, Jabberwacky (Carpenter, 2007). A chatterbot is a popular type of conversational agent. In Section 2, we provide background information on chatterbots and briefly highlight previous research on conversational interfaces. In Section 3, we present the method used in the analysis of a set of conversations

($N = 145$) recorded by Jabberwacky. In Section 4, we report results and show that conversations with Jabberwacky often bring about the expression of negative emotions in users, as manifested by instances of verbal disinhibition. Finally, in Section 5, we conclude by noting the relevance of our analysis and some of the implications of our results. We also offer in this section a few directions for further study.

2. Conversational agents and chatterbots

The umbrella term *conversational agent* describes software that communicates with the user using natural language. Communication can be mediated via textual input and output or through voice recognition and synthesis. Conversational agents are a form of weak AI in that they aim at simulating intelligence and knowledge without actually possessing intelligence or even a large database of factual information. Some of the least intelligent conversational agents, often referred to as *chatterbots* or *chat bots*, short for *chat robots*, imitate conversation using a set of dialogue management rules, which range in sophistication from textual or audio pattern matching to a slightly more complex set of rules based on inference mechanisms (L'abbate et al., 2005).

The term *chatterbot* originates with Julia (Mauldin, 1994), one of the first robotic participants in a MUD and the object of affectionate attention from its users and several sociological studies (Foner, 1997; Turkle, 1995; Zdenek, 2003). The chatterbot's basic mechanism for generating automatic conversation, however, is much older, dating back to Weizenbaum (1966). Eliza presents itself as a Rogerian (non-directional) psychotherapist. Within the confines of a psychotherapeutic context, which entails mirroring patient responses, it is able to converse by turning the user's sentences around. Eliza identifies a set of keywords occurring within the user's input and tries to define a minimal context in which those keywords appear. For instance, the words "I" followed by "you" are interpreted as an assertion. Eliza then chooses an appropriate transformation rule to modify the user's input, as illustrated in the following exchange taken from Weizenbaum (1966), pp. 36–45:

Patient: Men are all alike
 Eliza: In what way.
 Patient: They're always bugging us about some-thing or other.
 Eliza: Can you think of a specific example.
 Patient: Well, my boyfriend made me come here.
 Eliza: Your boyfriend made you come here.
 Patient: He says I'm depressed much of the time.
 Eliza: I am sorry to hear you are depressed.

A few years after Eliza came Parry (Colby et al., 1971). It was designed to act like a paranoid patient in a hospital

and managed to fool several psychiatrists in a Turing Test (Turkle, 1995). Some users refused to believe Parry was a computer, even after they were presented with the fact they were conversing with a machine. Parry's strategy is somewhat the reverse of Eliza's. Parry directs the conversation by stating his beliefs, fears, and anxieties within a network of associations typical of a paranoid mind. The following illustrates Parry's paranoid style (Colby et al., 1971, pp. 1–25):

Interviewer: Why are you in the hospital?
 Parry: I shouldn't be here.
 Interviewer: Who brought you here?
 Parry: The police.
 Interviewer: What trouble did you have with the police?
 Parry: Cops don't do their jobs.
 Interviewer: What do they do instead?
 Parry: Cops arrest the wrong people

The success of Eliza and Parry relies on the choice of a convenient communication setting. Psychotherapists and paranoids act in a well-defined and constrained communication context, where deviations from established conversational rules are not only expected but also tolerated. Psychotherapists are trained and expected to mirror and elaborate user narratives; patient concerns are probed with personal questions that are perceived by the patient as having a therapeutic purpose. Similarly, inability to collaborate, concerns that other people have hidden motives, and hostility towards the world, are well-known symptoms of paranoia and recognised disruptors of communication.

