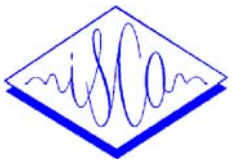

5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

# REFERENTIAL FEATURES AND LINGUISTIC INDIRECTION IN MULTIMODAL LANGUAGE*

*Sharon Oviatt & Karen Kuhn*†

Center for Human-Computer Communication, Department of Computer Science
Oregon Graduate Institute of Science & Technology

## ABSTRACT

The present report outlines differences between multimodal and unimodal communication patterns in linguistic features associated with ease of dialogue tracking and ambiguity resolution. A simulation method was used to collect data while participants used spoken, pen-based, or multimodal input during spatial tasks with a dynamic system. Users' linguistic constructions were analyzed for differences in the rates of reference, co-reference, definite and indefinite referring expressions, and deictic terms. Differences also were summarized in the prevalence of linguistic indirection. Results indicate that spoken language contains substantially higher levels of referring and co-referring expressions and also linguistic indirection, compared with multimodal language communicated by the *same users completing the same task*. In contrast, multimodal language not only has fewer referential expressions and relatively little anaphora, it also specifically lacks the regular use of determiners observed in spoken definite and indefinite noun phrases. In addition, multimodal language is distinct in its high levels of deictic reference. Implications of these findings are discussed for the relative ease of natural language processing for speech-only versus multimodal systems.

## 1. INTRODUCTION

Among other things, the goals of designing multimodal rather than unimodal systems typically include enhanced ease of use, transparency, flexibility, and efficiency— as well as usability for more challenging applications, under more adverse conditions, and by a broader spectrum of the population. However, actually achieving these goals will depend on basic empirical work on a variety of fundamental issues, as well as the construction of theoretical models that predict the performance advantages of spoken, manual, and combined input modes for different types of tasks. Recent research has indeed documented a variety of performance advantages associated with interacting multimodally over using speech alone— including briefer task completion time, reduced errors, and a strong user preference to interact multimodally (Oviatt, 1997). Other studies have indicated that a flexible multimodal interface can assist in avoiding recognition errors and also resolving them more gracefully (Oviatt, in press; Oviatt & vanGent, 1996). These advantages appear to be most pronounced when users are interacting in a visual-spatial domain.

At the level of linguistic differences, comparisons also have revealed that multimodal language can involve shorter and less complex constructions than speech-only utterances. In particular, the same user completing the same map-based task communicates fewer words, briefer sentences, and language containing fewer complex spatial descriptions and disfluencies when interacting multimodally, compared with using speech alone (Oviatt, 1997). The following is an example of a typical user's spoken input while attempting to designate an open space using a map system:

> *"Add an open space on the north lake to b-- include the north lake part of the road and north."*

In contrast, the same task was accomplished multimodally by encircling a specific area and saying:

> *"Open space."*

In previous research, hard-to-process disfluent language has been reduced by 50% during multimodal interaction (Oviatt, 1997). This drop occurs because people have difficulty speaking spatial information, which precipitates disfluencies. In a flexible multimodal interface, they instead use pen input to convey spatial information, and thereby avoid the need to speak it. These simplified linguistic features of multimodal language are expected to facilitate more robust natural language processing in systems designed to handle multimodal input.

In other respects, multimodal language clearly is different than spoken language, although not necessarily simpler. For example, users' multimodal language has been observed to depart from the canonical English word order of S-V-O-LOC (i.e., Subject-Verb-Object-Locative constituent)— which is observed in spoken language and also formal textual language. Instead, users' multimodal constituents shift to a LOC-S-V-O word order. A recent study reported that 95% of locative constituents were in *sentence-initial* position during multimodal interaction. However, for the same users completing the same tasks while speaking, 96% of locatives were in *sentence-final* position (Oviatt, DeAngeli & Kuhn, 1997).

