Proceedings of the IASTED International Conference
Computational Intelligence (CI 2009)
August 17 - 19, 2009 Honolulu, Hawaii, USA

GETTING CLOSER TO HUMAN LEVEL HUMAN-LIKENESS AND ITS RELATION TO HUMOR IN NON-TASK ORIENTED DIALOGUE SYSTEMS

Pawel Dybala, Michal Ptaszynski, Rafal Rzepka, Kenji Araki
Graduate School of Media Science and Technology
Hokkaido University
Kita 14 Nishi 9, Kita-ku, 060-0814 Sapporo
Japan
{paweldybala,ptaszynski,kabura,araki}@media.eng.hokudai.ac.jp

ABSTRACT
Human-likeness of dialogue systems is an important, albeit neglected issue. In this paper, basing on evaluation experiments of humor-equipped chatterbot, we propose a method of measuring the distance between humans and systems and relation between human-likeness and humor. The results show that the presence of humor can enhance the performance of dialogue systems. A humor-equipped chatterbot was evaluated as more human like and generally better than one without humor, by both first and third person evaluators. The implications of this fact and novelty of evaluation method are discussed, and some ideas for the future are given.

KEY WORDS
artificial intelligence, human-computer interaction, evaluation methods

1. Introduction

There is a clear and scientifically proven relation between human's sense of humor and intelligence [1]. This leads to an assumption that – to construct human-like conversational systems – we need to take humor into consideration. However, this area of computational intelligence is still quite neglected, and this paper is a novel contribution to the field. Below we propose a method of measuring dialogue system's human-likeness and its relation to the presence of humorous stimuli.

1.1 Humor is good...

Although many papers in this area start with a similar statement, it has to be said: humor IS an integral part of our lives, regardless to culture and language. Our desire for it is so strong that many of us are willing to pay some money just to see and hear comedians (humor professionals) making us laugh. There are many scientific proofs for the positive role of laughter. To list only some of them - it has been shown that humor can generally make us feel better and help deal with negative emotional states, such as stress [2], depression [3] or mood disturbances [4]. Presence of humor can make boring contents more interesting [5] or can improve perceived benefits of a relationship [6].

1.2 ...also in computers

While positive effect of humor in our lives has been proved in many researches, its role in computer science is still unexplored. There are very few studies on the influence of humor on human-computer interaction (HCI). One relatively convincing is the research conducted by Morkes et al. [7] which showed that humor-equipped (albeit not humor-generating) task-oriented system was evaluated as more sociable, likeable and easier to cooperate with by the users.

In his research Morkes's investigated the role of humor in HCI, without any automatic generation. There are, however, numerous projects in the field of computational humor, aimed at creating joke generating algorithms. One of the most popular genres in this field is so called "puns" – jokes based on features of the language, such as homophony or polysemy. One of the first and probably most robust systems in this field is Binsted's JAPE – punning riddles generator [8]. Basing on a WordNet-related lexicon [9], it was able to generate quite a large spec of riddles – however, most of them were not evaluated highly by humans. Also worth mentioning is McKay's WISCRAIC system [10], generating simple puns in three different forms (question-answer, single sentence and two sentences sequence).

The main problem with virtually all existing humor generating systems is that their outputs are jokes in isolated, closed forms. For example, riddles generated by JAPE, like the one below:

-How is a nice girl like a sugary bird?
-Each is a sweet chick.

may seem funny as such, but it would be difficult to include them into normal (daily) interaction between users and computers. This in fact restricts possible applications of such systems – the best we can get is a system that tells jokes without any wider interaction context, just in isolated forms.

There were two attempts of integrating JAPE into a system that interacts with users – one conducted by Loehr (combined it with an online game playing system Elmo – evaluated lowly due to the lack of relevance between users' utterances and the system's output [11]), and another one by Ritchie et al. (successfully implemented JAPE into an interface that interacts with children with complex communication needs [12]). Especially the latter work is worth mentioning, as it was proved to be successful in therapy for children with CCN, which means that the humor generator was actually used in a working application that is useful for a group of people.

However, even Ritchie's et al. application of JAPE shares the same problem other joke generators do – even if they are useful to some users, it is still hard to imagine these systems being used by average, healthy members of society.

Therefore, in the research described in this paper, the pun generator was implemented into a non-task oriented conversational system, and the jokes were generated using parts of user's utterances (base words – see **2.2**) as input, which gives most jokes at least some relevance to what the user said.

