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

Mao Xuetao
François Bouchet
Jean-Paul Sansonnet

LIMSI-CNRS,
Université Paris-Sud XI

{xuetao, bouchet, jps}@limsi.fr

AISB 2009

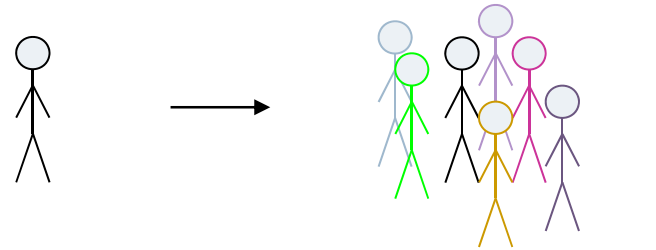
April 7th 2009

Outline

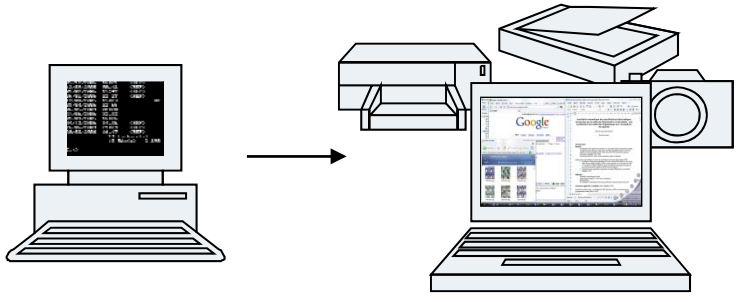
- Context: assisting novice users with ECA
  - The increasing need for assistance
  - Assisting novice users with ECA
  - Help systems comparison
  - Dialogue system or chatbots?
  - Key issues
- Methodology
- Results
- Conclusion

The increasing need for assistance

- **Users** evolution:
  - In number:
    600 millions (2002) →
    2 billions (2015 – projection)
  - In variety:
    from computer scientists to everyone



- **Hardware** evolution (Moore's law):
  - Application fields
  - Interaction fields



- **Software** evolution:
  - More numerous
  - More complex: in public applications
    150 « basic » actions (in menus);
    60 dialogue boxes ;
    80 tools (through icons).
    (Beaudoin-Lafon, 1997)

Assisting novice users with ECA

- **Assisting**: *« An Assisting Agent is a software tool with the capacity to resolve help requests, issuing from novice users, about the static structure and the dynamic functioning of software components or services »* (Maes, 1994)

- **Conversational**: interaction in unconstrained natural language (NL)
  Why?
  Frustrated (novice) users spontaneously express use NL
  (⇔« thinking aloud effect » (Ummelen & Neutelings, 2000))

- **Embodied**: given a graphical more or less realistic appearance
  Why?
  Increased agreeability and believability – « Persona Effect » (Lester, 1997)