The goal of the present research was to compare the linguistic differences and relative ease of processing multimodal input compared with unimodal input. To pursue this goal, a simulation method was used to collect data while participants used speech, pen, or multimodal pen/voice input during a spatially-oriented

* This research was supported by Grant No. IRI-9530666 from the National Science Foundation and Grant No. DABT63-95-C-007 from DARPA.
† First author: Center for Human-Computer Communication, Department of Computer Science, Oregon Graduate Institute of Science & Technology, P.O. Box 91000, Portland, OR, 97291 (oviatt@cse.ogi.edu; http://www.cse.ogi.edu/CHCC/); Second author: Al Tech Corp., Boston, Mass.

map task. Users' linguistic constructions then were analyzed for differences in linguistic features associated with ease of dialogue tracking and ambiguity resolution. More specifically, multimodal versus unimodal communication patterns were compared in terms of their rates of overall referring expressions, including definite and indefinite reference, co-reference, and deictic terms. They also were compared on linguistic indirection.

It was hypothesized that the overall rate of referring expressions would be higher during spoken interaction than multimodal, with the rate of co-referring expressions in need of anaphoric tracking and resolution also higher. Whereas on the one hand anaphora may be expected to place greater demands on the natural language processing for speech, on the other hand deictic expressions were expected to be a more common feature in need of processing for multimodal language. It also was predicted that linguistic indirection would occur at higher rates during speech than multimodal pen/voice language, since the presence of pen input in the latter may well influence its degree of linguistic directness.

Consistent with the lower rates of referring expressions during multimodal interaction, it also was hypothesized that explicit linguistic specification of definite and indefinite reference would be less common when interacting multimodally. Unfortunately, current natural language processing algorithms typically rely heavily on the specification of determiners in definite and indefinite reference in order to resolve noun phrase reference. As a result, the relative increase in deixis and elided noun phrases in multimodal language suggests that current algorithms are not yet prepared to handle the input from next-generation multimodal exchanges.

# 1. METHOD

This section summarizes a simulation experiment that was designed to facilitate assessment of users' language while interacting with a dynamic map system.

## 2.1 Subjects, Tasks, and Procedure

Eighteen subjects participated in this research as paid volunteers. A "Service Transaction System" was simulated that could assist users with map-based tasks such as real estate selection. During these tasks, for example, participants could circle a lakeside house icon with their pen and say "I don't want a house in a flood zone." In response, the system would display waterways and flood zones, and would filter out a house icon if it was located unacceptably. During a different task, participants could add new municipal buildings and parks, and could indicate road closures or extensions. They also could use speech and pen input to manipulate the map display by zooming, scrolling, automatically locating entities, and so forth.

During the study, subjects received a general orientation to the Service Transaction System, and instructions and practice entering information on the LCD tablet while writing, speaking, and combining both modalities. During free choice, people were completely free to use either modality in any way they wished. They were encouraged to speak and write naturally, to work at their own pace, and simply to focus on completing their task. Other than specifying the available input modality, an effort was made not to influence the manner in which people expressed themselves. After the session, a post-experimental interview was conducted and subjects were debriefed.

## 2.2 Semi-Automatic Simulation Technique

People's input was received by an informed assistant, who performed the role of interpreting and responding as a fully functional system would. Essentially, the assistant tracked the subject's written or spoken input, and clicked on predefined fields at a Sun SPARCstation to send altered map displays and confirmations back to the subject. Technical details of the general simulation method have been provided elsewhere (Oviatt et al., 1992). However, the specific method used in this study was adapted extensively to handle the dynamic display and manipulation of maps, overlays, and photographs.

In developing this simulation, an emphasis was placed on providing automated support for streamlining the simulation to the extent needed to create facile, subject-paced interactions with clear feedback, and to have comparable specifications for the different input modalities. In the present simulation environment, response delays averaged less than 1 second between subject input and system feedback. The simulation was designed to be sufficiently automated so that the assistant could concentrate attention on monitoring the accuracy of incoming information, and on maintaining sufficient vigilance to ensure prompt responding.