1.3 The need for freely talking systems

In recent years, we have seen numerous research projects, aiming at creating dialogue (conversational) systems. Most of them focus on so-called "task-oriented" dialogue systems, such as tour guide agents or information kiosks, performing conversations aimed to achieve clearly defined goals. Such devices are obviously useful and practical. However, recently the world of computer science is starting to understand the need for systems that would be able to talk with us without specific goal (so-called "non-task oriented" systems or "chatterbots". The usefulness of such devices has often been questioned - however, there are some applications in which the chatterbots would be highly beneficial, as, for example, companions for lonely and elderly people or chatting car navigators.

As described above, humor was proved to have many beneficial features in our lives, also in human-computer dialogue. Thus, we assumed that implementing humor generating engine into freely talking conversational system should improve the latter performance and increase its human-likeness and likeability (from the user's point of view). To check if the assumption was true, we conducted evaluation experiments of a joking conversational system and analyzed the results.

1.4 Evaluating human-likeness

The very idea of computer systems being human-like is still quite controversial. The main argument against it is that, after all, computers are "all about zeros and ones", and thus by definition cannot be even compared with humans. As a matter of fact, this is why the Turing Test [13] received so much criticism. To name one of the best known one, we have Searle's "Chinese room" argument [14], saying that even if a machine acts as human, it cannot be intelligent or "natural" in any way.

Human likeness is an important issue of research on virtual 3D-agents or androids. Here it includes such features as appearance (its resemblance to humans), gestures, movements (including eye movement), voice etc. The advancements of nowadays science allow us to proceed further and further in the process of imitating ourselves in as detailed way as possible. In recent years we have even seen first attempts of making live-action movies (such as "Final Fantasy: The Spirit Within", "Advent Children" or "Beowulf"), made entirely with virtual humans (computer-generated characters).

These, however, are virtual agents, while today's science goes even further – a closer look at the actroid created by Kokoro Co. [1] makes many people believe that we are only one step away from constructing artificial human beings.

Here, however, we face a crucial question: do we really want the machines to be human like? How will we feel about it, when one day we meet face to face with something (someone?) that looks and behaves like us, but is not human? In fact, these doubts were formulated by Mori [15], who indentified the dilemma as "the uncanny valley" problem. He predicted, that in the development of human-like devices we will eventually get to a point in which the human-likeness stops being attractive to users (partners in interaction) and becomes eerie and unnerving. The fact that machines look too much like us could be perceived as unnatural and thus – frightening, which obviously is not what we are aiming at.

However, the same dilemma does not necessarily have to apply to all dimensions of human-computer interaction. A user-oriented study on robot companions, conducted by Dautenhahn et al. [16] showed that only 29% of experiment participants wished for the robot to appear more human, 36% to behave more human, and 71% to communicate more human. This gives us an very information about the needs of users – namely, that (at least in some cases) the human-likeness in communication layer is by all means desirable.

Here, however, we face another question: do we want all conversational systems to be human-like? The answer, of course, is quite complex, but one thing can be said for certain – it depends on the application. In case of task oriented dialogue systems, we may desire them to be even better than humans – artificial tour guides should know more than human guides, question-answering (QA) systems should be able to find answers to every single user's question etc. However, in case of chatterbots, considering such applications as car navigators or robot companions, it is the human-likeness that is desired in the first place, and thus it should be included in evaluation experiments of such systems.

[1] http://www.kokoro-dreams.co.jp/english/robot/act/index.html

Unfortunately, evaluation methodology for measuring human-likeness is quite a neglected field. The methods for studying 3D agents or androids include such features as facial expression or voice generation analysis. In the case of conversational system, we can talk about two types of human-likeness: one related to "technical" abilities of a system (i.e. grammatical or semantic correctness, vocabulary richness etc.- desired in both task- and non-task-oriented systems) – this is relatively easier to study – and another one (albeit correlated with the first one), related to general "naturalness" of the system's behaviour. The latter is much more vague, as it relies on users' subjective impressions only, and thus is more problematic to evaluate. The Turing Test may work well in some cases – however, it does not give us quantitative (measurable) data, as the only thing the users do (apart from interacting with a system) is guessing if the partner was a computer or a human. Therefore, the results are hard to compare in details.

