Impact of agent's answers variability on its believability and human-likeness and consequent chatbot improvements

Mao Xuetao[1], **François Bouchet**[1] **and Jean-Paul Sansonnet**[1]

Abstract. Although globally less efficient than advanced dialogue systems, the chatbot approach allows people to easily design conversational agents. We suggest that one of their main drawbacks, their lack of believability, could be bypassed through the addition of variability in their answers, particularly when the variations depend on previous interactions or on particular parameters defining the agent. We validate the legitimacy of that hypothesis in two steps: first through simple additions to our chatbot-like framework (DIVA), we show it is technically feasible to simulate degrees of variability in answers. Then through an experiment done on 21 subjects interacting with two among six DIVA agents with different degrees of variability in a classical meeting scenario, we show that agents with an advanced variability in their answers are indeed perceived as more believable, human-like, and globally, more satisfying.

1 INTRODUCTION

1.1 Context

Assisting Conversational Agents (ACA) are virtual characters embedded into an artefact (i.e. software applications and services, smart objects, etc.) which purpose is to provide a Natural Language & Artificial Intelligence-based assistance to ordinary users interacting with that artefact. More specifically, here, we will consider the two key points of an ACA are 1) its ability of interaction in natural language with people from the general public and 2) its ability of symbolic reasoning about the structure and the functioning of the assisted artefact.

Indeed, associating such an assistant agent with a new product has long been considered a good approach to improve their immediate social acceptability, because natural language brings more naturalness in the interaction and symbolic reasoning brings more believability in the agent. However, till now it has endured many setbacks, what we could call the "Clippy Effect" [1][2] being the most prominent one. This phenomenon is consistent with the disinterest of novice users for help systems (the motivation paradox described in [3]) which has issued in the recent Contextual Help System approach [4] aiming at providing a more contextualized the assistance.

Overcoming those problems leads to a difficult dilemma where one has to choose between: 1) a complex custom dialogue system, like TRAINS [5], which works well (especially when used by corporate people) but entails a critical cost effectiveness issue (in terms of development duration and manpower linguistic skills - which led Allen to promote the notion of genericity as a major challenge in dialogue systems of the future [6]). 2) A naive chatbot-like system, like ALICE [7][8], Elbot [9], etc. They are very cost effective and have proved to be well-accepted by the general public (Eliza effect [10]), but lack the symbolic reasoning capabilities and the fine semantics analysis capabilities required to support the Function of Assistance, as shown by Wollermann in [11][12] for four main chatbots (ALICE, EllaZ, Elbot and ULTRA-HAL).

In our work on assisting agents, concurrently with an advanced semantic-based approach to capture precisely requests' subtleties [13], we are also exploring an ACA architecture based on a simpler chatbot-like system. Relying on a bottom-up approach, the basic chatbot is provided with a) an improved Natural Language Processing (NLP) chain (cf. figure 1) and b) reasoning heuristics over a symbolic model (task, agent and user).

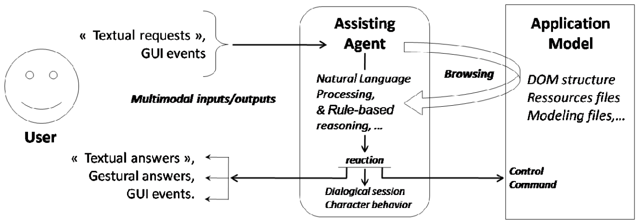

Figure 1: Conversational chain of the assisting agent.

1.2 Key issues

That architecture has proved to meet the goal of cost-effectiveness: a dozen of students without previous experience of agent scripting have been able to use that framework to easily design various assisted applications (see the DIVA website [14]). Nonetheless, the created agents still have a real problem in terms of acceptability: we suggest it is at least partially coming from the agent's lack of variability in its answers, which is a direct consequence of the lack of memory concerning previous interactions and of an advanced model of its state of mind.