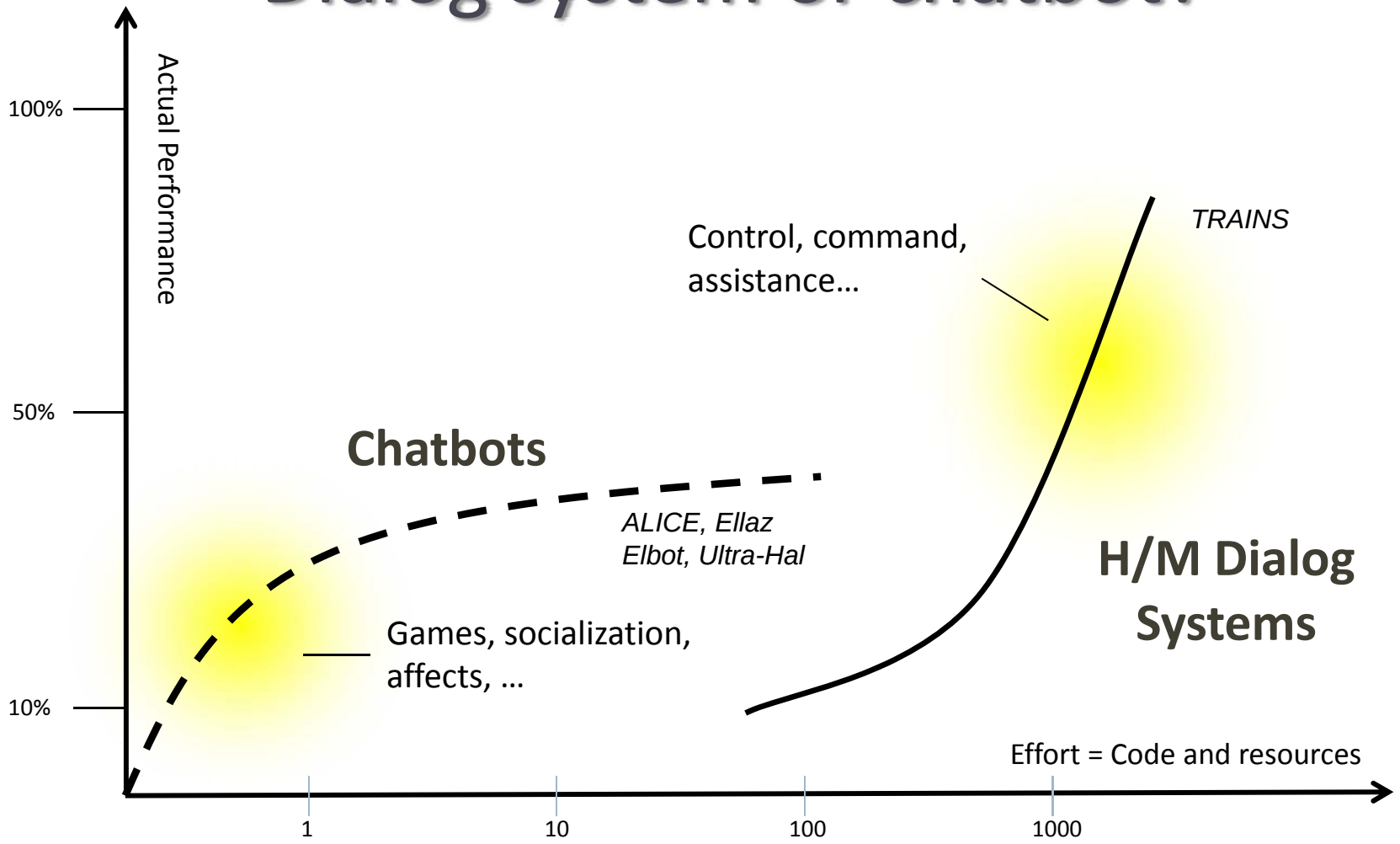
Help systems comparison

Help system	Reactivity	Vocabulary	Task-oriented	Dynamic	Personalized	Proactive
Paper documentation	-	-	-	-	-	-
Electronic documentation	+	-	-	-	-	-
FAQ, How-to, Tutorial	+	=	+	-	-	-
Contextual Help Systems	+	=	=	+	-	-
Assisting Conversational Agent	+	+	+	+	+	=

- **Reactivity**: how fast is it for the user to open the help system when it needs it?
- **Vocabulary**: are there strong constraints or limitations on the words the user has to know to efficiently use the help system? (ex: specific keywords/grammar constructions for NL)
- **Task-oriented**: does the help system explain procedures and not only define concepts?
- **Dynamic**: does the help system change according to the application state?
- **Personalized**: does the help system change according to the user?
- **Proactive**: does the help system appear only when asked for or can it anticipate the user needs (without being intrusive)?

Conclusion**:** Assisting conversational agents *potentially* seem to be the most efficient way to help novice users.

Dialog system or chatbot?



Chatbots are limited in terms of **genericity** (need to rebuild everytime) (Allen, 1995) and **linguistically** (Wollermann, 2006) – but how far can we push the approach?