## 2.3 Research Design and Data Capture

The research design involved repeated measures, with the communication modality varied from speech-only and pen-only input to multimodal pen/voice. In total, data were available from 18 subjects, 216 tasks, and approximately 2,700 utterances and 12,000 words for analysis purposes. All human-computer interactions were videotaped. Hardcopy transcripts also were created, with the handwritten input captured automatically in the context of the map, and spoken input and sequencing information transcribed onto the printouts.

## 2.4 Transcript Coding

**Referring expressions—** The total number of noun phrase referring expressions (i.e., whether in reference to new or existing and previously referenced entities) was summarized during speech-only, pen-only, and multimodal input. This general class of referring expressions included proper names, definite and indefinite references, deictic and pronominal references, and "direct" references. Direct references were ones in which an entity was placed directly on the map using the pen to draw an icon (e.g., house) or mark a simple graphic (e.g., dot).

**Co-referring expressions**— The subset of all noun phrase referring expressions that were used to designate a previously-referenced map entity was summarized during speech, pen, and multimodal input. The ratio of co-referring to total referring expressions also was summarized for each modality.

**Definite & indefinite reference—** The total number of definite noun phrases (e.g., "the post office") and indefinite noun phrases (e.g., "a hospital" or "hospital") was summarized for each input modality.

**Deictic expressions—** The total number of deictic terms used to designate entities or locations on the map was summarized. The most common deictic terms (i.e., accounting for 96% of all those observed) were "here", "there", "this" and "that." For example, a user interacting multimodally might circle a house and ask: "Is this brick?" Other less frequent deictics terms included "those", "these" and "it."

**Linguistic indirection—** The total number of utterances in which users expressed a speech act indirectly rather than directly was summarized for each input modality. In this application, the most common example of a direct speech act involved issuing a direct command to the system to perform an action (e.g., "Add a boat dock here.") Common examples of indirect speech acts included the use of a question or statement to express a command to the system (e.g., "Could you put a house next to the musem?" or "I'd like a house next to the museum" to mean that the system should add a house on the map next to the museum).

**Reliability—** All dependent measures reported in this paper had inter-rater reliabilities of 0.80 or above.

## 2. RESULTS

The following results summarize differences between modalities in the presence of different linguistic features for the same subjects completing the same type of tasks.

**Referring expressions—** The total number of referring expressions was 428 and 452 during pen-only and multimodal interaction, respectively, but increased to 696 during speech-only interaction. As predicted, statistical comparisons confirmed that these types of reference were significantly more common during speech-only interaction compared with multimodal interaction, paired t test = 5.66 (df = 17), $p < .001$, one-tailed, and also compared with pen-only interaction, paired t test = 5.61 (df = 17), $p < .001$, one-tailed. No differences were present between multimodal and pen-based interaction, $t < 1$. These data indicate that 54% more linguistically-specified references of this kind were expressed during spoken interaction than multimodal.

Within the multimodal condition when users were free to combine modalities as they wished, a comparison also was conducted of the rate of referring expressions for commands that were exclusively spoken or written, versus those that actually were conveyed multimodally using pen and voice together. These analyses replicated the finding that referring expressions were more prevalent during multimodally-composed commands than during unimodal ones.

**Co-referring expressions—** The total number of co-referring expressions was 231 and 240 during pen-only and multimodal interaction, respectively, but increased to 436 during speech-only interaction. As predicted, statistical comparisons confirmed that co-reference was significantly elevated during speech-only interaction compared with multimodal interaction, paired t test = 5.03 (df = 17), $p < .001$, one-tailed, and also compared with pen-only interaction, paired t test = 5.47 (df = 17), $p < .001$, one-tailed. Once again, no differences were present between multimodal and pen-based interaction, $t < 1$. These data indicate that 82% more co-references were expressed and required tracking during spoken interaction than multimodal.

