This paper was drafted in 2003. It is clearly dated in some of its contents. But I still think that the phenomenon described is interesting, and I also think that the view of "action" and "interaction" as two fundamentally distinct conceptual or even ontological categories has some merit in for instance discussions of the so-called Uncanney Valley phenomenon. All comments most welcome.

The 'skyrocketing expectations' phenomenon in dialogue interfaces, or why 'action' should not be called 'interaction'

Nils Dahlbäck

Department of Computer and Information Science,
Linköping University, SE-581 83 Linköping, Sweden

nils.dahlback@liu.se

Abstract: This paper presents a hypothesis concerning preferred ways of using computers called the *action-interaction* hypothesis, and which states that we as human beings prefer computer interfaces that are clear-cut equivalents of physical manipulation in the physical world, or dialogue based interaction with other intentional agents; the middle ground confuses the users, triggers unexpected user behaviour, and makes the users less comfortable with using the systems. The hypothesis is primarily based on own and others non-published experiences when developing natural-language dialogue systems, but some published observations are presented as support of the suggested hypothesis. In the final section some suggestions for the design of speech-only and multi-modal computer interfaces with dialogue modules are presented.

Keywords: style, guide, HCI, Zurich

1 Introduction

In this paper I will not present any new research, in the sense of descriptions of new modes of interaction with computers, or presentations of evaluations or other kinds of empirical studies. Instead I want to present a phenomenon that I previously only heard described in informal conversations between especially designers and developers of natural language dialogue interfaces, and which I believe is worthy of some wider recognition. The paper is structured as follows. First I describe the phenomena as such. In the next part a possible explanation for it is put forth, and in the final section I suggest some consequences for the design of multimodal dialogue systems.

What I will present here is based on my experience form working with research on natural language dialogue systems since the mid-80'ies (e.g. Dahlbäck & Jönsson, 1988, 2003), both own observations and discussions with fellow workers in the field. But what I present here is more of a hypothesis that I have formed based on this experience. And it is presented here in just that spirit of a possible hypothesis interesting enough to be corroborated or falsified by making it known to a larger audience.

2 Skyrocketing expectations

One well-known difficulty when developing natural language dialogue systems, both speech only and multimodal ones, is to make the users understand the limits of the system's interactive and communicative abilities, to ensure that they will not use a language beyond the system's capacity. Guidelines for such designs often suggest that for instance, the linguistic output from the system should not include utterances that the system could not interpret as input, (though it has been shown that in some cases this is not a necessary requirement, Dahlbäck, 1991).

If the linguistic capabilities of the system are simple enough, making the user stay within the limits of the system's linguistic power is usually not difficult. Examples of this are simple speech systems over the telephone, where the user is given a short menu of alternatives, like e.g. "If you want to enquire about departure times, say 'departure', if you want to enquire about prizes, say 'prizes', and

…", or "Please say from where you want to depart". In cases like these, users rarely try to engage in any real dialogue, but instead often restrict the input to words or short utterances. So in this respect these systems work fine (even though the user appreciation of them are often somewhat limited to put it mildly).

And also for systems with more advanced natural language dialogue capacity, users often stay within the limits of the systems capacity. But once the system gets a little bit better in its conversational abilities, there seem to be no limit to what the users expect to be able to say to it!

This is something which up till now has primarily been seen in Wizard of Oz-studies (Dahlbäck, Jönsson, Ahrenberg, 1992). I have seen this happen myself, or heard about this from workers on natural language dialogue interfaces for mobile robots, for advisory systems and a host of other applications. But when discussing this with the developers, the common reaction is that this is due to some shortcomings of the design, or that they in some other respects have not been able to communicate to the users the limits of the system's powers.

It is sometimes also claimed that this occurs because the users are not familiar with the limits of the technology, and that the problems with users not being able to stay within the limits of the systems range of linguistic and communicative abilities will diminish or vanish with more experience with them.

I certainly believe that there is some merit to this explanation, but I am not convinced that this is the only thing that is happening here, and that the only reason for these skyrocketing expectations is a lack of experience on part of the users, and faults in the designers' work with the system. As an, admittedly speculative hypothesis, I would like to put forth another complementary explanation, which if true I believe has some consequences for future design work on systems of this kind.