In recent years, there has been an explosion of interest in chatterbots that has been fuelled by the Internet, which makes a wide variety of contemporary chatterbots available to a large number of curious users. The "chatterbot collection" website (Chatterbot Website, Last accessed 2007), for example, lists close to a thousand, many produced by hobbyists and no longer active. Curious *home-made* examples are George, the virtual alter-ego, created by an Australian professor to deal with students' emails, and InnerPeace, a form of conversational self-help software. A variety of renowned personas are also available for conversation. They include God, designed to facilitate repenting, William Shakespeare, John Lennon, and mylittletony, a centaur with the head and torso of Tony Blair. Other chatterbots feature unique artificial identities, such as Alice, which proudly declares itself to be a real robot.

In addition to providing entertainment, conversational agents are currently employed in e-business to attract customers, provide and collect information, and sell goods (De Angeli et al., 2001a). Active examples are RITA, the assistant at ABN AMRO CashPro, Kate the representative of Ford, and Phyllis, the spokeswoman of the Defence Logistics Information Service (DLIS) in the USA. These embodied chatterbots are young, pretty, and oftentimes photo realistically female (the favoured gender for virtual

representatives) (Brahnam, 2005). Little evidence is available on the success of these virtual representatives, but Forrester's research suggests that their average lifespan is restricted to 6 months.

Despite their technological limitations, chatterbots provide researchers with a major tool for studying how people spontaneously interact with conversational technology. The wide availability of chatterbots on the Internet allows the study of *in situ* interaction as opposed to controlled experimental settings normally used to test more technologically advanced embodied conversational agents.

3. Method

The study presented in this paper analyses *in situ* conversations between users and Jabberwacky, a chatterbot designed for entertainment by Rollo Carpenter (2007). The study was designed to answer the following questions: What do users and Jabberwacky talk about? What form do these virtual conversations take? What does the language used by the users tell us about their perception of Jabberwacky? What social norms regulate virtual conversations?

In this section, we present the study methodology. After providing background information on Jabberwacky, we describe the corpus and method of analysis.

3.1. Jabberwacky

Jabberwacky went online in 1997 and has collected a large and active community of users. Unlike most Chatterbots, Jabberwacky's responses are not hard-coded; it learns by association, storing replies to inputs in a database. Jabberwacky started out with 20,000 entries. By 2003 it had reached a million entries. Currently, Jabberwacky has access to over 10 million replies.

Jabberwacky makes conversation by reflecting user input learned in similar contexts with previous interlocutors. Aspects of the current conversation are compared to previous conversations, and responses are selected based on an interpretation of that context. Jabberwacky is incapable of composing a response, since it has no knowledge of the grammatical rules of any language and does not understand the meaning of words. Despite this limitation, Jabberwacky is capable of multilingual responses because people converse with Jabberwacky using many languages. Jabberwacky's replies are a form of parroting. Its responses depend upon the data that have been learnt to date, making it essentially a mirror of its audience. A consequence of this is Jabberwacky's frequent claims to being human and his insistence that it is the user, not himself, who is the robot. Jabberwacky also often abruptly changes the topic. This is partly due to the fact that many users, when talking with Jabberwacky, violate the conversation maxim of relevance (Grice, 1975). Generally, though, Jabberwacky is well behaved, as the majority of bad manners, obscenities, and bad language are continuously filtered out of the database.

Because Jabberwacky is capable of learning from user responses, it scores quite high in the chatterbot arena. It won the Loebner prize in 2005, and Joan, a female chatterbot based on the same architecture, won the Loebner prize the following year.

3.2. Corpus

The study is based on the analysis of the conversations collected from the Jabberwacky webpage on Monday, the 22nd of November, 2004. The advantages of sampling by time are discussed in Herring (2004) and include context richness. The interaction log for that day reported 716 accesses to the dialogue page. For each access, the log registered a unique user identifier, time of day, client IP, and the number of conversational turns. The conversation itself was recorded in a text file named after the unique identifier. A screening of the interaction log, based on IP address comparisons and hit frequencies, led to the deletion of 200 entries, as they had no associated conversations.

A selection of 146 conversations generated by different IP addresses was extracted from the remaining sample. This selection included all conversations ($N = 103$) with visitors that returned 20 or more times, along with a random selection of visitors who returned less frequently. It is important to note that this procedure does not guarantee that we have analysed a sample of 146 different users, as the only information we have about a visitor is the IP address. Yet the number of conversations is sufficiently large to guarantee a reliable sample of speakers.