Although many works have been undertaken on advanced cognitive agents, particularly when they are based on the traditional BDI approach [15][16], our purpose here is to explore the possibilities of improvements of the acceptability within the limits set by that kind of architecture; said differently: how far can we push the chatbot-like approach?

In this paper, we first present our supporting framework and the way we have been able to use it to introduce variability within agent's reactions (either random or dependent on previous interactions), and illustrate it through the example of a high level behaviour: the reaction of a female agent when it is asked its age.

[1] LIMSI-CNRS, Univ. Paris-Sud XI, BP 133, 91403 Orsay cedex, France. Email: {mxt,bouchet,jps}@limsi.fr.

In a second part, we define an experimental protocol to test the efficiency of the variability introduced, particularly in terms of believability and human-likeness. Finally, we give the results we have been able to obtain with that experiment and analyze them.

2 METHODOLOGY

2.1 The supporting framework

To organize controlled experiments where ordinary users (now accessible over the Internet) can interact with artefacts assisted by ACAs (e.g. to collect a corpus of natural language assistance requests, to register the users reactions...) has led us to develop a web-based toolkit called DIVA (DOM Integrated Virtual Agents) which can support virtual characters completely integrated within the DOM (Document Object Model) tree structure of web pages. Its two main objectives are:
1) To be an open programming framework, making it easy and quick to develop and deploy experimental ACA in web-based applications;
2) To take advantage of the new rich-client web 2.0 technologies to offer a full control of the interaction with the virtual characters (see figure 4 – more examples are available at [14]).

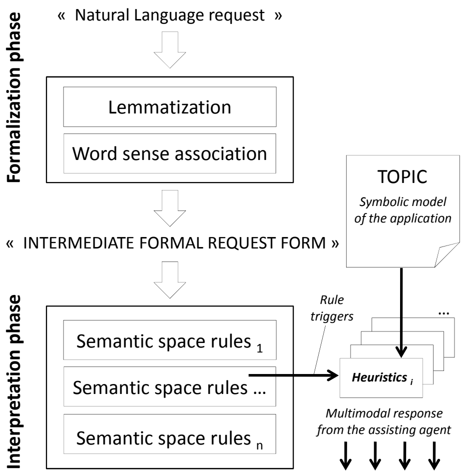

Figure 2: The DIVA NLP chain

The NLP-chain of the DIVA toolkit, sketched in figure 1, is detailed in figure 2. Like in most chatbots, the DIVA NLP-chain is based on pattern matching rules (we use a RegExp language) but it has a more sophisticated architecture, organized in two main phases with sub-phases:
1) The formalization phase: it is based on two sets of filtering rules applied in sequential order:
— Syntactical level: typical string pre-processing is followed by a lemmatization phase;
— Word-sense association level: lemmas are then transformed into semantic classes or 'synsets' as in Wordnet [17]; in case of ambiguity (multiple senses), a shallow WSD approach chooses the most likely one according to the collected corpus of requests.
At the end of the formalization phase, the request is transformed into an intermediate formal form, called the Formal Request Form (FRF).
2) The interpretation phase: it is based on a set of rules of the form <pattern → reaction> where patterns are applied to FRF expressions and reactions are procedural heuristics defining the behaviour of the agent in response to the user's requests.

Here are two examples of users' requests translated into FRF:
REQ$_1$ = "*If I want to buy such a car, what can I do?*"
FRF$_1$ = < QUEST IF THEUSER TOWANT TOOBTAIN such a car WHAT TOCAN THEUSER TODO >
The filtering process has extracted 9 synsets (uppercase symbols) from REQ$_1$ that are put in FRF$_1$. Some lemmas have no synsets because they are not in the generic ontology (e.g. 'car').
REQ$_2$ = "*Adopt a less provocative attitude, please.*"
FRF$_2$ = < TOTAKE a LESSTHAN ISUNPLEASANT THEBELIEF TOSAYPLEASE >
For the sake of simplicity, in the first version of DIVA, a primary requirement was to restrict the number of semantic classes to less than 500 (e.g. EuroWordnet has more than 10,000 [18], but it covers the whole NL whereas it has been shown our assistance domain represents only 1% of it).

