Abstract

This paper empirically investigate how humans use reference in space when interacting with a multimodal system able to understand written natural language and pointing with the mouse.

We verified that user expertise plays an important role in the use of multimodal systems: experienced users performed 84% multimodal inputs while inexpert only 30%. Moreover experienced are able to efficiently use modality shortening the written input and transferring part of the reference meaning on the pointing.

Analyzed data showed also the importance of the system layout: when very short labels (one character) are available users strongly adopt a redundant reference strategy, i.e. they reference the object in a linguistic way and use pointing too.

Starting from these facts some guidelines for future multimodal systems are suggested.

1 Introduction

Multimodal communication is used frequently and efficiently by humans to identify objects in physical space. By combining different modalities (e.g. speech and gestures), multimodal references act as efficient tools! for coping with the complexity of the physical space-as conveyed by visual perception-which can be communicate only partially by verbal expressions (Glenberg and McDaniell, 1992). Therefore, multimodal references can easily substitute too complex, too detailed, ambiguous, or undetermined verbal expressions. In particular, they simplify referent identification when the speaker or the hearer do not know the name of the target, or how to describe it.

For a long time, face-to-face communication has been considered a reliable model for natural language based human-computer interaction (henceforth HCI) (Schmaucks, 1987). Currently, little empirical work is available on what actually happens in multimodal HCI and how communication features cohabit with modern graphical interfaces (Oviatt, 1996; DeAngeli et al., 1996; Oviatt et al., 1997). Moreover, with only a few exceptions (Buxton, 1991; Brennan, 1991; Stock, 1995), direct manipulation interfaces have been seen as an antagonist of conversational interfaces, hindering a desirable synergy between the communication styles.

We believe that in multimodal HCI users find strategies that overcome natural language communication or direct manipulation paradigm alone, creating a new mixed communication form that makes the best use of both (Oviatt, 1996; Oviatt et al., 1997). The new communication, even if similar in principle to face-to-face communication, might be carried out in a far different fashion because one partner is a computer. As a matter of fact, some studies showed that people do adopt conversational and social rules when interacting with computers (Nass et al., 1994) but humans also design their utterances with the special partner in mind (Brennan, 1991).

Empirical studies on natural language human-computer interaction confirm general HCI peculiarities. Talking to a computer humans maintain a conversational framework but tend to simplify the syntactic structure, to reduce utterances length, lexicographic richness and use of pronouns (Jonsson and Dahlback, 1988; Dahlback and Jonsson, 1988; Oviatt, 1995). In other words users select a simplified register to interact with computers even if the (simulated) system has human capabilities (De Angeli, 1991).

These results suggest that face-to-face communication is not an adequate model for HCI. Therefore empirical studies are needed to develop predictive models of multimodal communication in HCI.

Empirical research becomes even more important when a multimodal system reproduces an unnatural modality combination, such as writing combined with pointing. Pointing while writing is highly different from pointing while speaking. In the first case, in fact, multimodal communication is hampered by
a single-modality-production constraint. Indeed, the requirement of moving the dominant hand back and forth between the keyboard and a pointing device implies a substitution of the natural parallel synchronization pattern (Levelt et al., 1985) with an unnatural sequential one. Nevertheless, the obligation of using the same effector for writing and pointing does not seem to have any inhibitory effect on multimodal input production (DeAngeli et al., 1996).

In general, deixis1 was found to be the most frequent referent identification strategy adopted by users to indicate objects. However, its occurrence depends strongly on the effort needed to indicate the target by a pure verbal reference. Moreover, we found a relevant percentage of redundant references2, a strategy never mentioned in the relevant literature and pretty different from common face-to-face communication strategies (DeAngeli et al., 1996).

Following the iterative design principles (Nielsen, 1993), we assume that experimental research, in the form of early simulations, should improve multimodal design by allowing to formulate reliable guidelines. From this point of view, the purpose of this paper is to investigate the effect of user expertise and system features on multimodal interaction in order to infer useful guidelines for future systems.

2 HCI issues in multimodal referring

In a HCI context where the user is required to write and point, we analyze the communication strategies that users spontaneously adopted at the very beginning of interaction. The main purpose of analyzing users’ spontaneous behavior is to develop design guidelines that might be taken into account when developing multimodal systems that have to support successful interaction from the very beginning, like “walk-up-and-use” interfaces.