Dialog system or chatbot?

- **Advantages:** easy, light, precise
  - They are **easy to develop**: no large semantic analyzer, no complex reasoning tools;
  - They are **light to deploy** in a web-based environment → client architectures can be envisioned;
  - They provide **robust** natural language reactions (Evasive list effect – ELIZA effect);
  - They are tailored and well-suited for the field of **ludo-social** chat;
  - When associated with a given application, they **can** be customized to **be extremely precise**.

- **Drawbacks:** lack of genericity, linguistical limits
  - Minimalistic/**ultra-customized** model of the application;
  - **Minimalistic model** of the dialogue session and of the users;
  - **No semantic analyzer** → lack of precision in the requests (grammar, speech acts, …);
  - **No formal requests** → class reactions are directly linked to specific linguistics patterns;
  - **No generic reasoning tools**, especially when the function of assistance is concerned.

Key issues

Hypothesis: variability improves user's perception of the ECA

1. Technical feasability: is it **possible** to handle variability with a chatbot architecture?

2. Need: do people **notice** variability?

3. Effect: does it **affect** the perception users have of the agent? And if yes, how?

4. Can it be **useful** for assistance?

Outline reminder

- Context: assisting novice users with ECA

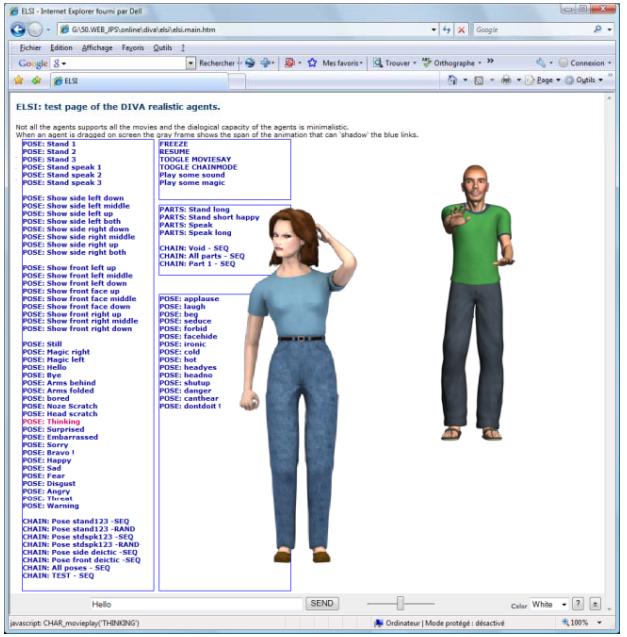
- Methodology
  - Experimental framework: DIVA framework overview
  - Experimental framework: DIVA NLP-chain
  - Experiment principles
  - Experimental protocol
  - Questionnaires