3 Action vs. interaction

My argument is based on a distinction between two modes of 'being-in-the-world', acting and interacting. As biological creatures we have not developed to fit in and adapt to today's complex technological society. Before the advent of modern technology, we lived in a world that comprised of two basic ontological entities, physical objects, either natural or man-made, and other intentional agents, the prototypical case being other human beings. We lived, in a sense, in two different worlds, a physical and a social. In the physical world we *act* on objects; we move them, we transform them, we even destroy and digest them. In the social world we *interact* with other people. On a basic biological and even ontological level these are very different kinds of activities. Perhaps the most important is that in the social world, the others are also acting agents, which not only can move on their own, but which also have their experience of the world, including their experience of me. It is only with these agents that we can communicate, in the strict sense of the word. The word 'communication' is, after all, derived from the Latin 'communis', and to communicate means then more or less 'making (something) common'. So when moving a stone I am not communicating, nor am I interacting. I am acting. (But what is then human-computer interaction using a direct manipulation interface? I will get back to this issue below.)

Of course, the distinction is not as clear cut and simple as I have sketched it here. The two categories are graded or radial, as Lakoff (1987) and others would claim is true for all natural categories. The prototype here is interaction and communication between two grown up adults sharing the same cultural background. Children are different, as are animals. But note also that the further we move from these clear cut or prototypical categories of dead matter and grown up adults, the more uncertain do we become in whether we should treat what we encounter as belonging to either of these two kinds of entities. And even if we have a common understanding within a culturally homogenous group on particular cases, others might have very different opinions on this, as can be seen in the very heated philosophical and political discussion on issues like e.g. abortions or animals' rights.

4 Any empirical support?

As I stated in the beginning of this paper, the action-interaction distinction presented here should rather be seen as a hypothesis than a conclusion drawn from a host of empirical evidence. But I do believe that some research results give some support for the hypothesis. Let me here just mention some of these.

First, in their well known work on social responses to media, Reeves and Nass (1996) have shown that users of interactive media, and especially interactive computer systems, reproduce the same kinds of social responses to the system that we do with other people. We are polite to computers, prefer them to have personalities similar to our own, etc. But note that the participants in these studies are using software that makes them engage in something which is a close approximation of interaction, in the

sense used in this paper. The systems studied give advice or recommendations, comments on suggestions from the user etc. This is very different from the basic physical manipulation-like activities performed in a GUI. And the only attempt I am familiar with that tried to reproduce results from Reeves and Nass research, in this case personality similarity-attraction, on a pure graphical interface (Karsvall, 2002) failed to obtain any such effects (though it should be admitted that the limited size of the evaluation in this case should lead us not to overstate the importance of this observation).

Reeves and Nass also stress that the users' responses observed by them is *not* because the users do not understand the difference between computers and people; it is not some kind of anthropomorphism. The users reproduce these social responses even though they intellectually know that the computer is not a person.

An interpretation of the results from Reeves and Nass' research using the framework presented in this paper, would say that since the computer is more like an intentional agent than a physical object in the cases studied by Reeves and Nass, the users spontaneously reproduce prototypical reactions from interacting with social intentional agents. Since we have no natural categories between the objects of the physical world and the intentional agents in the social world, we jump for the category most appropriate to the experience encountered and reproduce behaviour patterns appropriate for this.

And my claim is of course that this is exactly what happens when the users' expectations of natural language dialogue systems skyrocket too.

My other example comes from a recently defended thesis at Linköping University (Qvarfordt, 2003). In this work users' subjective experience when using a multimodal interface was studied under three conditions, no spoken feed-back, limited spoken feed-back, and elaborated human-like feed-back (the basic system was implemented, but some parts of the spoken interaction was simulated using a Wizard-of-Oz method). The evaluation on a number of parameters such as control, cooperation, habitability etc, showed what Qvarfordt calls an "all or nothing" attitude. Users preferred the no speech feed-back and the full human-like speech feed-back, whereas the limited spoken feed-back was less well perceived.

This to my mind again supports the action-interaction distinction; the users prefer to use the computer in a mode which is pure action, like acting in a physical world, or a mode which closely resembles interaction with human, i.e. interacting in a social world. The middle ground is avoided.

As a final comment in this section, let me point out that the present hypothesis gives an additional explanation (or, perhaps better, a re-formulation of the explanation) for the success of graphical user interfaces over the typed interfaces when they first appeared with the early Xerox and Apple direct manipulation interfaces. The claim would simply be that this design avoided making the user uncertain on what could and could not be done when working with the computer. The typed interface did not trigger any clear cut expectations of what could and could not be done.

5 Consequences for the design of multi-media interfaces?

The hypothesis presented in this paper suggests that users prefer using computers when they as clearly as possible resemble either acting in a physical world or interacting with other intentional agents in a social world. Hybrid or less clear cut designs or modes of using the computer are less well received.

A corollary of this view is that it is not advised to expand the features of a well-functioning multi-modal interface with some additional natural language dialogue interfaces. This is likely to confuse the users of what they can and cannot do, something which was clearly seen in a recent walk-up-and-use demonstration of a multi-modal dialogue system (BirdQuest, Jönsson & Merkel, 2003) at our lab. Instead a clear separation should be made between the dialogue part of the system and 'standard' GUI part.