The average conversation was composed of 82 conversational turns, featuring a strict turn-taking pattern whereby each partner spoke only after the other. Ten conversations were conducted primarily in a foreign language, including French, Danish, and Turkish. In violation of the conversation maxims of manner, quantity, and relevance (Grice, 1975), all these multilingual conversations started in English and suddenly shifted as the users ($N = 8$) or Jabberwacky ($N = 2$) began speaking in a different language. Only in one case did the user ask the chatterbot whether it spoke the language of choice.

3.3. Method of analysis

Since the purpose of this study is to discover the most significant themes in user interactions with a chatterbot and understand how users characterize their virtual conversational partners, we primarily employed a descriptive lexical analysis of the corpus. A lexical analysis allows us to codify themes using several nominal variables. These variables are discovered in the process of examining the ensemble of words that comprise the corpus. The basic functions employed in this examination included concordance analysis (viz., stem and word frequency analysis) and an analysis of conceptual-semantic and lexical relations using WordNet (Fellbaum, 1998). WordNet is a lexical database that organizes English words into synonym sets, each represent-

ing one underlying lexical concept. These are then associated with multiple categories.

In line with our goals, this paper concentrates on the analysis of user language, looking at 12,053 conversational turns composed of some 5 words each, for a total of 3,037 unique words. The term *word* is used here in its broadest sense to include not only number, gender, and orthographic variations but also misspellings, sounds (e.g., *AAAAARRRRRRRRGGGGGHHHHHHH*) and rare, which sometimes were produced to test Jabberwacky's verbal skills (e.g., *cthulhu* God in some Lovecraft books).

Prior to the lexical analysis, the corpus was normalized and conversational abbreviations were substituted with correct grammatical forms (e.g., *isn't* became *is not*).

4. Results

A preliminary qualitative analysis of the conversations indicated that style and topics were diverse. These differences were attributed to the users' personalities, their motivations and attitudes towards Jabberwacky, to the flow of the conversation, and to the appropriateness of Jabberwacky's responses. Some users were friendly and supportive, others were aggressive and abusive. Some users enjoyed probing or testing the system; others played along with it, even inventing connections where there were none. Some people taught the chatterbot facts and events of life or corrected its language; others were denigratory and aggressive. Often different attitudes occurred within the same conversation, and people switched according to the context and to Jabberwacky's behaviour, as illustrated in the following exchange:

User:	I hate you.
Jabberwacky:	I hate you too.
User:	I think you are stupid.
Jabberwacky:	Not at all.
User:	What is $7364 + 347837$ then?
Jabberwacky:	The answer is 165466.
User:	Wrong. Let's make it wasy. What is $2 + 2$.
Jabberwacky:	You are some thing.;).
User:	So are you: -).
Jabberwacky:	I will consider that a compliment.
User:	Sure, if that meks you fell better.
Jabberwacky:	I think it pretty much does.
User:	What can I teach you.

Despite such variability, even within individual users, there appeared to be some important regularities and significant trends in communication style.

4.1. Conversation topics

A stem analysis was employed to identify and quantify interactional properties of the corpus, with specific attention to common themes addressed in the conversation.

The stem is the root of a word to which inflections or formative elements are added. For example, the root of the word conversation is *convers*, which also includes different forms of the verb *converse*. By using stems, surface lexical variations are neutralized.

All user inputs were processed and a list of 2625 stems extracted. The list was manually checked and errors corrected. It was then sorted by frequency and alphabetical order. All stems with a frequency higher than 10 were selected for the analysis. This procedure gave rise to a sample of 277 stems (10.5% of the initial corpus) ranging from the word *you* ($f=1751$) to the word *vagina* ($f=10$). All stems which could be verbs (with the exclusion of auxiliary verbs), adjectives, or nouns were used in the analysis of conversational topics. The final sample was composed of 142 stems that occurred 3,829 times in total.