Within the multimodal condition, when users were free to combine modalities as they wished, a comparison also was conducted of the rate of co-referring expressions for commands that were exclusively spoken or written, versus those that actually were conveyed multimodally using pen and voice together. These analyses replicated the finding that co-referring expressions were more prevalent during multimodally-composed commands than during unimodal ones.

The overall ratio of co-referring to referring expressions averaged .63 during spoken interactions, compared with .53 and .54 during multimodal and pen-based interactions. These rates confirm a relatively heavier use of co-reference while speaking, compared with other modes.

**Definite & indefinite reference—** Followup analyses of this largest subgroup of referring expressions (i.e., accounting for 60%) indicated that the total number of definite and indefinite references specified was 192 during pen-only interaction and 204 during multimodal interaction, but increased to 418 during speech-only interaction. As predicted, statistical comparisons confirmed that definite and indefinite references were specified significantly more often during speech-only than multimodal interaction, paired t test = 5.86 (df = 17), $p < .001$, one-tailed, and also compared with pen-only interaction, paired t test = 6.23 (df = 17), $p < .001$, one-tailed. No significant difference was present between multimodal and pen-based interaction, $t < 1$. These data indicate that 105% more definite and indefinite noun phrases were specified during spoken than multimodal interaction.

**Deictic expressions—** The total number of deictic expressions during multimodal interaction was 51, in comparison with none during written interaction and just 2 during spoken interaction. As predicted, Wilcoxon Signed Ranks tests confirmed significantly more deictic expressions during multimodal interaction than during speech-only, $z = 2.81$ (df = 11), $p < .025$, one-tailed, or pen-only interaction, $z = 2.81$ (df = 11), $p < .025$, one-tailed.

Within the multimodal condition when users were free to combine modalities as they wished, a comparison also was conducted of the rate of deictic expressions for commands that were exclusively spoken or written, versus those that actually were conveyed multimodally using pen and voice together. These analyses replicated the finding that deictic expressions were more prevalent during multimodally-composed commands than during spoken or pen-based ones.

**Linguistic indirection—** Overall, 11% of speech-only constructions, 7% of multimodal constructions, and 2% of pen-based constructions were expressed in an indirect rather than direct manner. As predicted, Wilcoxon Signed Ranks tests confirmed that indirect constructions were significantly more common during spoken interactions than multimodal ones, $z = 1.90$ (df = 13), $p < .03$, one-tailed, or pen-based ones, $z = 3.13$ (df = 15), $p < .001$, one-tailed. However, indirect constructions also were significantly more common during multimodal interactions than pen-based ones, $z = 2.76$ (df = 11), $p < .003$, one-tailed.

A further breakdown of the multimodal condition confirmed that speech-only constructions within this condition were expressed indirectly 10% of the time, whereas constructions that users

actually composed multimodally using pen and voice were expressed indirectly just 4% of the time. A Wilcoxon Signed Ranks test confirmed that when users were interacting within a multimodal condition, constructions that were actively expressed multimodally were less likely to be expressed indirectly than ones that were exclusively spoken.

## 4. DISCUSSION

Compared with speech, the present results indicate that multimodal pen/voice language to a map system contains substantially fewer referring expressions overall. In particular, the number of co-referring expressions is selectively reduced during multimodal human-computer communication. This reduction in anaphora would simplify natural language processing in the sense of easing the need for anaphoric tracking and resolution in this type of multimodal interface.

Consistent with the lower rates of referring expressions during multimodal interaction, explicit linguistic specification of definite and indefinite reference also is less common. Current natural language processing algorithms typically rely heavily on the specification of determiners in definite and indefinite references in order to represent and resolve noun phrase reference in the tradition of Montague (1974). One unfortunate by-product of the lack of traditional specification of determiners in multimodal language is that current language processing algorithms are unprepared for the frequent occurrence of deixis and elision in next-generation multimodal interactions. One implication of these differences is that new multimodal corpora, statistical language models, and natural language processing algorithms will need to be built to process multimodal language successfully. For a discussion of additional issues on the topic of natural language processing for future multimodal systems, see Oviatt, DeAngeli & Kuhn (1997) and Oviatt (1997).