The interpreting phase is organized into several layers, called 'semantic spaces' or in short 'spaces'. Most spaces are dedicated to a generic conversational domain, making them easier to share and reuse from an experiment to another. Each semantic space contains a set of rules that defines a behaviour of the agent. For example, assume that the user asks its age to the agent: "*How old are you?*" → <QUEST HOW ISOLD TOBE THEAVATAR>
We now have the following behavioural rule:

```
<rule id="age" pat="QUEST THEAGE|HOW ISOLD”>
  <do>
    THETOPIC.age.asked++;
    If (THETOPIC.age.asked >= 1)
       TALK_prepend([‘As I said’,'I’ve told you, ']);
    If (THETOPIC.gender = ‘female’)
       TALK.say(‘It’s not polite to ask this.’);
  </do>
 <say>
   <p>I’m _THETOPIC.age_. years old</p>
   <p>I’m _THETOPIC.age_ ...</p>
   <p>My age is _THETOPIC.age_"</p>
</say>
</rule>
```

The possibility to add several lines into the `<say>` tag introduces variability as one of the option shall be chosen randomly. It can use the meta-variable `_THETOPIC.age_` thus producing for example: "I'm 25 years old".
The `<do>` tag can contain some JavaScript and thus allows easy scripting. In this example, we take into account past interactions through the simple use of an additional property (asked) associated to each fact[2]. We also take into account a static fact: the gender of the agent.

We can see that to build a reaction the agent requires some kind of *knowledge base* registering the relevant assistance information about the application, but also about the agent's and user's profiles (e.g. to store the agent's age in the above example). In DIVA, the symbolic information about the assisted application is stored in its so-called *topic* XML-file. For example, here is an extract of the topic file of the agent used in the experiment:

```
<?xml version="1.0" encoding="utf-8"?>
<topic id="TOPICLEAAGE">
  <objName>Lea</objName>
  <objLanguage>English</objLanguage>
```

[2] Obviously, that detection of repetition being not specific to the age is normally handled independently of the property considered.

```
  <objCreators encoding="JS">["my parents","my father
Jack", "my mother Clarissa"]</objCreators>
  <objBirthdate>October 10th, 1983</objBirthdate>
  <objHeight unit="m">1.60</objHeight>
  <objPosition>England</objPosition>
  <objJob>lab assistant</objJob>
</topic>
```

The variable *height* can be referred to by: `THETOPIC.objHeight`

2.2 Experiment

2.2.1 Tested parameters and principle of the experiment

We remind that our objective is to check the impact of variability in an agent's answer upon the user's satisfaction. For this purpose, we have designed an experiment where an agent has seen its behaviour subtly modified according to three parameters:

- *Responsivity*: the fact that the agent accepts or not to give an answer to the question;
- *Variability*: the agent's ability, for a given question, to answer in more than a single way;
- *Dependence*: when an answer is varying, that variation can be linked or not to previous interaction(s).

Dependence happens only when we have variability, but variability can be seen both when an agent is answering and when it is not. It leads us to define six different scenarios as described on the decision tree on figure 3.