- Results

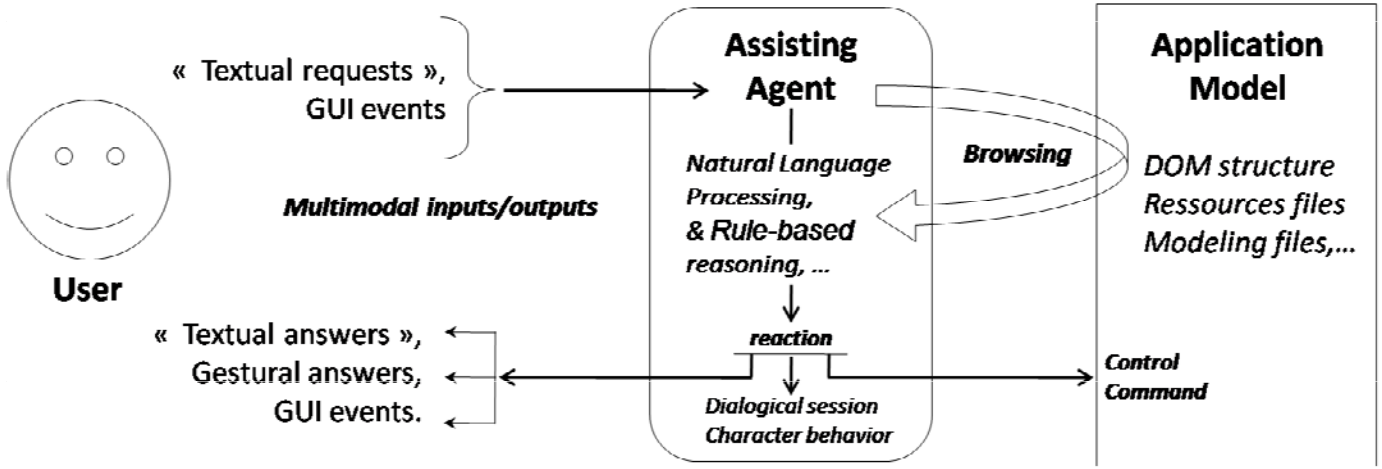
- Conclusion

DIVA framework overview

- DOM Integrated Virtual Agent:
  - Open programming framework
  - High level of interaction (AJAX)



1. Embodied Agents Elsi & Cyril:
2. Natural Language Processing chain:



Experimental framework: DIVA NLP-chain

1. Formalization phase

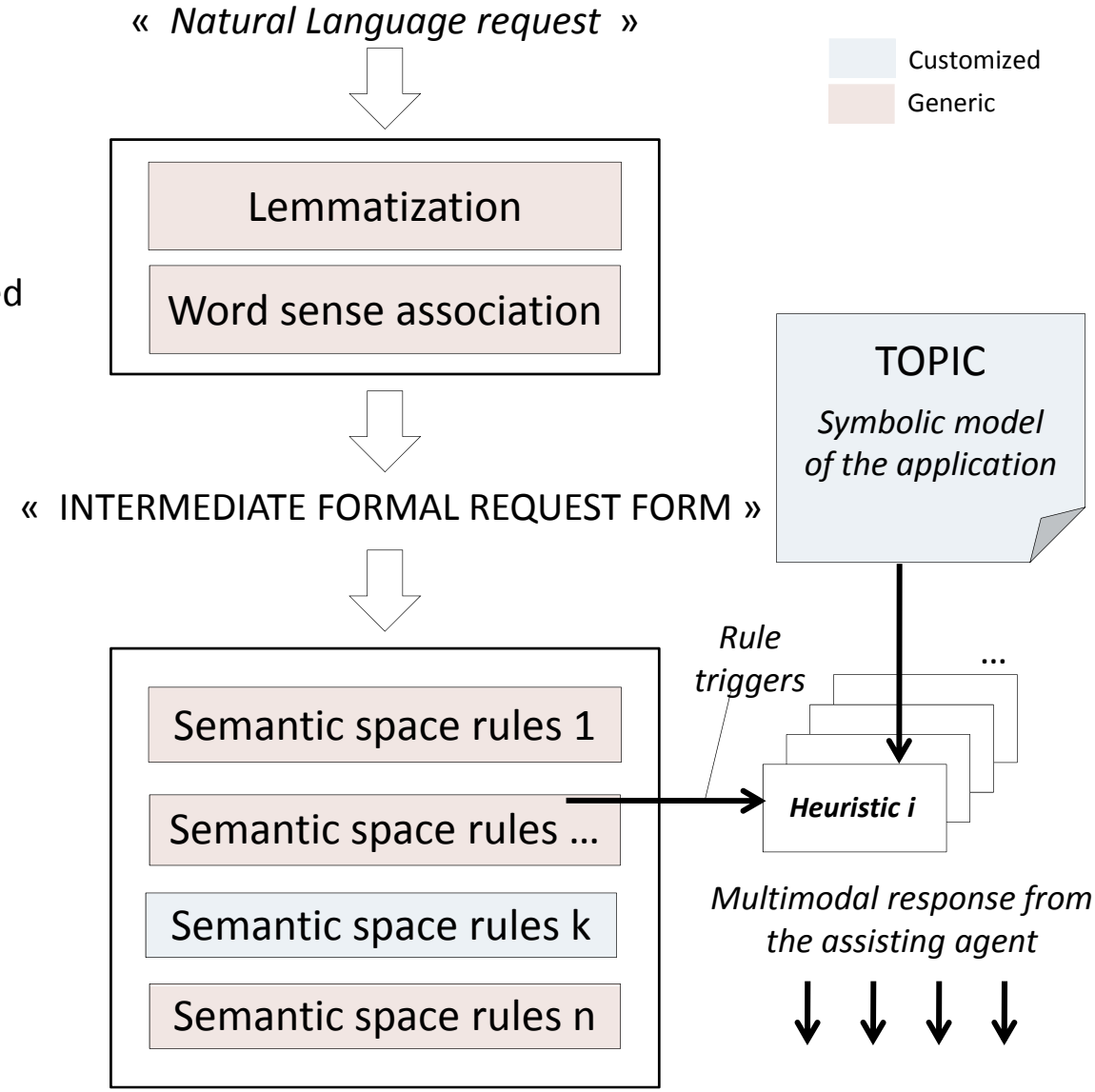
1. Sentences are preprocessed and words are lemmatized;
2. A semantic class (KEY) is associated with each word