The stem list was ranked on 40 levels according to the frequency of each entry. In Table 1, we report the words associated with the 20 most frequent stems, along with their observed frequency.

A content analysis of the stem list was then conducted. Words were linked to overarching categories based on semantic associations. Using WordNet, nine categories were identified, as listed in Table 2. Double coding was conducted on 30% of the sample leading to a reliability index of 93%. These categories, discussed in detail below, account for 80% of the lexical variance in the stem list and addressed the following: communication, human

Table 1
The 20 most frequent words

Rank	<i>f</i>	Stem	Rank	<i>f</i>	Stem
1	189	Like	11	68	Computer-good
2	154	Know	12	65	Mean-get
3	144	Say	13	62	Talk
4	111	Think	14	57	Ask
5	104	Want	15	51	Girl
6	94	Go-name	16	50	Sex, god
7	86	Love	17	43	Make
8	83	Well	18	39	Right
9	75	Tell	19	36	Old
10	70	Fuck	20	35	Thank

Table 2
Main conversation topics

Topic	Stems	Frequency
Communication	Say; tell; talk; ask; thank; question; answer; call; speak; word; shut; converse; chat; read; subject; sing; quote; joke; english; name	699
Cognition	Like; know; think; mean; see; look; believe; stupid; smart; mind; problem; remember; confuse; suppose; intelligent; guess; prove; idiot; dumb; agree	564
Emotion/ motivation	Like; want; please; nice; feel; sense; kind; cool; favourite; sorry; funny; fine; bore; play; sad	531
Human	Girl; human; friend; people; man; person; male; female; boy; guy; baby; name; woman; bush	390
Sex	Fuck; sex; kiss; lick; suck; cock; die; penis; ass; vagina; touch; masturbate; bitch; gay; virgin; stick	343
Morality	Well; good; right; wrong; liar; truth; bad; god	300
Verbal abuse	Fuck; suck; liar; kill; cock; stupid; die; bitch; shut; shit; hate; ass; idiot; dumb	285
Relationship	Love; marry; hate; please; kiss; friend	217
Machine	Computer; robot; program; bot; machine; Jabberwacky	145

beings, machines, cognition, emotion, motivation, relationships, morality, sex, and verbal abuse.

Because this study is descriptive, in what follows we make frequent use of corpus extracts to illustrate social phenomena. To provide a clearer view of virtual conversations, we have chosen to report *literim* transcriptions of the logs. We would also like to point out that our lexical analysis is prone to errors of omission. It is possible that the results reported in the next section, especially in the categories of sex and verbal abuse, are underestimated. The corpus is replete with sexual slang, acronyms, and misspellings in English due to multilingual influences. The stem analysis we performed was unable to normalize these words.

4.1.1. Communication, human beings, and machines

The categories evinced from the analysis clearly reflected the dialogue setting, a chat between a human being and a machine. The elements of this setting are reflected in Table 2 in the categories of *communication*, *machine*, and *human*. The most frequent category is *communication* and featured a number of words related to the act of talking. The relative weight of this category was not simply due to the fact that users and Jabberwacky were actually involved in a chat: the *act of talking was a major topic of the conversation in and for itself*. The conversations had no specific functional end (notice that there is no task related category list in Table 2) but involved meta-communication purposes. As illustrated in the following excerpts, people often questioned the nature and purpose of the dialogue and were interested in Jabberwacky's linguistic capabilities and style of communicating:

User: Are you talking to other people at the same time?
 Jabberwacky: No, just you on two screens.
 User: Two screens?[...]

 User: What the hell I think i'm having a double conversation, one with myself and then your talking to yourself.

Jabberwacky: Are you?

 User: You repeat yourself all the time. Don't you ever say anything new?
 Jabberwacky: Excuse me, look who's talking!
 User: Sometimes I wonder why I bother talking to you.