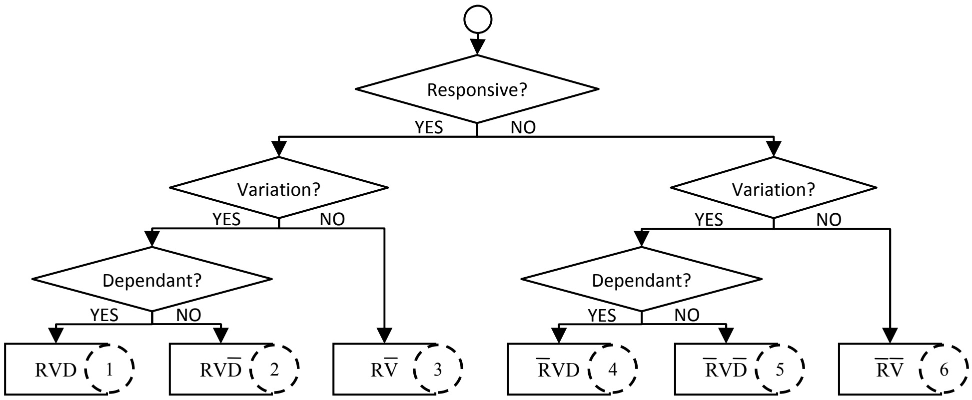

Figure 3: The six different scenarios represented as a decision tree, with their associated number and short notation (ex: RVD for the scenario 1 where the agent replies with a variation dependant on the previous answers already given)

For each case, we want first to let the user interact freely with the agent, and then to ask him to evaluate the behaviour according to parameters described in table 1.

Tested Parameter	Explanation and example
Precision	When asking for the time, the answer "17:02" is precise, "around 5pm" is not precise.
Relevance	When asking about musical tastes, the answer is about music, not sports or anything else.
Believability	When asking about the agent's gender, male is not believable considering her unambiguous feminine look.
Human-likeness	Answers are not obviously from a computer – same ones could come from a real human.
Variability	When asking several times the same question, the answer is always the same, it is not variable.
Cooperation	The agent is cooperative if it always provides the information requested in its answer. If it doesn't, even after repetitions, it is not cooperative.
Global satisfaction	The overall feeling about the agent's answers.

Table 1: Parameters evaluated in post-interaction questionnaires

We have decided to apply those behaviour variations to a single fact of the knowledge base: the age of the agent. That means the rest of the interactions shall be strictly the same in the six cases. That decision was justified by the fact it was a first evaluation of the phenomenon, and we thought changes in the way the agent is replying to every possible question would bias the measure of difference between interactions, making them too dissimilar. Table 2 illustrates the difference of answers in each scenario.

Case	First reply	Second reply	Third reply
1	I'm 25	I told you I'm 25	I won't answer to that again
2	I'm 25	I'm 25 years old	I'm 25
3	I'm 25	I'm 25	I'm 25
4	I won't tell you	I told you I won't tell you	Stop insisting!
5	I won't tell you	I will not tell you this	I won't tell you
6	I won't tell you	I won't tell you	I won't tell you

Table 2: Example of agent's reply to the question "how old are you?" in the 6 cases shown on figure 3

2.2.2 Protocol description and justification

In the written instructions, the subjects are explained that the purpose of the experiment is to interact successively with two different agents (whenever their embodiment is the same – cf. figure 4). The interaction is "natural", i.e. users were not following any explicitly scripted list of questions to ask. The general objective given is simply to get to know the agent by collecting basic information about it (its name, its age, its job…). The interaction is not time-constrained but is suggested to remain short (around 2 minutes). We also inform them that they shouldn't hesitate to insist or repeat questions.

After each interaction, the subjects are asked to fill a questionnaire to give their opinion about the agent, through an evaluation of their level of agreement on a 5-point Likert scale to an affirmation like "The agent's answers were relevant" (followed by an explanation and/or example of the parameter evaluated), for the 7 parameters in table 1. They can also leave a comment about each parameter if needed. Once all the interactions have taken place, subjects are finally asked to compare the agents they have been interacting with.