2. Interpretation phase

Interpretation rules are of the form:

Pattern → Reaction

Where reactions are expressed as procedural heuristics achieving reasoning tasks over the description of the application (the topic file).



Experimental framework: DIVA NLP-chain

« How old are you? »

DIVA:

1) **Formalization**:

```
<QUEST HOW ISOLD TOBE THEAVATAR>
```

2) **Interpretation**:

```
<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>
```

variability

genericity

Classical chatbots (ALICE – AIML):

```
<category>
  <pattern>HOW OLD ARE YOU</pattern>
  <template>
    <set_it>I</set_it> am 25 years old
  </template>
 </category>
```

1. Matches a user input containing the **exact pattern**
2. Handles a **minimalistic** model of the session (IT)
3. Sends an entirely **predefined** answer

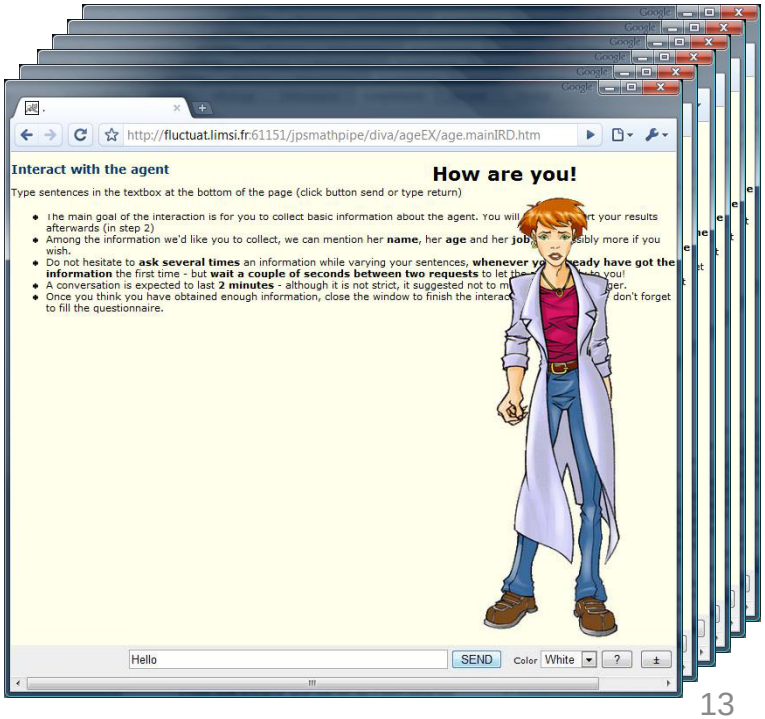
Experiment principles (1)

- Three (linked) parameters actually tested:
  - **Responsivity**: the requested information is in the answer
  - **Variability**: twice the same question can lead to different answers
  - **Dependence**: variability with a memory of previous questions

- Differences: one only answer when requested its **age**.