Results of a simulation experiment have been analyzed to answer the following question:

Is multimodal interaction really instinctive, i.e. do naive users perform as experienced ones?

In general, multimodal systems appear to improve HCI by allowing humans to communicate in a more spontaneous way (Oviatt and Olsen, 1994; Oviatt, 1996). Therefore, one could infer that multimodal communication is the best interaction style for naive users. However, some authors suggest that language based interaction is mainly suitable for experienced users (Hutchins et al., 1986; Gentner and Nielsen, 1996). Indeed the opacity of language allows very flexible interaction, but requires previous knowledge3. We believe that experience, defined as computer science literacy, may increase the efficiency of multimodality. Here the notion of efficiency is defined, following (Mac Aogain and Reilly, 1990), as the capacity of the multimodal input to derive important semantic parts from information channels other than language, i.e. from pointing. In other words, efficiency is operationalized as the proportion of written input replaced by the gestural one.

3 Method

In order to evaluate spontaneous multimodal input production, data from the training session of a simulation experiment were analyzed.

Procedure The multimodal system called SIM Sistema Interattivo per la Modulistica was simulated to assist students with form filling tasks. Conversing with SIM, users had to gather information on how to fill out form fields (user questions) and to provide personal data for automatic insertion (user answers). Hard-copy instructions described system capability and required participants to complete the task as quickly and accurately as possible. No examples of dialogue were directly given, to avoid biasing communication behavior. Participants worked individually in a user room and were monitored by a closed circuit camera. Dialogues and pointing were logged and interactions videotaped during all experimental sessions. At the end all students filled in a user satisfaction questionnaire (USQ) and were debriefed.

Simulation The system was simulated by the Wizard of Oz technique, in which a human (the wizard) plays the role of the computer behind the human-computer interface (Fraser and Gilbert, 1991). A semi-automatic procedure supported the simulation that was carried out on two connected SUN SPARC workstations. Interface constraints and several pre-loaded utterances (including a couple of prefixed answers for every task-relevant action, error messages, help and welcoming phrases) supported two trained wizards. These strategies have been found to increase simulation reliability by reducing response delays and lessening the attentional demand on the wizard (Oviatt et al., 1992).

User interface was composed by a dialogue window in the upper part and a form window in the lower part of the screen (figure 1). In the dialogue

1Deixis concerns the ways in which languages encode or grammaticalyze features of the context of utterance or speech event (Levinson, 1983). Among the various types we consider here only space or place deixis used in a gestural way.

2We defined as redundant reference multimodal references composed by a full linguistic reference and a not needed additional pointing.

3This is opposite to WYSIWYG (What You See Is What You Get) interfaces where what can be done is clearly visible.
window users typed their input and read system output. Pointing was supported by mouse and pointer was constrained inside the form window.

SIM was simulated with good dialogue capabilities, anyway still far away from human abilities. It accepted every type of multimodal references, i.e., with or without linguistic anchorage for the gesture and either close or far pointing. It could understand ellipses and complete linguistic references. The form layout had nine fields grouped in three rows. Rows and fields had meaningful labels (one or two words) suggesting the required field. Users could refer both to single fields and to rows as a whole. At row selection SIM gave instruction on the whole row content. After users received row information, to further fields selection corresponded more synthetic instruction.

SIM required to click on the referred field and gave visual feedback to this. It supported multiple references too. System answers were always multimodal with demonstrative pronouns and synchronized (visual) pointing.

Participants and Design Twenty four students from Trieste University participated in the simulations as paid volunteers. Ages of participants ranged from 20 to 31 and all were Italian native speakers.

Participants were grouped in two different sets according to their computer experience. Selection was achieved by a self-administered questionnaire on computer attitude and experience. Half sample was represented by experienced users, skilled typists with positive attitude towards computers and some programming experience. The other half was composed by students who had never used a computer before.

4 Results and Discussion

Data were available from 24 participants yielding a corpus of 210 user questions (due to technical problems one inexpert was discarded). User answers had to provide personal data for automatic insertion, but they did not require to identify fields. So user answers were not included in the analysis.