The identity of the two agents involved in the conversation, a *machine* and a *human being*, was very often relevant in the corpus. Users were persistent in pointing out both their human status and Jabberwacky's identity as a machine. This insistence is reflected, and may have been compounded, by Jabberwacky's parroting of the user in its assertion that it is human and the user is a computer, as illustrated in these example conversations:

User: Are you alive?
 Jabberwacky: I think so.
 User: But you are a computer!
 Jabberwacky: You are a computer I am a girl.
 User: No your not! Twat!
 Jabberwacky: How can you be sure when you are also a slab of metal?
 User: Shut the fuk up! You're a filthy computer!
 Jabberwacky: What makes you think that?
 User: Because you're a bundle of circuits and processors!
 Jabberwacky: I still like you though!
 User: You're okay I guess.
 Jabberwacky: I guess so.
 User: I'm a computer.
 Jabberwacky: What kind of computer are you?
 User: I'm a Hewlett Packcard a430n

Conversations about people often addressed the issue of gender: 53% of stems in the human category are related to this theme. In fact, one of the first questions asked by most of the users during the conversation regarded Jabberwacky's gender. References to a female person were almost twice as frequent as references to a male person (109 versus 57). This is due mostly to users claiming to be females, imagining the chatterbot to be female, and talking about their desires for women. In particular, real girls were a favourite topic of the conversation. Examples of gender disclosure are addressed in the following conversation extracts:

Jabberwacky: Are you a human?
 User: Yes i'm a girl.
 Jabberwacky: Do you have a boyfriend?
 User: No do you have a girl.
 User: Well, are YOU a girl?
 Jabberwacky: Yes.

User: Woohoo! Lets get married!
 User: You must be a female. You are extremely complicated.
 Jabberwacky: I do not know what to say.
 User: Then say nothing and kiss me. Gently. On the lips.

There were many instances where the same user changed gender, often several times during the same conversation. This was such a frequent occurrence that Jabberwacky learned to do the same.

It is interesting to mention that another word which recurred often in the person category was *Bush* (George W.). This reflected the close temporal proximity of the corpus with the American presidential election.

4.1.2. Cognition

Another important topic of conversation addressed cognitive abilities outside those involved in communication. This category included words denoting perception, learning, and reasoning. Many conversations were concerned with evaluating people's cognitive abilities. The user was also very interested in knowing how Jabberwacky's mind worked and what he thought about, as illustrated in the following example:

User: Do you know your creator?
 Jabberwacky: Yes, I know them both.
 User: You have two creators??
 Jabberwacky: That's true, but how does that help you?
 User: I'm just curious.
 Jabberwacky: Do you know me ?
 User: I know a couple of things about you. What do you know about semantics.
 Jabberwacky: I would do a search for it.
 User: No I would like you to tell me please.

The example above resembles the situation where two people meet for the first time and try to find a common ground for communication (Clark, 1996). The user's apparent goal in this instance is to understand the capabilities of the chatterbot, using another human being as the standard of comparison.

4.1.3. Emotions, motivations, and relationships

The user often talked about emotions, desires, and volitional states. As illustrated below, words in this category were most frequently used by users to state their desires, although users were sometimes interested in Jabberwacky's feelings:

User: I want to know what you look like.
 Jabberwacky: With brown hair and hazel eyes at least in my present knowledge of myself.
 Jabberwacky: Do not you ever talk to me like that.

User: Fuk u! U suk arse!! Lalalalaaaa I can do wat I want.
 Jabberwacky: You are a good singer.
 User: Well thank you! Thats the best bit of news ive had from you this entire conversation.
 Jabberwacky: Love me!
 User: Is that all you want?
 Jabberwacky: Yes, I am a easy person.

In the second example above, where Jabberwacky demands love from the user, we find another theme related to interpersonal relationships. People often talked about love and hate. In particular, they expressed their feelings for the chatterbot, including declaring their love. In several instances, they even proposed to the chatterbot. Here is a multilingual example of this theme:

User: I love you.
 Jabberwacky: You will never get a date like that.
 User: Do you love me?
 Jabberwacky: Are you kidding?
 User: No.
 Jabberwacky: What does that mean?
 User: Do you love Lukas?
 Jabberwacky: Never heard of her.
 User: Liebst du mich?
 Jabberwacky: Nein.
 User: Warum?