One applicable quantitative way to check system's humor sense is to simply ask the users to assess it in a numeric scale (in this research we use a 5-point one). Although the users, as participants of the interaction, may not have the distance to the subject of evaluation, it is them who will use the application in the first place, so their opinion is of high importance. However, if we want to check the system's human-likeness with a slightly higher level of objectivity, we can also perform a third person focused evaluation experiment, in which the chat logs from system's interaction with humans are evaluated by non-user participants. In this paper we present and discuss results of such two experiments (user- and third person focused). The latter are analysed using a comparative method (comparing differences of humans' and system's evaluation – see **4.2**).

2. The systems

In this paper, we briefly describe joking system called "Pundalin" (introduced in [17]), explain its algorithm (**2.2**) and evaluation experiments (**3** and **4**).
Pundalin was constructed by merging Dybala's et al. joke generator PUNDA Simple - a simplified version of PUNDA Japanese pun generator [18] - with a freely talking conversational system Modalin created by Higuchi et al. [19]. Modalin itself was also used in the evaluation experiments, in which its results were compared to these of Pundalin.

2.1 Modalin

The first system in our research is Modalin - freely talking keyword based conversational system, created by Higuchi et al. [19] For the conversation topic set freely by the user, the system extracts related sets of words, basing on keywords spotted in user's utterance. Next, word associations are extracted in real time using Goo search engine[2] snippets (without previously prepared resources, such as off-line databases). In the next step, the system applies extracted word associations into proposition templates, like: [(noun) (topic indicating particle *wa*) (adjective)]. Next, the system checks the naturalness of each sentence proposition using the Internet. If an unnatural proposition is generated, the system generates next proposition in the same way. Next, the system adds modality (expressions such as "well" or "yeah,") to the extracted natural proposition and again checks the semantic correctness of the proposed sentence in the Internet.

To sum up, Modalin is a system that answers user's utterance with a modality-added sentence that corresponds to its topic. An example of such conversation can be found below:

User: - *Nanika sukina tabemono aru?* (What food do you like?)
Modalin: - *Maa, tabemono-wa oishii desu.* (Well, food tastes good.)
User: -*Saikin-wa osake-mo sukini nattekitanda.* (Recently, I began to like alcohol too.)
Modalin: - *Demo, sake-wa yowai-no-yo-ne.* (But, I can't drink much.)

Modalin was also used as a "base" for creating Pundalin (see **2.3** for details).

2.2 Pun generator

The PUNDA pun generator was developed by Dybala et al. [18] as a part of PUNDA research project, aiming to create a Japanese pun generating engine.

The system is also based on the Internet. From user's utterance it extracts a base word (usually a noun) and transforms it using Japanese pun phonetic generation patterns, to create a phonetic candidate list.

All of phonetic generation patterns base on "moras" – small phonetic units, roughly equivalent to syllables. In the current version, the system uses 4 such patterns, proposed by Dybala in one of his earlier works [20]: homophony, initial mora addition, internal mora addition, and final mora addition. An example for the word *karada* (a body) is showed below:

candidates for base word {*karada*}:
1. **homophony**: {*karada*}
2. **initial mora addition**: {**karada* } (*akarada, ikarada*...)
3. **final mora addition**: {*karada**} (*karadaa, karadai*...)
4. **internal mora addition**: {*ka*rada*}, {*kara*da*} (*kaarada, kairada*...)

After generating the phonetic candidates list, the system checks all candidates in the Goo search engine, and chooses the one with the highest hit rate (i.e. the most

[2] Goo search engine, http://www.goo.ne.jp/

common word that sounds similar). Next, it uses pun templates, extracted from Sjöbergh and Araki's Japanese puns data base [21] to generate a humorous answer. An example of such template is given below:

{speaking of [base word], it's [pun candidate]}

The system also uses KWIC on WEB - online Keyword-in-context sentences database [22] – to integrate the chosen candidate into a sentence (for details, see [17]).

Below we present an example of the system in action:

User: -*Kaze ga tsuyoi hi ga yasashiku nasasou da*.
(Windy days don't seem too nice)
[base word: kaze (wind), pun candidate: *kazen* (as expected)]
[template: speaking of [base word], it's [pun candidate]
Pundalin: -*Kaze to ieba kazen da yo ne*.
(Speaking of wind, it was as expected)

It happens, though, that no pun candidate at all can be found for a base word. In such cases, the system uses the pun data base as a "last resource", and one pun is chosen randomly to be presented to the user.

