Human–Computer Communication
Wizard Of Oz–Experiments In SmartKom

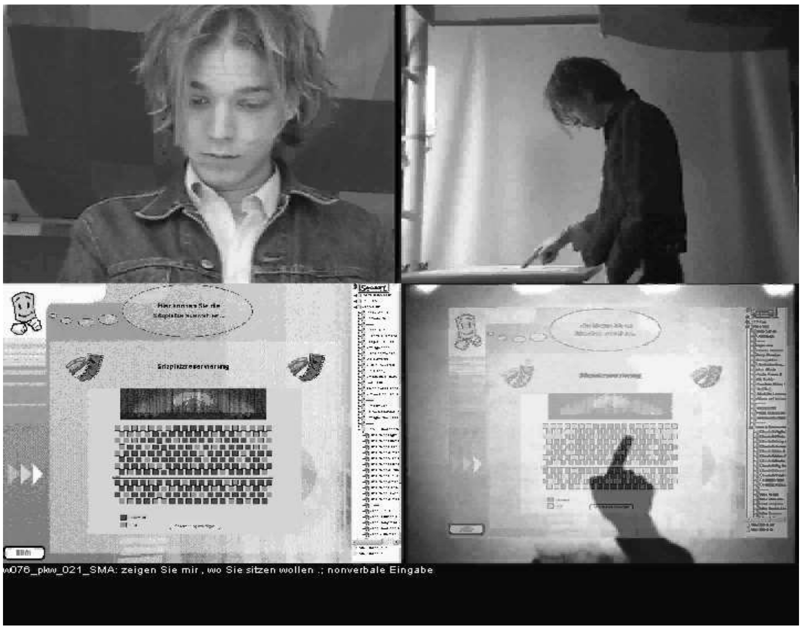
Silke Steininger
Institute of Phonetics and Speech Communication,
Schellingstr. 3, 80799 Munich, Germany

The SmartKom Project

The goal of the SmartKom project [1] is the development of an intelligent computer–user interface that allows a computer novice as well as an expert to communicate almost naturally with an adaptive and self–explanatory machine. The system does not only allow input in the form of natural speech but also in the form of gestures. Additionally the emotional state of the user is analyzed via her facial expression and prosody of speech. The output of the system comprises a graphic user interface and synthesized language. The graphic output is realized as a computer screen that is projected onto a graph tablet. [2]

To explore how users interact with a machine, data is collected in so–called "Wizard–of–Oz experiments" [3], [4], [5]. How these experiments are implemented is explained below. In each Wizard–of–Oz session spontaneous speech, facial expression and gestures of the subjects are recorded. The speech is transliterated, interesting user–states (anger, joy etc.) and gestures are labeled. The recordings together with the transliterations and labeling are distributed via DVD to the project partners. They need the data to train speech, user–state and gesture recognizers and to learn how humans interact with a multimodal dialogue system.

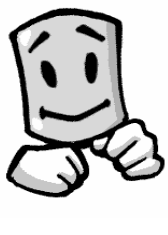
What Is A „Wizard–of–Oz Experiment"?



Goal of the experiments is the creation of a convincing simulation of a system that does not yet exist and to gather data of subjects that interact as naturally as possible with the simulation. In our Wizard–of–Oz experiments, the subjects have to solve certain tasks with the help of the system (like planning a trip to the cinema). They are made to believe that the system they interact with is already fully functional. Actually, many functions are only simulated by two "wizards", human beings who control the system from another room.

Our goal is to make the system as realistic as possible. In order to fool the subjects, we used the following methods:

– When designing the **display,** care was taken to give it a look different than internet sites, well known Windows applications or other familiar computer software. For example, there are no underlined links and there are new forms of objects the users could interact with. The display–pages were generated using HTML. One of the wizards presents the sites via a navigation bar that is hidden from the view of the subjects. [6]



– The information is presented by a **web persona** which can show four different states (awake, thinking, asleep and k.o.). It's name is „Aladdin". Aladdin greets the users, makes suggestions, offers help and information and apologizes if he doesn't understand the user.
– The **speech output** is spoken live by one of the wizards. The wizards are trained extensively to be able to speak monotonous and to avoid lapses. Additionally, the voice is distorted to give it a more machine–like character. The spoken

sentences are chosen by the wizard from a pre−generated list, but improvisation is allowed.