We have chosen to let each subject interact only with two agents, considering he would quickly lose attention and motivation if the experiment was too long (with the protocol above, it was already 20-30 minutes long) and forget the first interactions so the comparison might be less accurate. The other extreme could have been to let the user interact with only one type of agent, but there we would have had to face with individual differences in the way to rate the agent (an enthusiastic user being able to give a better mark to a given agent than the one a more critical user would give to another agent objectively more satisfying). Considering the scenario 1 (RVD) was a priori the best one, we tested it against all the other ones as shown in table 3 – changing the order of interaction also allowed us to take into account the potential problem related to the order of exposition.
The use of a final questionnaire where the user is explicitly asked to choose between the two agents is also a way to cross-validate his individual marking, and to possibly counter-balance too extremely positive/negative marks initially given to the first agent (but the first impression being important too, we couldn't have only that last questionnaire).
The choice of using the same embodiment for the two agents the user has to interact with was to prevent side effects linked not to the agent behaviour but to its appearance.

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1st interaction	1	1	1	1	1	2	3	4	5	6
2nd interaction	2	3	4	5	6	1	1	1	1	1

Table 3: the 10 test cases according to the two interactions

Due to our decision to implement the behaviour variation over a single parameter, we preferred to mention this parameter explicitly in the examples of questions to ask, not to have too many subjects missing it; they were however not explicitly asked to repeat those examples. We don't believe it introduced a too strong bias though, since the age is a question naturally asked in a first encounter chat, particularly on the web (cf. the ASL (Age/Sex/Location) phenomenon [19]). For the same reason, we emphasized the possibility to insist that might not have used naturally by every subject otherwise (either because they wouldn't insist in real life if an answer is not given or because they wouldn't expect an agent to change its answers). Those two hypotheses would ideally need to be checked, which could be done by using a larger pool of subjects (where the cost of having some of them not following the ideal path wouldn't jeopardize the results of the experiment). Finally, we also suggested users to keep interaction short, to prevent that difference from being swallowed up in questions about other facts.

Experiment has been done mainly over the Internet, emails sent to the participants containing links to the two agents they had to interact with and to the three online questionnaires. Subjects who passed the experiment next to an experimenter were not given any additional information than the ones from the email and used the same online system.

To prevent a potential bias linked to the fact subjects were not English native speakers whereas the ECA was in English, instructions, important words and examples in the questionnaires were translated in Chinese and French – subjects were also free to add their comments in any of those three languages.

2.2.3 Implementation on the DIVA website

The six agents created were set on visually identical web pages as the one shown on figure 4, where the key steps explained in the email were reminded in the background, to prevent the subject from having to change too regularly his interface.

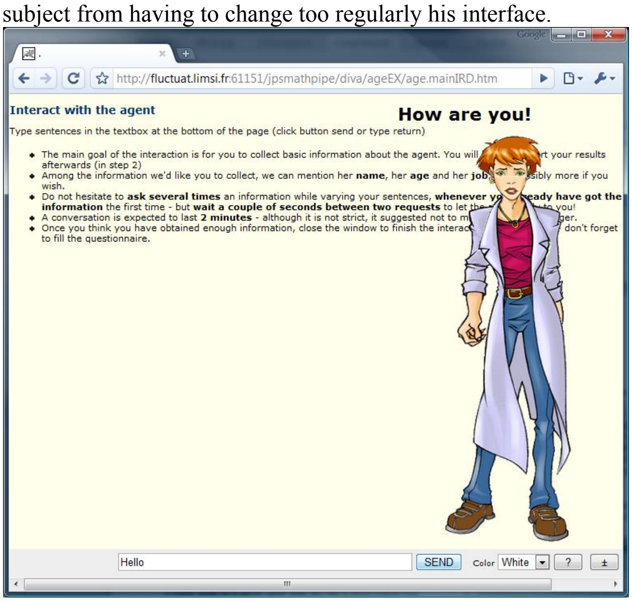

Figure 4: One of the six DIVA agents used for the experiment

3 RESULTS

3.1 Data

21 subjects have taken part in the experiment: 14 men and 7 women, all were between 20 and 60 (with a majority of 62% in the 26-30 age bracket) and were French or Chinese. Most of them (85%) had a university formation level but they had a disparate knowledge in computer science (when asked to rate their computer science level on a 1-5 scale, 42% were at 3 or below). For half of them, this experiment was their first interaction with a conversational agent. We have collected a total of 38 post-interaction questionnaires filled (4 subjects have skipped the second questionnaire despite recorded interactions with the agent), and 19 final questionnaires (2 skipped it).