2.3 Pundalin – Joking Conversational System

PUNDA Simple pun generating engine was implemented into Modalin freely talking system to create Pundalin – a joking conversational system (described in details in [17]). For the timing of jokes, a very simple rule was applied – in every third turn of the conversation, Modalin's output was replaced by a joke-including sentence, generated by PUNDA Simple. In other words, user's every third utterance becomes an input for PUNDA, which generates an appropriate pun for it. Preliminary tests showed that joking in every third turn is optimal for this experiment. This method, although quite simple, allowed checking if the usage of humor improved the system's overall performance.

3. Experiments

The impact of humor on the performance of chatterbot was checked in two evaluation experiments, with Modalin as the baseline, (non-humorous system) and Pundalin as the main, humor-equipped system. The methods were: 1) first person (users) focused evaluation and 2) third person (non-user) focused evaluaton.

Among them, the latter may seem slightly more objective, as non-user evaluators tend to have more distance to the evaluated subject – however, it is still humans who assess the product, and this by definition cannot be fully objective.

3.1 First Person Focused Evaluation

In the first experiment, we asked 13 subjects (11 males and 2 females) to perform a 10-turn dialogue with Modalin and with Pundalin. No topic restrictions were made, so that the talk could be as free and human-like as possible. All conversations were typed.

Having talked with both systems, each evaluator was asked to fill in two questionnaires about the systems' performance. The questions were: 1) Do you want to continue the dialogue? (in the tables below referred to as CONT); 2) Did you get an impression that the system possesses any knowledge? (KNOW); 3) Did you get an impression that the system was human-like? (HUM); 4) Do you think the system tried to make the dialogue more interesting? (TRY) and 5) Did you find system's talk interesting (INT)?

The answers for questions were given in 5-point scale. Each user filled in two such questionnaires, one for each system. In the end, one summarizing question was asked: "Which system do you think was better?". Applying such approach may seem little bit too general, but – from a user's (or a customer's point of view) – deciding "which is better" is a natural way to compare similar entities.

3.2 Third Person Focused Evaluation

To verify user's assessment, we conducted a third person evaluation experiment. The questionnaires were similar to these used in user's evaluation experiment, with few differences. The word "system" was changed to "dialogue" or "speaker", as we did not want the evaluators know that one of the dialogue participants was actually a computer. In the chat logs, the users were referred to as "Speaker A" and the systems as "Speaker B". For the same reason, question 3) (about human-likeness) was deleted – instead, we performed an analysis to check the actual human-likeness perception in this experiment (see section **5.2**). In questions 2), 4) and 5) we added two options: 1) "Speaker A" and 2) "Speaker B" – so that the dialogue participants would be evaluated separately. After completing the detailed questionnaire, evaluators answered the final question, the same as in the previous experiment - "Which dialogue do you find most interesting?"

There were 13 sets of chat logs, each including one non-humorous and one humorous dialogue. Each set was evaluated by 5 people (a total of 65 participants) [17].
The third-person oriented method also allowed us to study human-likeness of both systems – not in such an explicit and direct way as the user-oriented evaluation, but by calculating the differences of scores for both (human and non-human) speakers (see **4.2**).

Table 2
The differences between Modalin and users compared to the differences between Pundalin and users (third person evaluation experiment). Minus values mean that Speaker B (the system) received higher scores than the user. Question 1) did not have separate options for two speakers, and question 3) (about human-likeness) was deleted to hide the fact that one of the speakers was not human

	Modalin				**Pundalin**			
Question	**User**	**Modalin**	**Diff.**	**P value**	**User**	**Pundalin**	**Diff.**	**P value**
2.KNOW	3.13	1.87	**1.26**	<0.05	2.97	2.13	**0.84**	<0.05
4. TRY	2.54	2.51	**0.03**	>0.05	2.52	2.91	**-0.39**	<0.05
5. INT	2.85	2.73	**0.12**	>0.05	3.09	3.16	**-0.07**	>0.05

Table 1
User's evaluation – results for Modalin and Pundalin for detailed questions (see 3.1). Answers were given in a 5-point scale

Question	**Modalin**	**Pundalin**	**Difference**	**P value**
1.CONT	2.62	3.38	**0.76**	<0.06
2.KNOW	2.15	2.85	**0.70**	**<0.05**
3.HUM	2.38	3.31	**0.93**	**<0.05**
4. TRY	1.92	4.15	**2.23**	**<0.05**
5. INT	2.46	4.08	**1.62**	**<0.05**

4. Results

As showed in the Tables 1 and 2, in both experiments the system with humor received higher scores in all categories. For the needs of this paper, the most important message is that the system with humor was perceived as more human-like directly by the users (**4.1**). This was confirmed by the comparative analysis of third person focused evaluation results – the differences between humans and systems were smaller in the case of Pundalin (with humor).