- For each **application** („cinema“ and „restaurant“) a flowchart was generated which shows how the users can navigate through the „system“. If they reach the limit of an application, they get the message „Sorry, this function is currently not available”.
- The wizards do not simulate a perfect performance. Complex sentences or confusing input lead to error messages or a wrong reaction by the „system“. It has to be noted that the **errors** that are simulated by the wizards are probably quite different from the ones the real system will produce.

Example of a session:

User: hallo Aladdin. ich bin zum ersten Mal in Heidelberg und m"ochte heut’ abend gern ins Kino gehen. k"onntest du mir

SMARTKOM: HIER SEHEN SIE EINE "UBERSICHT "UBER DAS PROGRAMM DER HEIDELBERGER KINOS.

User: so, nich’ dr"ucken.

SMARTKOM: ICH HABE SIE NICHT VERSTANDEN.

User: sind das alle Filme oder gibt es noch mehrere?

SMARTKOM: W"AHLEN SIE EINEN FILM.

User: ja, aber davon interessiert mich jetz’ kein Film. ich w"urde gern Lola+rennt sehen. l"auft der auch?

SMARTKOM: DIESE FUNKTION IST LEIDER NICHT VERF"UGBAR.

User: es is’ nich’ verf"ugbar, ob kein anderer Film l"auft oder ob jetz’ explizit der Film nich’ l"auft? Aladdin.

SMARTKOM: ICH HABE SIE NICHT VERSTANDEN.

[snip]

User: ja ich glaub’, ich m"ocht’ lieber Armageddon sehen. wo l"auft denn der, in welchem Kino? nehm’ ich doch das Kino

SMARTKOM: AUF DER KARTE SEHEN SIE DIE KINOS, IN DENEN DER FILM ARMAGEDDON L"AUFT.

User: ich m"ocht’ ins Kino Schlo"s.

SMARTKOM: WOLLEN SIE KARTEN RESERVIEREN?

User: ja, m"ocht’ ich gerne. f"ur Armageddon. wieso fragst du mich eigentlich, ob ich reservieren m"ochte und dann is’

SMARTKOM: EINE RESERVIERUNG IST LEIDER NICHT M"OGLICH.

User: ah, sch"oner Witz. na ja, dann

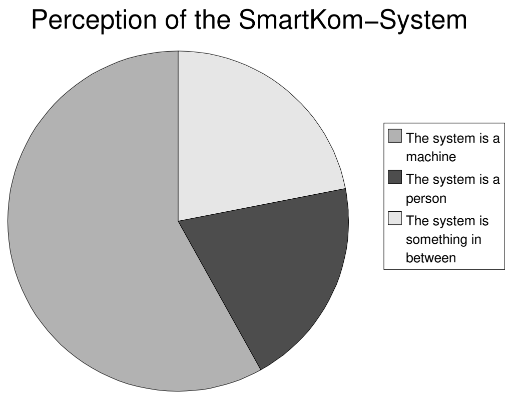
The Subjects And The Task

The first set of data (the cinema corpus) consisted of 70 sessions of about 4.5 min length each. 35 voluntary, naive subjects were recorded and were paid a small recompense. 18 subjects were female, 17 male. They were between the age of 19 and 60. Occupations: 24 students and 11 employed persons. The subjects were carefully instructed. They had to imagine themselves being in Heidelberg and using a new information booth for the first time. They were told to "Plan a trip to the cinema this evening". After the instruction two sessions were recorded (with a short break in between). Afterwards the subjects were led into a different room, where they were interviewed. They were, for example, asked about problems, what they had liked and disliked about the system, if it seemed more like a computer or a machine to them etc. Most subjects thought that the use of the system was easy. About two third thought the system was more like a machine, the rest judged it to be either something in between or rather more human. Even of the latter subjects, none reported suspicios that it actually was human.