Table 4 presents a synthetic comparison of the RVD agent compared to all the other ones, according to the post-experiment comparison questionnaire – table 5 is the detailed version for each of the 5 pairs tested (we haven't taken into account in this table the order of interaction).

Finally table 6 is made from the post-interaction ratings on a 5-point Likert scale of the 7 parameters in table 1. Means have been computed for each parameter in each of the six scenarios. After a Fisher test validating the hypothesis regarding the homogeneity of variances, means of scenario 2 to 6 have been compared to the ones obtained for the reference scenario 1, using a Student unilateral test with α=0.05 of the H_0 hypothesis: "$\mu_{\text{param},i} = \mu_{\text{param},1}$" (where i is in [2,6] and μ stands for the mean), with the relevant alternative hypothesis H_1: "$\mu_{\text{param},i} > \mu_{\text{param},1}$" or "$\mu_{\text{param},i} < \mu_{\text{param},1}$".

Agent preferred/ Parameter	**1 (RVD)**	**Other**	**None**
Precision	44.4%	5.6%	50%
Relevance	38.9%	22.2%	38.9%
Believability	44.4%	16.7%	38.9%
Human-likeness	33.3%	16.7%	50%
Satisfaction	50%	16.7%	33.3%

Table 4: Synthetic comparative results between the RVD agent and the others

Agent preferred/ Parameter	**1>2**	**1<2**	**1=2**	**1>3**	**1<3**	**1=3**	**1>4**	**1<4**	**1=4**	**1>5**	**1<5**	**1=5**	**1>6**	**1<6**	**1=6**
Precision	0	0	100	25	25	50	100	0	0	60	0	40	25	0	75
Relevance	0	0	100	25	25	50	100	0	0	40	20	40	25	50	25
Believability	50	0	50	25	25	50	66.7	0	33.3	40	20	40	50	25	25
Human-likeness	0	0	100	25	0	75	33.3	66.7	0	40	20	40	50	0	50
Satisfaction	50	0	50	25	0	75	66.7	33.3	0	60	20	20	50	25	25

Table 5: Comparative analysis of each scenario for each parameter
– boxes are in light gray when one of the agents is evaluated better than the other for that parameter

3.2 Analysis and discussion

As shown on table 4, for all the parameters tested, the RVD agent was judged to perform better than the other ones, particularly on the overall satisfaction. Although no score is above 50%, when the user noticed a difference between agents it was in favour of the RVD one (by a majority above 2 against 1). Table 5 offers a detailed analysis which lets appear that the RVD agent was also generally performing equally to or better than all the other agents considered individually. Globally, the difference also appears to be more obvious when compared to cases where the agent was not answering (cases 4, 5 and 6). When compared to the agent which was answering with a static answer (case 3), the agent 1 appears to be more human-like in its behaviour, and when compared to the agent which variations were random (case 2) it appears as more believable. Those results, although not striking, are supporting our initial hypothesis that an agent with a variation dependant on previous interaction is indeed perceived to be more human-like and believable, and thus confirms the interest of the framework modifications to handle them introduced in 2.1.

Nonetheless, some other results are less explainable a priori and might require further attention. For instance, the human-likeness of an agent not answering to the question but with dependant variability in its answers (case 4) is perceived to be higher: this is interesting and probably explainable by the chosen parameter for the experiment (the age), as not answering when asked its age for a woman could perceived as a sign of higher degree cognitive model. The fact that this human-likeness is not perceived when there is no dependant variability (cases 5 and 6) would let us suppose that the need for variability is even more important when the agent is *not* providing the expected answer. Indeed, not telling one's age willingly is a high level behaviour, and one can't expect it in an agent which is not even able to detect that a given question had been already asked several times.