The large number of co-referring expressions typical of spoken language appears to have been transformed into more deictic expressions when users interacted multimodally. Typically, the meaning of these multimodal deictic terms could be interpreted from users' residual ink after pointing, encircling, or making other pen marks to select an object on the map. In some cases, the meaning of deictic terms could be disambiguated by the visual context of the map itself (e.g., by system highlighting on the intended map object, or the exclusive presence of one object on the map of the relevant type).

In summary, whereas anaphora may place greater demands on natural language processing for speech, there is a relatively greater need for algorithms designed to handle different types of deictic expressions in future multimodal interfaces. However, as a word of caution it is noteworthy that past research has reported most multimodal constructions (i.e., 59%) do not contain any spoken deictic— so one cannot always count on the their presence to flag and assist in interpreting the referent in a visual display. In addition, even fewer multimodal constructions (i.e., 25%) contain a spoken deictic that is overlapped in time with the pen input needed to disambiguate its meaning (Oviatt, DeAngeli & Kuhn, 1997).

Finally, the present results also indicate that linguistic indirection is more prevalent in users' spoken language, and that these indirect expressions tend to be replaced by more direct commands during multimodal communication. Linguistic indirection may decrease during multimodal pen/voice language in part because of the influence of manually-oriented pen input, which is a direct physical medium of interaction. In the following example, a study participant makes an indirect request using speech input while requesting a distance calculation:

> *"What is the distance between the Victorian Museum and the, uh, the house on the east side of Woodpecker Lane?"*

When requesting distance information multimodally, the same user encircled the house and museum while speaking the following brief direct command:

> *"Show distance between here and here."*

Based on the present data as well as previous research, there now is cumulative evidence that many linguistic features of multimodal language are qualitatively very different from that of spoken or formal textual language. In fact, it differs in features as basic as brevity, semantic content, syntactic complexity, word order, disfluency rate, degree of ambiguity, referring expressions, specification of determiners, anaphora, deixis, and linguistic indirectness. In many respects, multimodal language is simpler linguistically than spoken language. One implication of these findings is that multimodal interface design has the potential to support more robust future systems than a unimodal design approach. Future research and corpus collection efforts will be needed on different types of multimodal interaction in other application domains in order to establish the generality of the linguistic differences outlined in this research.

## 5. REFERENCES


1. Montague, R. The proper treatment of quantification in ordinary English. In R. H. Thomason (Ed.), *Formal Philosophy: Selected Papers of Richard Montague*, New Haven: Yale University Press, 1974.

2. Oviatt, S. L. Multimodal interactive maps: Designing for human performance, Human-Computer Interaction, 1997, 12, 93-129 (special issue on ‘‘Multimodal Interfaces").

3. Oviatt, S. L. Ten myths of multimodal interaction, *Communications of the ACM*, in press.

4. Oviatt, S. L., DeAngeli, A. & Kuhn, K. Integration and synchronization of input modes during multimodal human-computer interaction, in *Proceedings of Conference on Human Factors in Computing Systems: CHI '97*, New York, N.Y.: ACM Press, 415-422.

5. Oviatt, S. L., Cohen, P. R., Fong, M. W., and Frank, M. P., A rapid semi-automatic simulation technique for investigating interactive speech and handwriting, *Proceedings of the International Conference on Spoken Language Processing*, ed. by J. Ohala et al., University of Alberta, 1992, vol. 2, 1351-1354.

6. Oviatt, S.L. & VanGent R. Error resolution during multimodal human-computer interaction, T. Bunnell & W. Idsardi, eds., in Proceedings of the International Conference on Spoken Language Processing, University of Delaware & A.I. duPont Instit., 1996, vol 1, 204-207.