Analysis Of The Data

The audio files were transliterated [7], the videos were labeled with regard to interesting user–states and gestures [8]. With these transliteration and label files the structure of the dialogue, the words used, the number and kind of gestures and the emotional states of the users can be studied. Additionally the subjects were interviewed. This data is analyzed qualitatively to answer questions about the WOZ–technique and how certain aspects of the simulation were perceived by the subjects [9].

Some Results

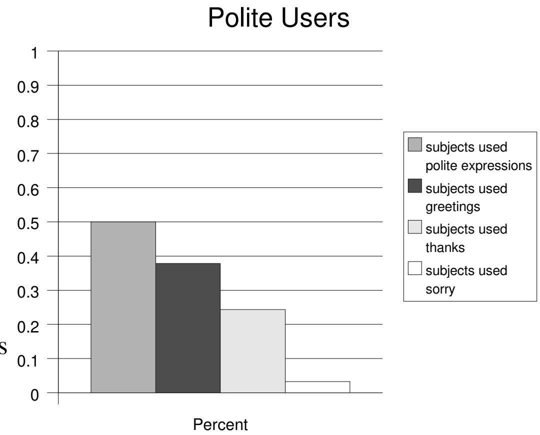


Apart from generating training material for recognizers, the experiments are expected to give insight into typical interaction patterns between users and a (rather intelligent) multimodal dialogue system. We hope to find some interesting theses about the behavior of subjects in such a situation which can then be tested in more controlled experiments.

Here are some of the interesting results:

1. **The system is mostly perceived as a machine**
In the interview most users stated that they had the impression of talking to a machine because of the stereotypical answers and because

2. **But some users perceive it more like a person and treat it appropriately:**

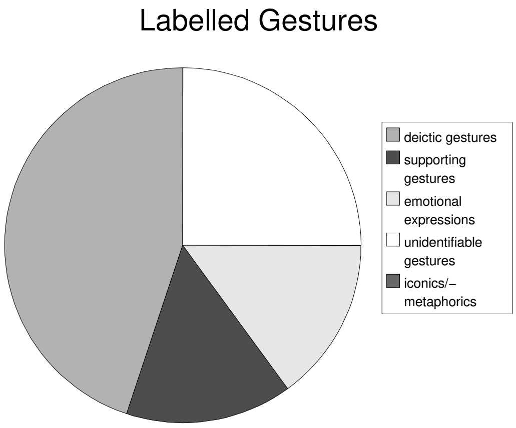


An artificial system that understands natural speech, natural gestures, has a voice and a face and reacts to the user more intelligently than known systems is treated by some of the users like an artificial dialogue partner – not like an object. They are polite: 37,8% used greetings like „hello“ and „good bye“, 24,4% used “thanks” and 3,3% even “sorry”.

3. **Body Language**

The body language of the users is more selective than in a human–human dialogue. Although subjects treat the system in some aspects like an artificial person (see above) their gestures are adapted to the communication situation: They stand in front of a display and want solve a task. Therefore they use mostly constructive gestures on or above the display and not much other body language. The labeled gestures were composed of:

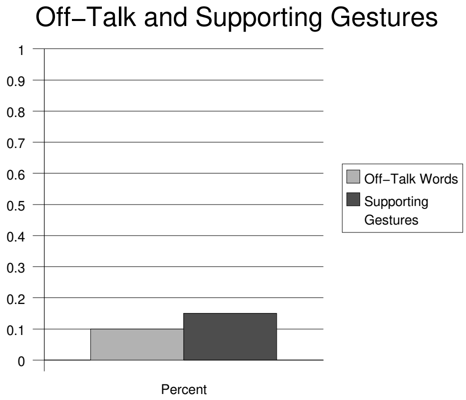
– mostly deictic gestures (about 45%)
– deictic gestures that were not addressed at the system but helped the subject make up her mind („supporting gestures“) (15%)
– emotional expressions (about 15%)
– unidentifiable gestures like posture changes, manipulators etc. (about 25%)
– No iconics or metaphorics
– By comparison: In narrative speech with a human listener McNeill found between 17–57% iconics, 3–15% metaphorics, 1–6% deictics [10]



4. “Off−Talk” and “Supporting Gestures”

Apart from the expected interactive behavior a form of (verbal and non−verbal) »monologue« behavior could be observed, which seemed to serve purposes like turnholding or the planning of the next interact ion.