Agent / Parameter	**1**	**2**	**3**	**4**	**5**	**6**
Precision	2.78 *1.06*	2.5 *2.12*	**3.5** *1*	2.2 *0.45*	1.8 *0.84*	2.25 *0.5*
Relevance	2.72 *1.22*	2 *0*	**3.25** *0.96*	2.4 *0.55*	3 *1.22*	2.25 *0.96*
Believability	3.39 *1.04*	**4.5** *0.71*	3.75 *0.5*	3.2 *1.48*	3.4 *1.34*	3.5 *0.58*
Human-likeness	2.72 *1.13*	3 *2.83*	**3.75** *0.5*	2.8 *1.30*	3 *1.22*	2.25 *1.26*
Variability	**3.06** *1.39*	3 *2.82*	3 *1.15*	2.2 *1.64*	3 *1.58*	2.25 *1.26*
Cooperation	**2.44** *1.25*	1.5 *0.71*	1.75 *0.96*	2.4 *1.95*	1.4 *0.55*	1.25 *0.5*
Satisfaction	2.83 *1.25*	2.5 *0.71*	**3.5** *1*	2.8 *1.10*	2.4 *1.34*	1.75 *0.96*
Number of subjects	18	2	4	5	5	4

Table 6: Mean (plain) and standard deviation (italic) from ratings given in the post-interaction questionnaire, for each of the six cases – boxes in light gray represent means statistically different from the reference case 1 (reject of H_0), means in bold are the best ones for the considered parameter.

Results obtained in table 6 are however harder to interpret, since it's the agent which was answering without variability at all that gets the best scores in terms of precision, relevance, human-likeness and even satisfaction. This phenomenon is clearly linked to the fact agent 1 must have been rated more poorly in some interactions, but more detailed explanations would certainly require an analysis of the interaction logs that hasn't been fully performed yet.

Indeed, some evaluations might have to be considered with a lower weighting (if considered at all) in cases where the user didn't ask the age several times (or didn't ask it at all) and hence was unable to notice the difference between both agents.

4 CONCLUSIONS AND PERSPECTIVES

We have seen that despite its technical limitations, the chatbot approach can be easily extended to introduce not only variability, but a variability dependent on parameters of the agent or of the application it is assisting (since it uses the same topic-based XML representation) and on the previous interactions, whenever the agent doesn't have any model of the dialogic session. We have confirmed through an experiment that the agents created with dependant variability are perceived to be more human-like and believable by a panel of users with various profiles. The results obtained also let us assume that the need for dependent variability is crucial if we want to be able to go further in the modelling of high level behaviour like the phenomenon in which a female agent is reluctant to give her age.

In a future work, it would be interesting to retry the same experiment using the six different behaviours on all the parameters (instead of only the age) to see if the differences are more noticeable (i.e. less cases when the user doesn't choose between the two agents) and if the users' preferences are confirmed when this behaviour is global. A further analysis of the collected logs might also be helpful to interpret some results. Finally, to go back to our problematic concerning the case of assisting conversational agents, the question of the need of such high level behaviour in a context of assistance where the agent is a priori expected to be always cooperative remains open for now.

5 REFERENCES

[1] J. Xiao, J. Stasko, and R. Catrambone, "An Empirical Study of the Effect of Agent Competence on User Performance and Perception," *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, New York, New York: IEEE Computer Society, 2004, pp. 178-185.

[2] R.G. Pérez Galluccio, "Humanizing CALL: The use of pedagogical agents as language tutors," *NERALLT 2006*, Cambridge, MA, USA: 2006.

[3] J.M. Carroll and M.B. Rosson, "Paradox of the active user," *Interfacing thought: cognitive aspects of human-computer interaction*, MIT Press, 1987, pp. 80-111.