In fact, just this conclusion is also drawn by Ibrahim and Johansson (2002) in their user evaluation of a multimodal dialogue system for interactive TV applications, where they claim that the speech interactive module requires a clear separation between the dialogue system and the visual output. Their suggestion is an interaction model consisting on not two but three basic entities; the user, the dialogue component and the visual presentation (though I have here taken the liberty of translating their conclusions into a terminology more in accordance with the one used in this paper).

What makes this work especially interesting is that it was *not* concerned with a computer interface, but with a TV application. This makes it less probable that the users' preferences were a reflection of their expectations of previous encounters with similar systems without a speech interactive module (as would have been the case if the study was on a computer interface), since interactive TV is not something which have been around long enough to foster clear expectations of how they 'usually' work.

Another suggestion emerging from the hypothesis presented in this paper is that NL-dialogue systems, and especially multi-modal such systems (in distinction to speech only systems) should present some agent or avatar that represents the interacting agent. But, on the other hand, to set the level of expectations right, it is perhaps better to not have a picture of a grown-up adult as a form for this representation.

My reason for believing that this is perhaps less necessary for speech-only systems is that most of us by now have some experience of similar situations, since most of us have had to leave messages on answering machines. But note also that most people find it awkward to leave voice messages to machines, perhaps giving some additional support to the hypothesis presented in this paper.

6 Summing up

I have in this paper presented a response pattern observed by me and colleagues of naïve users of systems which enables them to interact with computers using natural language dialogues, in either multi-modal or speech only interfaces, and where they at some point up the ladder of improved dialogue capability in the system suddenly seem to expect it to handle full-blown dialogues, after previously having been very careful to adapt their language to the (assumed) limits of the computers ability to understand natural language.

The basic hypothesis is that we as human beings are biologically developed to live in a world where we either move around in a physical space or communicate with other intentional agents. We either *act* or *interact*. And while present day computer interface technology does not restrict us to develop similes of these two categories, given our tendencies to prefer clear-cut cases of either action or interaction, the hypothesis states that this is something which should be avoided.

Most of the inspiration leading up to this hypothesis stems from conversations with friends and colleagues sharing the kinds of experiences from user studies that never seem to find their way into scientific publications. But also from some corroborative observations from published work were presented.

But, as I stated in the first part of this paper, what I have presented here is not any conclusive result. It is a tentative hypothesis, which to my mind both has the potential to explain some observed regularities in users' reactions to the experience of interacting with computers using also natural language and suggests some issues to consider in future development of such systems.

I am convinced that most readers have noticed a problem with the view presented here. The suggested *action-interaction* hypothesis also suggests that the common parlance of our field is less appropriate. We talk about human-computer 'interaction' with graphical user interfaces, and those interfaces also have 'dialogue boxes'. If my hypothesis is correct, this could be seen as a problem. But I have no ambitions or hope to change this (nor do I have any constructive suggestions for alternatives). My ambition has been more modest than that; to present a hypothesis which I find interesting enough to be worthy of scrutiny and possible refutation form a larger audience. I look forward to hearing your reactions to this.

References

Dahlbäck, N. (1991) *Representations of Discourse*. Ph.D. thesis. Linköping University

Dahlbäck, N. & Jönsson, A. (1988), Talking to a Computer is not Like Talking to Your Best Friend, *Proceedings of The first Scandinivian Conference on Artificial Intelligence*, Tromsø, Norway, March 9-11, 1988.

Dahlbäck & Jönsson Experiences with and lessons learned from working with a modular natural language architecture. *Proceedings of HCII2003*, Crete, Greece, June 2003.

Dahlbäck, N:, Jönsson, A. & Ahrenberg, L. (1998) Wizard of Oz Studies -- Why and How, In M. Maybury and W. Wahlster (1998) *Readings in Intelligent User Interfaces* Morgan Kaufmann Publishers.

Ibrahim and Johansson (2002) Multimodal dialogue systems for interactive TV applications. In Proceedings of ICMI'02, Pittsburgh, PA, USA.

Jönsson & Merkel 2003 Some issues in Dialogue-Based Question-Answering, Arne Jönsson and Magnus Merkel, *Working Notes from AAAI Spring Symposium,* Stanford, 2003.

Karsvall 2002 Pershonality preferences in graphical interface design. *Proceedings of NordiCHI 2002*, Aarhus, Denmark.

Lakoff (1987) *Women, Fire and Dangerous Things*. Chicago: Chicago University Press.

Reeves & Naess (1996) *The Media Equation*. Cambridge: Cambridge University Press.5