- 6 female agents, visually identical

- Interaction through chatbox at the bottom of the window



Experiment principles (2)

« How old are you? »

	Responsive	Variable	Dependent	1st reply	2nd reply	3rd reply
1	✓	✓	✓	I'm 25	I told you I'm 25	I won't answer to that again
2	✓	✓	✗	I'm 25	25 years old	I'm 25 years old
3	✓	✗	-	I'm 25	I'm 25	I'm 25
4	✗	✓	✓	I won't tell you	I said I won't tell you this	Stop insisting!
5	✗	✓	✗	I won't tell you	It's a secret	I will not tell you
6	✗	✗	-	I won't tell you	I won't tell you	I won't tell you

Experimental protocol

- User's objective: retrieving information about an agent
  - Free chat
  - Suggestions:
    - Examples given: name, age, job…
    - Short interaction (< 2 minutes)

- Interaction with **two agents**:
  - Case 1 or Case [2..6]
  - Case [2..6] or Case 1

- Three **questionnaires**:
  - One after each interaction (5-point Likert scales)
  - Final comparative questionnaire

Questionnaires

- 7 parameters evaluated:
  - **Variability**: not always answering the same way ⇔ noticing variability
  - **Cooperation**: if information requested could be obtained ⇔ noticing responsiveness

  Only after interaction

  - **Precision**: « 25 years old » / « young »
  - **Relevance**: the agent remains in the topic of conversation
  - **Believability**: the agent being a female is believable
  - Human-likeness: same answer could come from a human being
  - Global satisfaction: overall feeling about conversation

Outline reminder

- Context: assisting novice users with ECA
- Methodology
- Results
  - Raw results
  - Comparative questionnaire results
  - Post-interaction questionnaire results
- Conclusion

Raw results

- 21 subjects, over the internet
  - Sex: 14 men / 7 women
  - Age: 20-60 (62% in 26-30)
  - Origin: Chinese/French mainly
  - Studies: university level (85%)
  - Computer science knowledge: disparate (42% below 3/5)
- 38 post-interaction questionnaires
- 19 final questionnaires

Comparative questionnaire results

- Globally: 1 vs all
  if a difference is made, **1 is prefered**, for every parameter

- Individually: 1 vs [2-6]
  if a difference is made, **1 is prefered**, except:
  - 4 (¬RVD) is perceived as more human-like
  - 6 (¬R¬V) is perceived as more relevant

- Discussion:
  - Not giving the age of a woman is not problematic: parameters interdependancy
  - Variability is even more crucial in that case (4 vs 5-6): expectation of a high level behavior

Post-interaction questionnaire results

- Sample too small to obtain many statistically significant results
- Many expected results:
  - Satisfaction: RVD > ¬R¬V
  - Cooperation: RVD > 5, RVD > ¬R¬V
  - Precision: RVD > ¬RVD, RVD > ¬R¬V
- Some unexpected ones:
  - Precision: RVD < R¬V
  - Believability: RVD < RV¬D
  - Human-likeness: RVD < R¬V
- Discussion:
  - Variability can make the agent look more imprecise
  - If the rest of the behavior doesn't follow, it is interpreted as mistakes

Conclusion

- **Possibility** to handle variability with a chatbot architecture
- Users **notice** variability in agents
- Agents with variability are perceived as:
  - more believable,
  - more human-like…

  …but **coherence** is crucial!

- Can it be useful for assistance?
  - Indirectly yes:
    - chat is important (~40%) even for assisting agents only *(Bouchet&Sansonnet, 2007)*
    - improved user's satisfaction
    - reduced « motivational paradox » *(Carroll&Rosson, 1987)*
  - Directly? Upcoming experiment
- Variant: behaviours affecting every parameter
- Study of parameters influence on each other (ex: gender/age)