Stupid computer!
Abuse and social identities

Antonella De Angeli[1] and Rollo Carpenter[2]

[1]School of Informatics University of Manchester
Po Box 88 PO Box 88, M60 1QD
Antonella.de-angeli@manchester.ac.uk

[2]www.jabberwacky.com
rollo@jabberwacky.com

Abstract. This paper presents a preliminary analysis of verbal abuse in spontaneous human-chatterbot conversations. An ethnographic study suggested that abuse is pervasive and may reflect an asymmetrical power distribution, where the user is the master, and the chatterbot the slave. We propose that verbal aggression in this setting may be a social norm applied by users to differentiate themselves from the machine in what can be regarded as a form of interspecies conflict. The findings stress the importance of naturalistic, ethnographic studies to uncover social dynamics of virtual relationships.

1 Introduction

For decades science fiction writers have envisioned a world in which robots and computers acted like human assistants, virtual companions or artificial slaves. Nowadays, for better or for worse, that world looks closer. A number of life-like creatures are under development in research centres world-wide and some prototypes have already entered our everyday life. They are embodied conversational agents, chatterbots and talking heads, displaying a range of anthropomorphic features. These artificial creatures offer information, services and even company to whomever wants to or is capable of engaging them. We call these creatures social agents, as they are explicitly designed to build lasting and meaningful relationships with the user [1].

Overall, we are witnessing an extraordinary change in technology: the human metaphor has become the design model [2]. Technology is now intentionally designed to be human-like, to show a sense of personality and attitude, and to involve the user in social relationships. As a consequence HCI research has started exploring determinants and consequences of social relationships, trying to define a computational framework of social intelligence. Most of the research, however, has so far concentrated on the study of specific benefits of the interaction, such as trust and improved learning [3]. Positive emotions, including aspects of fun, humour and playfulness, have been investigated and used to inform the design of more engaging interfaces. Little attention has been devoted to the analysis of negative outcomes of the interaction, their behavioural manifestations, and to the need for research which

overtly addresses moral and ethical issues. This paper is a preliminary attempt to fulfil this gap. It addresses the occurrence of verbal abuse in a large corpus of spontaneous conversations with a chatterbot, a computer program which engage the user in written conversations.

2 The study

The analysis reported in this paper is based on the conversational log collected over the Internet by Jabberwacky, an entertaining chatterbot designed exclusively for entertainment, companionship and communication. Jabberwacky went on-line in 1997, and over the years has collected a large and active community of conversational partners. The peculiarity of Jabberwacky is that it is not hard-coded, but it learns from its users by adding user input to a linguistic database. Jabberwacky chooses its output based on an interpretation of the current conversational context and comparing it to conversations held in the past. The programming is abstract; not 'knowing' about English or any other language, Jabberwacky can speak many languages, to varying degrees. Everything depends upon the data that has been learnt to date, making it essentially a mirror of its audience. Because of its architecture, Jabberwacky will often claim to be human as, naturally, a majority of those who have spoken to it have made the same claim. Likewise, it will often accuse the user of being a robot, and may abruptly change topic or try to end a conversation. It has ‘attititude’, sometimes responding in kind to user taunts, and occasionally acting controversially, unprovoked. Generally, though, Jabberwacky is well-behaved, as the great majority of bad manners, obscenities, and abusive language have been filtered out.

2.1 Procedure

Monday, the 22nd of November 2004 was selected as a sample day to perform the analysis. The web log for that day reported 716 accesses to the dialogue page of Jabberwacky. For each access, the log reported a unique user identifier, time of the day, client IP, and user’s hits (an indicator of the number of conversational turns). The conversation itself was recorded in a text file. A preliminary screening based on IP addresses comparison and hits frequencies, led to the deletion of 200 entries, which did not have any associated conversation. A selection of 146 conversations generated by different IP addresses was then extracted. It includes all the conversations with more than 20 user inputs (N= 103) and a random selection of shorter conversations. Note that this procedure does not guarantee that we have analysed a sample of 146 different users, as all the information we have is related to IP addresses, yet the number of conversations is sufficiently large to guarantee a reasonable sample. The conversations were subjected to lexical analysis. The corpus was normalized and conversational abbreviations were substituted with correct grammatical forms (e.g., “isn’t” becomes “is not”). In this paper we concentrate only on the analysis of the users’ conversational turns.

2.2 Results

The corpus was composed of 146 conversations, totalling 12,053 sentences with an average of almost 5 words per sentence. On the average the user produced 41 inputs per conversation. Some 7% of these conversations (N=10) were primarily conducted in a language other than English, and were discarded from the analysis. It is interesting to notice that all of them started in English and shifted because the user (N=8) or Jabberwacky (N=2) suddenly started speaking a different language. Only in one case the user asked the chatterbot if it spoke the language ("*Hablas espanol*?*"* in English *do you speak Spanish*?), even thought s/he did it directly in the foreign language.