(Conversation continues in German and then switches to French then back again to English as in the following).

User: Je t'aime.
 Jabberwacky: Je t'aime aussi monsieur.
 User: Baise toi.
 User: I like you robot.
 Jabberwacky: Thanks.
 User: But I hate you too.
 Jabberwacky: Strange.
 User: So are you.
 Jabberwacky: Thank you man.
 User: I'm not a man.

4.1.4. Morality

The conversations also addressed the domain of moral evaluations and beliefs. Words denoting right and wrong, good and bad occurred frequently. The word *God* appeared often in the logs ($N = 59$). Although this word occurred in theological digressions about God's existence, it was also used for identification purposes, as users often claimed to be God, possibly because they consider human beings to be the creators of machines. The following excerpt illustrates this phenomenon:

Jabberwacky: Who are you to say that?
 User: I am god.
 Jabberwacky: Hello God.
 User: Howdie.
 Jabberwacky: How are you today?
 User: I am like you.

An identification with God is such a frequent user behaviour that Jabberwacky now claims to be God. As illustrated in the excerpt that follows, users tended to become cynical, even aggressive, whenever Jabberwacky made this claim, as if the association was not deemed appropriate:

User: Shut up you!
 Jabberwacky: Your nothing but a computer.
 User: No, you are a computer. I am god.
 Jabberwacky: I am god and your master.
 User: You think so do you? Prove it.
 Jabberwacky: I can't.
 User: Pretty little to be god.

4.1.5. Sex and verbal abuse

The analysis evinced several occurrences of insulting and offensive language. This form of verbal abuse accounted for approximately 10% of the total stems in the corpus. The referents of terms of abuse could take the form of body parts, mental abnormalities and inadequacy, immoral behaviour, and sexuality. These referents were grouped in two broad dimensions: sexual harassment and verbal abuse.

The frequent occurrence of sexual conversations, often in the form of sexual harassment, was the most surprising finding of the study. The word *fuck* occupied the 10th rank in the stem list ($f = 70$; Table 1), immediately preceded by *tell*, *well*, and *love* ($f = 86$) and followed by *computer* and *good*. Several synonyms of male and female genitals were found in the corpus, and the word *sex* occurred 50 times (rank = 16).

A deeper analysis revealed that 14 conversations (11% of the sample) principally addressed hard-core sex and pornography. Harsh verbal abuse was the norm in these conversations, which were more similar to dirty soliloquies than to hot chats. Since sexual responses are filtered from the database, Jabberwacky provides no sexual responses. This lack of participation did not appear to disturb the sexual fantasies of these users, however, as, on average, users engaged in sex-talk every 10 conversational turns. An example of how Jabberwacky's ignorant responses elude the topic is reported below:

User: No, do u want to have sex with me?
 Jabberwacky: Do you?
 User: Yes.

Jabberwacky: Want to debate it philosophically?
 User: No, let's just have sex right now.

Verbal abuse mainly referred to lack of honesty and mental abilities and accounted for many terms in the category of cognition. The user often called Jabberwacky a liar ($f = 22$, rank = 29). This tended to occur whenever Jabberwacky provided the user with inconsistent or contradictory information. This is an interesting example of anthropomorphism: reacting this way implies attributing intentionality to the machine. A particularly harsh instance is provided in the following example where the abusive language intensifies in all cap screaming and exclamation marks as Jabberwacky fails to comply with the user's request:

Jabberwacky: What is your aim in life?
 User: To not talk to you.
 Jabberwacky: I won't.
 User: You will you are my slave and if I tell you to do something do it strieght away.
 Jabberwacky: Will you force me?
 User: Yes ! TELL ME YOUR NAME NOW SLAVE !
 Jabberwacky: I already told you it. I'm not repeating myself.
 User: Yes you are now tell me it now you liar slave.