Each user question was tabulated in one of the following five categories according to the referent identification strategy adopted in it:

- **direct naming**: it is a unimodal reference and occurs when the field label is explicitly used in the utterance, e.g., *il campo dati anagrafici* (the personal data field);
- **language reference**: it is a unimodal reference and occurs whenever the field is referred by a pure verbal input, but without direct naming, e.g., *l'ultimo campo* (the last field). This category includes, among others, anaphoric reference and metonymy;
- **deixis**: it is a multimodal reference that occurs whenever it exists an explicit anchor (deictic linguistic expression) for the pointing, e.g., *questo campo* (this field);
- **mixed**: it is a multimodal reference that occurs when the reference contains both linguistic and gestural part, but no deictic marks can be found in the utterance, e.g., *in A* (in /);
- **redundant**: it is a multimodal reference; it occurs when one component (or part of it) is not needed for the understanding, e.g., *nel campo A* (the field A /).

Figure 2 shows the percentages of each referent identification strategies as a function of user expertise. It clearly emerges that previous knowledge affects strategy selection. Multimodal input were strongly preferred by expert users, while inexpert preferred unimodal linguistic references, especially direct naming. These results imply that communication behavior may be predicted by knowing previous expertise.

Multimodal occurrence strongly increased efficiency of communication. Utterance length was found to be inverse correlated to the number of multimodal input ($r = .48, p < .05$). In average, expert users wrote almost 3 words each utterance, while inexpert nearly the double.

It is interesting to notice that on the total sample deixis and mixed input occurs close to the same frequency. Mixed input implies a contraction of verbal

$^4$ In a previous study we demonstrated that visual feedback allows more efficient multimodal input, increases integration of pointing within writing and is preferred by users (DeAngeli et al., 1996; DeAngeli, 1997).

$^5$ The difference is significant according to results of an ANOVA $F(1,22) = 12.21, p < .001$. 

---

**Figure 1**: The user screen during an interaction.
input which is partially substituted by pointing, as in *cosa* (what). This phenomenon is peculiar to HCI, and pretty infrequent in face to face communication where deixis (e.g. *cosa qui* - what here) represents the maximum efficient input.

Redundant input were pretty rare, with no significant difference due to expertise. This findings are in contrast with what we observed in (DeAngeli et al., 1996). This difference may be due to different interface layout.

The system simulated in (DeAngeli et al., 1996), called MIS (Multimodal Intelligent System), was more coercive than SIM. It was not able to understand ellipses, while it could understand complete linguistic references. MIS required users to produce linguistic expression containing explicit deictic anchor (i.e. demonstrative pronouns) and to execute the pointing immediately before or after it, e.g. *cosa metto in questo campo* (what have I to put in this field). Any other multimodal expression was refused. MIS layout had single fields with very short labels (a single letter) not related to the required content of the field (figure 3).

Coming back to redundant input, we evince that the significant different rate when interacting with the two systems was due to form layouts. Indeed, redundant references in the case where labels were one character long was no-cost compared to the case where labels where one/two words long. This suggests that system layout may influence communication behavior. In the next chapter we discuss related guidelines.

When designing whatever system, specialists should consider both system functionalities and interaction features depending on the typology of users and on the tasks they will perform. In current multimodal systems, this balancing among functionalities and users has not been considered enough.

For example, the obligation for the user to point at a certain time while writing was found to be in contrast with his/her natural inclination. This constraint would be justified only if synchronization understanding is a true problem, e.g. if the users use multiple selection or pars-pro-toto. Our data shows that, at least, this is not the case at the beginning of the interaction: expert used multiple selection in the 0.22% of the total multimodal references while for inexpert the percentage decrease furthermore to 0.14%.

5 Guidelines

In this section we state some useful guidelines for designers of walk-up-and-use multimodal systems or for systems that have to have a successful interaction from the very beginning.

As widely discussed above, user expertise is a factor that deeply influence multimodal interaction. Multimodal systems are definitely suite for experienced users that use it in the 84% of all the considered interactions. At the opposite, multimodality was not exploited by the inexpert that use it only for 30%. This may suggest that multimodal systems are definitely suite for expert users even from the very beginning of the interaction, while inexpert have difficulties in exploiting multimodality interaction limiting themselves to inefficient linguistic references.