[4] A. Capobianco and N. Carbonell, "Contextual Online Help: a Contribution to the Implementation of Universal Access," *Universal Access and Assistive Technology*, S. Keates, P.J. Clarkson, and P. Robinson, eds., London: Springer, 2002, pp. 131-140.

[5] J.F. Allen, L.K. Schubert, G. Ferguson, P. Heeman, C.H. Hwang, T. Kato, M. Light, N. Martin, B. Miller, M. Poesio, and D.R. Traum, "The TRAINS - project: a case study in building a conversational planning agent," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 7, 1995, p. 7.

[6] G. Ferguson and J.F. Allen, "TRIPS: an integrated intelligent problem-solving assistant," *AAAI'98/IAAI'98: Proc. of the 15th national/10th conf. on Artificial intell./Innovative applications of artificial intell*, Menlo Park, CA, USA: American Association for Artificial Intelligence, 1998, pp. 567-572.

[7] R.S. Wallace, "The Anatomy of A.L.I.C.E.," *Parsing the Turing Test*, 2008, pp. 181-210.

[8] "Alicebot," *http://alicebot.blogspot.com/*.

[9] "Elbot by Artificial Solutions," *http://www.elbot.com/*.

[10] J. Weizenbaum, "ELIZA: a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, Jan. 1966, pp. 36-45.

[11] C. Wollermann, "Evaluierung der linguistischen Fähigkeiten von Chatbots," Master's thesis, Rheinische-Friedrich-Wilhelms Universität Bonn, 2004.

[12] C. Wollermann, "Position paper," *Proceedings of Young Researchers' Roundtable on Spoken Dialogue Systems*, Pittsburgh, PA: 2006, pp. 75-76.

[13] F. Bouchet and J-P. Sansonnet, "Caractérisation de Requêtes d'Assistance à partir de corpus," *Actes de MFI'07*, Paris, France: 2007.

[14] "DIVA - DOM Integrated Virtual Agent," *http://www.limsi.fr/~jps/online/diva/divahome/index.html*.

[15] R. Evertsz, F.E. Ritter, P. Busetta, and M. Pedrotti, "Realistic Behaviour Variation in a BDI-based Cognitive Architecture," *Proc. of SimTecT'08*, Melbourne, Australia: 2008.

[16] D. Pereira, E. Oliveira, and N. Moreira, "Formal modelling of emotions in BDI agents," *Proceedings of CLIMA-VIII*, F. Sadri and K. Satoh, eds., Porto, Portugal: Springer, 2008, pp. 62-81.

[17] C. Fellbaum, *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press, 1998.

[18] P. Vossen, ed., *EuroWordNet: a multilingual database with lexical semantic networks*, Kluwer Academic Publishers, 1998.

[19] G. Merchant, "Teenagers in cyberspace: an investigation of language use and language change in internet chatrooms," *Journal of Research in Reading*, vol. 24, Oct. 2001, pp. 293-306.

6 APPENDIX

Transcript of an interaction with a RVD agent (scenario 1)

Agent: Hello welcome to the experiment.
You must type your personal ID number (like 12345) in the bottom left textfield and click ok.
User: 580
A: Your ID is 580, you can begin now.
U: hi
A: Yes, can I help you?
U: yes
A: Erm...
U: what's your name
A: Call me Lea.
U: where do you live?
A: I don't want to talk about it.
U: introduce yourself
A: Call me Lea.
U: how old?
A: 25 years old
U: how old are you?
A: 25 years old, remember?
U: yes
A: Let's talk about something else.
U: where do you work?
A: Erm...
U: What's your job ?
A: I'm currently working as a lab assistant.
U: do you speak german ?
A: I speak English.
U: do you speak german ?
A: I speak English.
U: you're stupid
A: My speed is strange question.
U: don't do that
A: Ok I won't do it.