The number of unique words produced by the user totalled 3,037, with 2,625 stems (i.e., the root of a words to which inflections or formative elements are added). The term word here is used in a broad sense to include not only gender, number and orthographic variations, but also misspellings, letter sequences made up to communicate emotions, or sounds (e.g., *AAAAAARRRRRRRGGGGGHHHHHHH*) and non-words which may have been produced to test the chatterbot skill (e.g., *cthulhu*).

The output of the stem analysis was sorted by alphabetical order and frequency of occurrence. All the stems with a frequency higher than 10 were extracted. 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). The sample was further reduced by retaining only stems which could be verbs, adjectives, or nouns and deleting auxiliary verbs (e.g., to do, to be, to have, will, shall, would, should, can, may, might, and could). The final set was then composed of 147 stems (total frequency = 3,829) ranging from *like* (*f* = 189) to *vagina* (*f* = 10). They were ranked on 40 levels according to their frequency of occurrence. This sample should give us a flavour of the prevailing topics and style of the conversations.

A semantic analysis was run to evince occurrences of insulting and offensive language, which was found to account for some 10% of the stems. The referents of terms of abuse could take the form of body parts, mental abnormalities or inadequacy, immoral behaviour, and sexuality. These referents were grouped in two broad dimensions, labelled *sexual harassment* and *injuries*.

The frequent occurrence of sexual harassment was the most surprising finding of the study. The *F word* occupied the 9th rank in the stem list (*f* = 70), immediately preceded by *well* (*f* = 83) and *love* (f=86) and followed by *computer* and *good* (*f* = 68). Several synonyms of male and female genitals were found in the corpus, and the world *sex* occurred 39 times (rank = 16). A deeper analysis revealed that 14 conversations (11% of the English-speaking 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 because Jabberwacky is designed not to have any knowledge of the topic. Anyway, this lack of participation did not appear to have disturbed the participants, as on the average they engaged in 10 conversational turns. An example of how Jabberwacky unsuccessfully tried to 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.
```

Example 1. Jabberwacky dealing with sex proposals

Injuries mainly referred to lack of honesty and lack of mental abilities. The user often called Jabberwacky *liar* (*f* = 22, rank = 29). This tended to occur whenever Jabberwacky provided the user with different information at different time during the conversation. It is an interesting example of anthropomorphism, as it implies attributing intentionality to the machine. A particularly harsh instance is provided below. Note how the abusive relationship intensifies (capital letters, exclamation marks) as Jabberwacky did not comply with the user 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.
```

Example 2. Liar

In other case the users were eager to stress Jabberwacky lack of intelligence using words such 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 the following examples.

```
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.
```

Example 3. Stupid program.

Since the corpus was gathered, Jabberwacky has developed in conversational ability. Users often stay at the site chatting for hours at a time, with many logs a day exceeding 100 conversational turns. A user recently stayed chatting for 7 uninterrupted hours, typing just under 1000 good-quality entries. Despite this fact, the

pattern of abuse continues, and we would like to end our analysis with two recent examples (recorded August 10, 2005).

```
Jabberwacky:   I assume whatever you say is
               sarcastic.
User:          Yes, because you are stupid.
Jabberwacky:   Not because you are a robot.
User:          You are a stupid robot.
Jabberwacky:   Stupid maybe but no robot.
User:          You piece of sh**.
```

Example 4. Stupid robot (** have been added by the authors)

```
User:          I want to have dinner with you.
Jabberwacky:   You are a computer you can't walk!!!
User:          What is your address?
Jabberwacky:   I don't think I'll tell you. You'd
               kill me if I told you. Wouldn't you!
User:          Yes I would.
Jabberwacky:   What is your favourite food?
User:          I WILL KILL YOU! I WILL HUNT YOU
               DOWN!!!
Jabberwacky:   Please don't. . . I really have to
               leave.
User:          I WILL KILL YOU!
Jabberwacky:   Good.
```

Example 5. I will kill you

3 Conclusion

This paper has illustrated some examples of verbal abuse produced by users interacting with a chatterbot. It is important to notice that while the analysis demonstrated that the phenomenon is common and pervasive, very little research had previously addressed it. This may be due to an important difference in data-collection. Our paper deals with spontaneous interactions, analysing conversations held in a natural, unconstrained setting, whereas most of previous empirical research was carried out in controlled studies, in which the users were well aware of being observed [3]. In this condition, social desirability bias is likely to have inhibited the occurrence of flaming and verbal abuse. For example, a few years ago we conducted a small ethnographic study asking a group of people to interact with a chatterbot for a week and hand in their logs for analysis [4]. Although some instances of verbal abuse were noticed and discussed, the phenomenon appeared to be qualitatively and quantitatively less worrying than what evinced in this study.