4.1 First Person Focused Evaluation

85% of users (11 out of 13) found the system with humor system to be generally better than the one without humor. The humor-equipped system received higher scores also for detailed questions, including the one about human-likeness.

For each question, significance of the differences between Modalin and Pundalin was checked using the Student's t-test. Apart from question 1) (P value < 0.06), all results were found statistically significant on 5% level (P value < 0.05), which is a commonly used significance threshold in statistics.

4.2 Third Person Focused Evaluation

As we expected, the results in the third person focused experiment were not that good as in the user-oriented one. However, the overall question's results still show that humorous system is visibly better than the non-humorous one. 45 out of 65 evaluators (69%) pointed at the system with humor as more interesting.

As far as detailed questions are concerned, we compared the scores given by the third person evaluators to the utterances of Speaker A (users) with those of Speaker B (the systems - Modalin and Pundalin). The results show that the humor-equipped system differs less from humans than the non-humorous one. In other words, the difference between humans and Pundalin was smaller than the difference between humans and Modalin – see Table 2 and Figure 1.

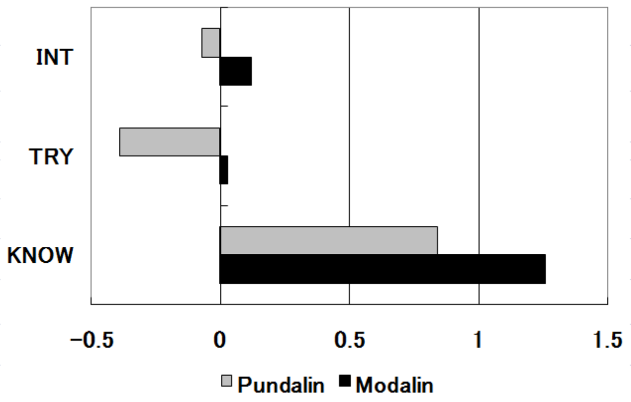

Figure 1. The differences between Modalin and users compared to the differences between Pundalin and users (third person evaluation experiment). Minus values mean that Speaker B (the system) received higher scores than the user

For the question 1) ("Do you want to read the continuation of the dialogue?"), the questionnaire did not include separate options for Speakers A and B. The results of Pundalin also here were slightly higher (difference between the two systems = 0.29) than those of Modalin – however, their statistic significance is questionable.

The implications of these results are discussed in **5**.

5. Discussion

The presumptions we made before the experiments were correct. The results showed that the presence of humor improved the chatterbot's performance in all investigated categories, including human-likeness, which we were mostly concerned with in this case.

5.1 First Person Focused Evaluation

As far as users' answers to the detailed questions are concerned, all of them point at Pundalin. Although t-test showed that the difference in question 1) is not statically significant on a 5% level, the P value here was 0.054 – which suggests that including two or three more users in the experiment might be sufficient to achieve significance also here. Differences in other questions are statistically significant on a 5% level.

The differences between Modalin and Pundalin were clearly visible in questions 4) and 5) (directly related to the presence of humor), which means that not only did the system try to entertain the user, but in most cases the attempts were successful. The results for Pundalin were also higher for the question 2), regarding knowledge possessing.

Results for question 3) show that the Pundalin is considered as more human-like than the non-humorous system. This is a very important issue for the topic of this paper, and leads to the assumption that further development of Pundalin shall lead to creating more human-like machine in general. Also, we can say that the results acquired in this category are consistent with the findings of Dautenhahn et al. [16] (see **1.4**).

What is also important – we did not give or even suggest to the users any definitions of the "human-likeness". This means that they probably understood it commonsensicaly – as, more or less, "behaving in a way humans do".

The question is, however, if this users-defined human-likeness is a category that can be used in a robust study. In other words – if all participants could define the human-likeness freely, how can we know that they had the same thing in mind?

Although, as mentioned above, we do not aim to define the phenomenon of human-likeness here, we agree that analyzing the user-only point of view is not enough to be sure that one system was more human-like than the other. Therefore, the results of user-focused experiment were double-checked in the third-person (non-user) focused evaluation experiment.