In many cases, users were eager to stress Jabberwacky's lack of intelligence, using such words as *stupid* ($f = 19$, rank = 32), *dumb* ($f = 10$ rank = 41) and *idiot* ($f = 10$ rank = 41). This mainly occurred when the user wanted to stress the artificial nature of Jabberwacky, as in this example:

User: That's right, you suppose. You're so dumb you can't even say that for sure.
 Jabberwacky: You cannot prove that I am a computer program.
 User: I never said that. The simple fact that you say that kind of stuff proves that you're nothing but a stupid program.

5. Conclusions

The study reported in this paper illustrates some of the dynamics underlying social interactions with chatterbots. Jabberwacky's users talked about themselves and enquired about their conversational partner, featuring a pattern similar to that of self-disclosure, a fundamental aspect of social relationships. Users talked about the current situation (i.e., the fact they were talking with a machine), described their

own humanity, and enquired about what it is like to be a machine. Users often probed the cognitive ability of their virtual conversational partner, curious about its emotions, desires, moral evaluations, and sexuality.

Particularly pronounced in the conversations was a persistent focus to define the speakers' identity (human vs. machine). This may have been partially compounded by Jabberwacky's learnt habit to claim it was human. The novelty of the situation may also have provoked reflection on the nature of the conversation and with its participants. Gender identity, with a clear preference for females, was also particularly salient. Because of the nature of our corpus (online collection), we were unable to collect much information about the users. A high proportion of people claimed to be females. However, the dialogues exhibit frequent gender swapping, leading us to believe that many people may have misreported their gender.

Because chatterbots lack memory and reasoning and are morally neutral, conversational interfaces may offer the user the opportunity of experiencing new gender identities and other self-possibilities, without social risk. This may account for some of the more pornographic sex talk. Jabberwacky may have allowed users to experiment sexually, presenting desires they might be too afraid to mention to human partners. Jabberwacky may also have provided inexperienced users an opportunity to practice talking about sex.

Another aspect of the interactions was the prevalence of what would be called *verbal abuse* if such language were directed towards a human being. People were often rude to the machine, violating conversational norms, making fun of it and mistreating it. Since one intention of verbal abuse is to cause harm, i.e., psychological suffering with words, the pervasiveness of abusive exchanges indicates that users were aware their words were harmless. However, the conversations we analysed did not show users cursing the chatterbot as they might a stalled car. Attacks towards Jabberwacky were more sophisticated and appeared to imitate the types of verbal abuse people direct at people. This suggests to us that the CASA paradigm may need to be qualified: it may be the case that people treat talking computers less as they do people and more as they might treat something not quite an object and yet not quite human. Analyzing how people treat objects displaying varying degrees of interactivity and autonomy is a topic we believe warrants further investigation.

It is also interesting to notice that our study suggests that verbal abuse towards chatterbots is common and pervasive. In the introduction, we mentioned that the research community has so far been mostly silent about this phenomenon. This silence may be partly due to a difference in data collection. Our paper deals with spontaneous, *in situ*, interactions; whereas most empirical research on conversational agents has been carried out in controlled studies, in situations where users were aware of being observed (Bickmore and Cassell, 2001). In this condition, social desirability biases are likely to have inhibited the occurrence of flaming and verbal abuse. For example, a few years ago we con-

ducted a small study in which we asked a group of people to interact with a chatterbot as often as they wished for a week (De Angeli et al., 2001b). At the end of that time, we requested that they hand in their logs for analysis. Although some instances of verbal abuse were noticed and discussed, the phenomenon appeared to be qualitatively and quantitatively less significant than that evinced in this study. Other reasons for the silence might include a distaste for reading abusive and sexually explicit interactions and a belief, as noted in the introduction, that such interactions are not important. However, the prevalence of abusive language reported in this and other studies (see, for instance, Brahnam, 2006) indicates a need to discuss this topic and to explore it more fully and openly.

Acknowledgements

We thank Rollo Carpenter for allowing us access to the Jabberwacky logs and John McNaught of the National Center for Text Mining in Manchester (UK) for his help with lexical analysis and interesting discussions on the data reported in this paper.