- **Speech**: Utterances occurred that were not directed to the system as a question, a feedback utterance or as an instruction. We named these utterances „Off−Talk“ [11] . Examples are reading aloud text from the display, swearing or soliloquy/thinking aloud. Off−Talk is observed in about two thirds of the dialogues. About 10% of all spoken words are Off−Talk words [11].
- **Gestures**: The gestures of the subjects show a similar phenomenon: Some deictic gestures are not meant as an instruction to the system. They help *prepare* a request: A searching movement, following a line while reading, pointing to something to keep it in mind. We named these gestures „supporting gestures“[8]. In a human−human dialogue there are similar gestures (ideographs for example) which sketch a thought. These gestures normally take place in front of the thorax. We assume that the supporting gestures we describe are such "thought sketching" gestures.



Summary

To gather realistic data of human−machine interaction, the Wizard−of−Oz technique is especially suited. If the data is recorded and annotated, a wealth of information can be won. For example, by comparing speech and gesture annotations, connections between the two channels can be studied. By interviewing the subjects and observing their behavior, their attitude towards an „intelligent system“ can be assessed. Controlled experiments can be realized using this technique, but because of the high expense they are better used as observation tools to gather as much data as possible which is analyzed in a second step. WOZ experiments cannot replace observations in the field − nonetheless they can get very close to the „real thing“.

Acknowledgments

This research is being supported by the German Federal Ministry of Education and Research, grant no. 01 IL 905. We give our thanks to the SmartKom group of the Institute of Phonetics in Munich that provided the Wizard−of−Oz data and did the labelling and annotations. Special thanks to Victoria Penide−Lopez for last minute analysis, Florian Schiel and Ute Kartal for helpful comments on the document.

Literature

[1] http://smartkom.dfki.de/index.html

[2] Steininger, S. (2000). Transliteration of Language and Labeling of Emotion and Gestures in SMARTKOM. Workshop Proc. of the Second International Conference on Language Resources and Evaluation: Meta−Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources. Athens, Greece, pp. 49−51.

[3] Wooffitt, R. (1997). Humans, computers and wizards. Analysing human (simulated) computer interaction. Routledge: London.

[4] Dahlbäck, N., Jönsson A., & Ahrenberg, L. (1993). Wizard of Oz Studies − Why and How. Knowledge−Based Systems, 6 (4), p. 258−266.

[5] Jönsson, A., & Dahlbäck, N. (1988). Talking to a computer is not like talking to your best friend. Proc. of the First Scandinavian Conference on Artificial Intelligence, Tromso, Norway, p. 297− 307.

[6] Beringer N. (2001): Evoking Gestures in SmartKom − Design of the Graphical User Interface. Gesture Workshop 2001, London, UK.

[7] Beringer, N., Oppermann, D., Burger, S.: Transliteration spontansprachlicher Daten − Lexikon der Transliterationskonventionen − SmartKom (Version 1). SmartKom Technisches Dokument Nr. 2, Februar 2000.

[8] Steininger, S., Lindemann, B., and Paetzold, T. (2001). Labeling of Gestures in SmartKom − Concept of the Coding System. Gesture Workshop 2001, London, UK.

[9] Steininger, S., Kartal, U., (2001). User−Befragung − WOZ Experimente Kino/Restaurant, SmartKom Memo Nr. 8 .

[10] McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. Chicago: University of Chicago Press.

[11] Oppermann, D., Schiel, F., Steininger, S., Beringer, N. (2001). Off−Talk − A Problem For Human−Machine−Interaction? Eurospeech 2001, Aalsbourg, Denmark.