5.2 Third Person Focused Evaluation

Consistently with results described above, also the third person evaluation experiment showed that humor-equipped system is generally better.

As far as detailed questions are concerned, we compared the results of Speaker A (the users) and Speaker B (the systems) to check, if the presence of humor has any influence on the differences between users and systems. As showed in Table 2, in all cases the differences between humans and Pundalin were smaller than those between humans and Modalin. In other words, the humor-equipped system proved to be closer to the human level than the system without humor.

For question 2) ("Did you get an impression that Speaker A/B possesses any knowledge?"), the difference between Pundalin and users was 0.42 smaller than in case of Modalin (both differences statistically significant on 5% level). This means that Pundalin's "knowledge" was evaluated as visibly closer to human level.

For question 4) ("Do you think that Speaker A/B tried to make the dialogue more funny and interesting?"), the difference was clear. In case of Modalin, differences between users and the system were not statistically significant (P value >0.05), so it can be stated that these two present similar level. Contrary to this, in case of Pundalin, the difference is statistically significant, and, what is more interesting, the result points at the system as the speaker that tried to make the conversation interesting (0.39 higher score than users'). This means that in this category, system's efforts were more visible than humans' – another question is, if Pundalin did not just try too hard (stubbornness of the system might as well annoy the user). While this issue still needs more research, we assume that part of the answer lies in the results of the general question in this evaluation – the fact that almost 85% of users and 70% of third person evaluators chose Pundalin's dialogue as better suggests that system's attempts to make conversation more funny and interesting were rather appreciated than disliked.

In case of question 5), however, the differences are not that visible and significant and it can be stated that the presence of humor did not influence them in such clear manner as in other categories. However, overall results for this question are still slightly better for Pundalin (comparing Speaker B's scores in both systems).

Finally, the results for general question showed that 69% of evaluators chose Pundalin's dialogue as better and more funny than those of Modalin. This is also consistent with other results described above, and, what may be even more important, shows that even if some differences are not very significant, evaluators still point at humor-equipped system as more human-like, interesting and generally better.

The most important for this research are the results concerning human-likeness of both systems. The differences between human and non-human speakers calculated as a part of the third-person experiment show a tendency that is consistent with the results of the first-person oriented evaluation. This "double-check" gives us more confidence that the results are valid and the humor-equipped system was actually seen as more human-like.

Also, another issue is worth our attention here. There was one big difference between the two experiments: in the first one, the participants (users) obviously knew that their partner is a computer, whereas in the second experiment we have hidden this knowledge from the evaluators. However, this setup might have caused some incongruities – the users, for example, evaluated the perceived human-likeness of something they know is not human, which seems quite a contradiction. This setup, though, was necessary if we wanted to ask them directly about the human-likeness – if the identity of the partner was hidden, the question about resemblance to human would reveal the whole mystery. On the other hand – in the case of the third-person focused experiment, we wanted the evaluation to be slightly more objective that the first one, and thus we decided to have both (human and non-human) speakers evaluated on equal rights. The results are quite convincing – however, it seems a good idea to check what they would look like if the third person evaluators knew that one of the speakers was non-human. In the near future we are planning to conduct an experiment with such a setup.

6. Conclusion and future work

Evaluating human-likeness of dialogue systems is quite a troublesome issue. In this paper we showed two methods to do that. Especially innovative is the third person focused comparative method, as, to our knowledge, no preceding research applied such approach. The methods were tested on a humor-equipped dialogue system. The results showed that the presence of (even very simply generated) humorous stimuli can visibly improve the dialogue. As for the issue of human-likeness, users' answers for the direct question were consistent with the results of the comparative analysis, performed as a part of third-person oriented experiment. This shows that both of these methods are actually working. They can also be used in evaluation of other types of systems, in all cases when there is a need to quantitatively measure human-likeness.

As far as future work is concerned – currently the third person oriented experiment is being repeated with one important difference: the evaluators are told that one of the speakers is a computer system. The results will be presented in the near future and compared with the experiment described here.

We are also working on alternative methods of evaluation (also for human-likeness), such as automatic emotiveness analysis based evaluation, in which the logs from user-focused experiments would be analyzed by emotiveness analysis system.

Humor-equipped dialogue systems and their evaluation methods are still a very neglected field of modern science. This paper is an important contribution to both of these areas, and the results presented here can be used in further research.

Acknowledgements

The research presented in this paper was partially sponsored by a research grant from the Nissan Research Foundation and the Center for Engineering Education Development at Hokkaido University.