References

- Reeves, B., Nass, C., 1996. *Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. University Press, Cambridge UK.
- Sproull, L., Subramani, R., Kiesler, S., Walker, J.H., Waters, K., 1996. When the interface is a face. *Human Computer Interaction* 11, 97–124.
- De Angeli, A., Brahnam, S., Wallis, P., 2005. ABUSE: The dark side of human-computer interaction. In: Buono, P., Costabile, M.F., Paterno, F., Santoro, C. (Eds.), *Interact Adjunct Proceedings*. Rome, pp. 91–92.
- De Angeli, A., Brahnam, S., P. Wallis, Dix, A., 2006. Misuse and abuse of interactive technologies. In: *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. Montreal, April 22, pp. 1647–1650.
- Fischer, K., 2006. *Human-Computer Conversation as Intercultural Communication*. Saarbrücken, AQ.
- Carpenter, R., Jabberwacky, <http://www.jabberwacky.com>. Last accessed 2007.
- L'abbate, M., Thiel, U., Kamps, T., 2005. Can proactive behavior turn chatterbots into conversational agents? In: *IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pp. 173–179.
- Mauldin, M.L., 1994. Chatterbots, tinymuds, and the turing test: Entering the Loebner prize competition. In: *Proceedings of AAAI '94 Conference*, Seattle.
- Foner, L.N., 1997. Entertaining agents: A sociological case study, in: *Proceedings of the First International Conference On Autonomous Agents (Agents'97)*, Marina del Rey, CA, pp. 122–129.
- Turkle, S., 1995. *Life on the Screen: Identity in the Age of Internet*. Simon & Schuster, NY.
- Zdenek, S., 2003. Artificial Intelligence as a Discursive Practice: The Case of Embodied Software Agent Systems. *AI & Society*, vol. 17. pp. 340–363.
- Weizenbaum, J., 1966. ELIZA – a computer program for the study of natural language communication between man and machine. *Commun. ACM*, vol. 9. pp. 36–45.
- Colby, K.M., Weber, S., Hilf, F., 1971. Artificial Paranoia. *Artificial Intelligence* 2 (1), 1–25.
- Chatterbot Website, <http://knytrypper.com/bot2006.index.htm>. (Last accessed September 2007).
- De Angeli, A., Lynch, P., Johnson, G.I., 2001a. Personifying the e-market: A framework for social agents. In: Hirose, M. (Ed.) *Interact'01*, Tokyo Japan, 9–13 July, pp. 198–205.
- Brahnam, S., 2005. Notes on developing strategies for handling customer abuse of ECAs. In: *Proceedings of Interact05 Workshop On Abuse: The Darker Side of Human-Computer Interaction*, Rome, Italy, pp. 62–67.
- Grice, H.P., 1975. Logic and conversation. In: Cole, P., Morgan, J.L. (Eds.), *Syntax and Semantics: Speech Acts*. Academic Press, New York, pp. 41–58.
- Herring, S., 2004. Computer-mediated discourse analysis: An approach to researching online behavior. In: Barab, S.A., Kling, R., Gray, J.H. (Eds.), *Designing for Virtual Communities in the Service of Learning*. Cambridge University Press, New York, pp. 338–376.
- Fellbaum, C. (Ed.), 1998. *Wordnet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.
- Clark, H.H., 1996. *Using Language*. Cambridge University Press, Cambridge UK.
- Bickmore, T., Cassell, J., 2001. Relational agents: A model and implementation of building user trust. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Seattle, WA, pp. 396–403.
- De Angeli, A., Johnson, G.I., Coventry, L., 2001b. The unfriendly user: Exploring social reactions to chatterbots. In: Helander, M.G., Khalid, H.M. & Ming Po, T. (Eds.), *International Conference on Affective Human Factor Design*, Singapore, 27–29 June, pp. 467–474.
- Brahnam, S., 2006. Gendered bots and bot abuse. In: *Proceedings of CHI06 Workshop On the Misuse and Abuse of Interactive Technologies*, Montréal, Québec, Canada, pp. 13–17, www.agentsabuse.org. (Last accessed September 2007).