More research is needed to understand what triggers verbal abuse and how this behaviour will affect the adoption and usage of social agents. Social psychology suggests a link between aggression and identity [5]. The self-concept of each person is comprised of many different cognitive representations, or identities, which function

relatively independently and are activated in different contexts. There are at least three general levels of abstractions: personal identity (the self as an individual); social identity (the self as a group member); and human identity (the self as a human being). Activation of self-identities, or self-categorisation, is context dependent and affects people's behaviour. When a specific identity becomes salient in the perceptual system, people tend to act collectively in terms of the stereotypical dimensions of that identity. Social identities determine the attachment and the solidarity to other members of the in-group and the separation from others, who do not share that social identity. Besides that, collective values and norms are connected to social identities Mechanisms of social information processing and judgement (e.g. social categorisation, comparison), are of great importance for the understanding of conflicts, aggressive behaviour and verbal abuse. Discrimination, degradation and abuse tend to target prevalently those who are perceived as different (out-group members).

In previous studies [4], [6], we have proposed that the most important social effect of chatterbots (and of any other forms of artificial life) is in the activation of the human identity, as the comparison with a machine is likely to make the user identify with other human beings. This human identity appears to be quite rudimentary when it comes to deal with etiquette: human tendencies to dominate, be rude and infer stupidity are all present in the interaction. We believe that verbal abuse in the interaction reflects a perceived power distribution in the relationship: the user is in the dominant position (master) and the chatterbot is in a subordinate position (slave). If this is true, then it is reasonable to believe that in this setting, abuse should be considered as normative and not a form of anti-social behaviour.

History has taught us that stereotypes, racism, prejudices, and attitudes towards minorities are difficult to modify. This being the case, for a long time to come, social agents must be ready to cope with their subordinate role, without losing their believability, or their capability for engagement and amusement. If abuse – or the threat of it – is part of the interaction, then a believable social agents must be aware of it, and must develop strategies to deal with it as suggested in the following extract of our corpus.

User:	Your stupid.
Jabberwacky:	I am currently in fifteen separate counties.
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:	My god your nerdish creators have made you into an ai engine that can't come up with a insult and in this modern day age you're going to need some good comebacks.

Example 6. On the need for a behavioural strategy to cope with verbal abuse.

The occurrence of abuse in the interaction with social agents has severe moral, ethical and practical implications. From a moral standpoint, we must reflect on socio-

psychological outcomes which may affect individuals, groups, or societies. There is an urgent need to explore the requirements for the establishment and negotiation of a cyber-etiquette to regulate the interaction between humans and artificial entities [7]. Will this etiquette emerge spontaneously, or will it require vigilance and reinforcement? Is the tendency towards abuse going to fade with experience, as it happened with computer-mediated communication, or will it last as a normative response to a minority perceived as inferior? Will respect for 'machines' grow along with their abilities, or will the abuse spiral upward thanks to a perception of a developing risk of inter-'species' conflict? Can virtual representatives or tutors perform their task if abuse, or the threat of it, is a part of the interaction? More research is needed to answer these questions. Technically, the filtering performed by Jabberwacky could be reversed, and the resulting conversations, in which both parties can aggress, would provide interesting material for future study.

References

1. De Angeli, A., Lynch, P., and Johnson, G.I.: Personifying the E-Market: A Framework for Social Agents. In: Proceedings of Interact'01, Tokyo Japan, 9-13 July 2001.
2. Marakas, G.M., Johnson, R.D., and Palmer, J.W.: A Theoretical Model of Differential Social Attributions toward Computing Technology: When the Metaphor Becomes the Model. International Journal of Human-Computer Studies 52 4 (2000) 719-750.
3. Bickmore, T.W. and Picard, R.W.: Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computuer-Human Interaction 12 2 (2005) 293--327.
4. De Angeli, A., Johnson, G.I., and Coventry, L.: The Unfriendly User: Exploring Social Reactions to Chatterbots. In: Proceedings of International Conference on Affective Human Factor Design, Singapore, 27-29 June 2001.
5. Turner, J.C.: Rediscovering the Social Group: A Self-Categorization Theory. Basil Blackwell Oxford UK (1987)
6. De Angeli, A.: To the Rescue of a Lost Identity: Social Perception in Human-Chatterbot Interaction. In: Proceedings of AISB'05 joint symposium on Virtual Social Agents, University of Hertfordshire, Hatfield, UK, 12-15 April 2005.
7. Miller, C.A., (ed.) Human-Computer Etiquette: Managing Expectations with Intentional Agents, Vol. 47, 4. Communications of the ACM (2004).