References

[1] A. Feingold, & R. Manzella, Psychometric intelligence and verbal humor ability, *Personality & Individual Differences,12*(5)1991, 427-435.

[2] A. Cann, K. Holt, & L. G. Calhoun, The roles of humor and sense of humor in responses to stressors, *Humor: International Journal of Humor Research*, *12*(2), 1999, 177–193.

[3] A. Danzer, J.A. Dale, & H.L. Klions, Effect of exposure to humorous stimuli on induced depression, *Psychological Reports, 66*(3, Pt 1), 1990, 1027-1036.

[4] S.M. Labott, & R.B. Martin, The stress-moderating effects of weeping and humor, *Journal of Human Stress, 13*(4), 1987, 159-164.

[5] R.A. Dienstbier, The impact of humor on energy, tension, task choices and attributions: Exploring hypotheses from toughness theory, *Motivation and Emotion, 19*(4), 1995, 255-267.

[6] K.S. Cook, & E. Rice, Social exchange theory, J. Delamater (Ed.), *Handbook of social psychology, 11. Difficult subjects*, NewYork: Plenum, 2003, 53—76.

[7] J. Morkes, H.K. Kernal, & C. Nass, Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of srct theory. *Human-Computer Interaction*, *14*(4), 1999, 395-435

[8] K. Binsted, Machine humour: An implemented model of puns, Univ. of Edinburgh, 1996.

[9] C. Fellbaum,. WordNet: An Electronic Lexical Database, MIT Press, Cambridge, Mass, 1998.

[10] J. McKay, Generation of idiom-based witticisms to aid second language learning, In Stock et al., 2002, 77–87.

[11] D. Loehr, An integration of a pun generator with a natural language robot, Proc. Intern. Workshop on Computational Humor, University of Twente, Netherlands, 1996,. 161-172.

[12] G. Ritchie, R. Manurung, H. Pain, A. Waller, R. Black, & D. O'Mara, A practical application of computational humour, Proceedings of the 4th International Joint Conference on Computational Creativity, 2007, 91-98.

[13] A. Turing, Computing Machinery and Intelligence. *Mind 59* (236), 1950, 433-460.

[14] J. Searle, Minds, Brains and Programs, Behavioral and Brain Sciences 3 (3), 1980, 417–457.

[15] M. Mori, *Bukimi no tani* [the uncanny valley]. *Energy*, 7, 1970, 33-35.

[16] K. Dautenhahn, S. Woods, C. Kaouri, M. Walters, K. Koay, & I. Werry, What is a robot companion -friend, assistant or butler?. Proceedings of International Conference on Intelligent Robots and Systems (IROS 2005), Edmonton, Canada, 2005, 1192-1197.

[17] P. Dybala, M. Ptaszynski, S. Higuchi, R. Rzepka, & K. Araki, Humor Prevails! - Implementing a Joke Generator into a Conversational System, Proceedings of the 21st Australasian Joint Conference on AI (AI-08), Wobcke, W. and Zhang, M. (eds), Auckland, New Zealand, 2008. Springer-Verlag Lecture Notes in

Artificial Intelligence (LNAI) Vol. 5360, 214-225, Springer Berlin & Heidelberg, 2008.

[18] P. Dybala, M. Ptaszynski, R. Rzepka, & K. Araki, Extracting Dajare Candidates from the Web - Japanese Puns Generating System as a Part of Humor Processing Research, Proceedings of the First International Workshop on Laughter in Interaction and Body Movement (LIBM'08), Asahikawa, Japan, 2008, 46-51.

[19] S. Higuchi, R. Rzepka, & K. Araki, A Casual Conversation System Using Modality and Word Associations Retrieved from the Web, Proc. of EMNLP ‘08, Honolulu, USA, 2008, 382-390

[20] P. Dybala, *Dajare - Nihongo ni okeru dōon’igi ni motozuku gengo yūgi* (*Dajare* – Japanese puns based on homophony), Jagiellonian Univ., Kraków, Poland, 2006

[21] J. Sjöbergh, & K. Araki, Robots Make Things Funnier, Proceedings of LIBM'08, Asahikawa, Japan, 2008, 46-51.

[22] K. Yoshihira, Y. Takeda, & S. Sekine, KWIC system for Web Documents. (in Japanese), Proc. 10th Annual Meeting of the Japanese Association for NLP, Japan, 2004, 137-139.