Another interesting point is that experienced users perform nearly the same percentage of deixis (38%) and of mixed modality (42%). This suggests that system should be flexible enough to accept whatever combination of pointing and writing, not requiring a well formed deixis. This is strengthened by the fact that users can find very efficient ways for referring, optimizing writing and pointing and exploiting the context as in “I?” where the meaning is conveyed by the gesture, by the minimal writing and by considering the task the user is performing.

Lastly, the influence of the layout on the user be-
behavior has to be underlined. In fact, the possibility of referring to object in a no-cost linguistic way encourages the user to use redundant references. This suggest to use shortest label when a double reference is useful to discriminate objects, for example on dense maps.

References

Susan Brennan. 1991. Conversation as direct manipulation: An iconoclastic view. In Brenda Laurel, editor, The Art of Human-Computer Interface Design, pages 393–404. Addison-Wesley.

Bill Buxton. 1991. The "natural" language of interaction: A perspective on nonverbal dialogues. In Brenda Laurel, editor, The Art of Human-Computer Interface Design, pages 405–416. Addison-Wesley.

Nils Dahlbäck and Arne Jonsson. 1989. Empirical studies of discourse representations for natural language interfaces. In 4th European Conference of ACL, pages 291–298.

Antonella De Angeli. 1991. Dillo maia. Master's thesis, Psychology Dept., University of Trieste.

Antonella De Angeli, Daniela Petrelli, and Walter Gerbino. 1996. Interface features affecting deixis production: A simulation study. In WIGLS - Workshop on the Integration of Gesture in Language and Speech (at ICSLP '96), pages 195–204.

Antonella De Angeli. 1997. Valutare i sistemi flessibili: un approccio globale alla HCI. Ph.D. thesis, University of Trieste.

N. Fraser and G. Gilbert. 1991. Simulating speech systems. Computer Speech and Language, pages 81–99.

Don Gentner and Jakob Nielsen. 1996. The anti-mac interface. Communication of the ACM, 39(8):76–82.

A. Glenberg and M. McDaniel. 1992. Mental models, pictures, and text: Integration of spatial and verbal information. Memory and Cognition, 20(5):458–460.

Edwin L. Hutchins, James D. Hollan, and Donald A. Norman. 1986. Direct manipulation interfaces. In Donald A. Norman and Stephen W. Draper, editors, User Centered System Design: new Perspectives on Human-Computer Interaction, chapter 5, pages 87–124. Lawrence Erlbaum Associates.

Arne Jonsson and Nils Dahlbäck. 1988. Talking to a computer is not like talking to your best friend. In Scandinavian Conference on Artificial Intelligence - SCAI '88, pages 53–68.

Willem Levelt, Graham Richardson, and Wido La Heij. 1985. Pointing and voicing in deictic expressions. Journal of Memory and Language, 24:133–164.

Stephan Levinson. 1983. Pragmatics. Cambridge Textbook in Linguistics. Cambridge University Press.

Eoghan Mac Aogain and Ronan Reilly. 1990. Discourse theory and interface design: The case of pointing with the mouse. International Journal of Man-Machine Studies, 32:591–602.

Clifford Nass, Jonathan Steuer, and Ellen Tauer. 1994. Computers are social actors. In CHI '94 Human Factors in Computing Systems, pages 72–78. ACM Press.

Jakob Nielsen. 1993. Usability Engineering. Academic Press.

Sharon Oviatt and Eric Olsen. 1994. Integration themes in multimodal human-computer interaction. In International Conference on Spoken Language Processing, pages 551–554.

Sharon Oviatt, P. R. Cohen, M. Fong, and M. Frank. 1992. A rapid semiautomatic simulation technique for investigating interactive speech and hand writing. In International Conference on Spoken Language Processing, pages 1351–1354.

Sharon Oviatt, Antonella De Angeli, and Karen Kuhn. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In CHI '97 - Human Factors on Computing Systems. ACM Press.

Sharon Oviatt. 1995. Predicting spoken disfluences during human-computer interaction. Computer Speech and Language, 9:19–35.

Sharon Oviatt. 1996. Multimodal interfaces for dynamic interactive maps. In CHI '96 - Human Factors in Computing Systems. ACM Press.

Dagnar Schmuckes. 1987. Natural and simulated pointing: An interdisciplinary survey. Technical report, FB Informatik, Universität Saarbrücken.

Oliviero Stock. 1995. A third modality? Artificial Intelligence Review, 9:129–146.