Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction

Andrew Kehler
Department of Linguistics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0108
kehler@ling.ucsd.edu

Abstract

We analyze a corpus of referring expressions collected from user interactions with a multimodal travel guide application. The analysis suggests that, in dramatic contrast to normal modes of human-human interaction, the interpretation of referring expressions can be computed with very high accuracy using a model which pairs an impoverished notion of discourse state with a simple set of rules that are insensitive to the type of referring expression used. We attribute this result to the implicit manner in which the interface conveys the system's beliefs about the operative discourse state, to which users tailor their choice of referring expressions. This result offers new insight into the way computer interfaces can shape a user's language behavior, insights which can be exploited to bring otherwise difficult interpretation problems into the realm of tractability.

Introduction

Despite recent advances in natural language processing (NLP) technology, computers still do not understand natural language interactions very well. Language is rife with complex phenomena that elude our understanding, and thus do not avail themselves of robust methods for interpretation. Solving these difficult problems, of course, is the basis for continuing research in NLP. On the other hand, we might also study the ways in which computer interfaces can shape a user's language behavior, and capitalize on these to reduce the complexity of certain interpretation problems.

In this paper, we consider the problem of resolving reference in human-computer interaction (HCI) with a multimodal interface, to which users can speak and gesture. We show that unlike normal modes of human-human interaction (HHI), reference resolution in an HCI system developed using standard interface design principles admits of a simple algorithm that nonetheless yields very high accuracy. We attribute this unexpected result to the implicit manner in which the interface conveys the system's beliefs about the operative discourse state, to which users tailor their choice of referring expression. This result offers new insight into the way that computer interfaces can shape a user's language behavior, insights which can be exploited to bring otherwise difficult interpretation problems into the realm of tractability.

Copyright © 2000, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Form of Reference, Cognitive Status, and Salience

From a computational linguistic standpoint, it would be nice if speakers always referred to entities using complete and unambiguous referring expressions, thereby rendering reference resolution unproblematic. Of course, this is not what competent speakers do. Instead, natural languages provide speakers with a variety of ways to refer to entities and eventualities when producing an utterance, including pronouns, demonstratives, lexical noun phrases, and proper names. For instance, a particular Four Seasons Hotel might be referred to as *it*, *this*, *that*, *here*, *there*, *this hotel*, *that hotel*, *the hotel*, or *the Four Seasons*. Importantly, these alternatives are not freely interchangeable, as each encodes different signals about the location of the referent within the hearer's mental model of the discourse – the referent's *cognitive status* – which are necessary for the hearer to identify the intended referent (Chafe 1976; Prince 1981; 1992; Gundel, Hedberg, & Zacharski 1993, inter alia). One reason why accurate algorithms for reference resolution are elusive is the lack of reliably computable methods for determining a potential referent's cognitive status.

To add concreteness to our discussion, we sketch a particular theory of cognitive status, due to Gundel et al. (1993), who propose a *Givenness Hierarchy* containing six cognitive statuses that referents can have and the types of referential expressions that signal them.[1]

in focus	>	activated	>	familiar	>	uniquely identifiable
it		*that* *this* *this* N		*that* N		*the* N

Each cognitive status logically implicates those to its right in the hierarchy; for instance, an *in focus* referent is necessarily also *activated*, *familiar*, and *uniquely identifiable*.

[1] We will restrict our analysis to definite reference, and thus only the first four statuses in Gundel et al.'s hierarchy.

Thus, a form that normally signals a given cognitive status can be used to refer to an entity with a higher one. However, in a survey of data across several languages, Gundel et al. found that with one exception, each form was found almost exclusively with the status with which it is correlated. The exception in English is the case of definite lexical noun phrases, which were found not only with *uniquely identifiable* referents, but with all higher statuses. Gundel et al. explain these facts using Grice's Maxim of Quantity (Grice 1975), which can be paraphrased as *Make your contribution as informative as required, but not more so*. The first part of the maxim explains why demonstratives are not typically found with referents holding a higher status, as their use conversationally implicates that the higher status does not hold. On the other hand, unlike demonstratives and pronouns, definite lexical noun phrases typically contain the descriptive content necessary to uniquely identify the referent, so an explicit signal of a higher status is unnecessary, per the second half of the maxim.

Theories of information status such as Gundel et al.'s are useful for characterizing the types of referential expression with which a referent is compatible, which helps explain why different referential expressions are used in different contextual circumstances. However, these theories do not contain the degree of specificity required to capture all the constraints required for a computational model for reference resolution: They lack formal, independent conditions for determining the status of a referent in a particular discourse situation, as well as a way to distinguish between several possible referents that hold the same cognitive status. Developers of computational models – who have centered largely on a single cognitive status (in focus) and its correlated referential form (pronominalization) (Sidner 1983; Lappin & Leass 1994; Grosz, Joshi, & Weinstein 1995, inter alia) – have addressed these questions by incorporating a notion of *salience* into their models, along with sets of linguistic factors that are used to compute approximations to degree of salience (e.g., topic continuity, grammatical role, parallelism, and semantic factors). The factors determining salience with respect to other cognitive statuses (e.g., activated), and other types of referring expressions (e.g., demonstrative reference), are even less well understood, and thus have not received as much attention (but cf. Sidner (1983), Passonneau (1989)).

In this paper, we are interested in modeling reference behavior across the spectrum of referential form types for the case of multimodal HCI, which is different from HHI in several respects. First, the discourse state is augmented by the existence of entities in a prominently displayed, shared situational context consisting of icons on a computer screen (compare this, for instance, with the diminished role of situational context in normal HHI settings, and the almost complete lack thereof in a telephone conversation). Second, referring is readily performed using gesture to this shared context, possibly (but not necessarily) augmented with a natural language referring expression. Although these factors might be expected to further complicate the discourse modeling problem, we present an analysis of a corpus of reference data which suggests that, in dramatic contrast to HHI, the interpretation of referring expressions can be computed with very high accuracy with a model consisting of an highly impoverished notion of discourse state and a simple set of rules which are not sensitive to the type of referring expression used. We detail an experiment and its results in the next section, and then explain how these results arise from the implicit manner in which the interface conveys the system's beliefs about the cognitive status of referents to which users tailor their choice of referring expressions.

A Multimodal Map Application and Experiment

The basis for our study is an implemented prototype multimodal travel guide application (Cheyer & Julia 1995; Julia & Cheyer 1997) that was inspired by a multimodal Wizard of Oz simulation (Oviatt 1996). The system provides an interactive interface on which the user may draw, write, or speak. It makes available information about hotels, restaurants, and tourist sites in Toronto that have been retrieved by distributed software agents from commercial Internet World Wide Web sites. The map application is implemented within a multiagent framework called the Open Agent Architecture (OAA) (Martin, Cheyer, & Moran 1999).

S:	What do we have here? (circles China Town) *China Town selected*
S:	Give me information about this (draws arrow to China Town) *Textual description displayed*
S:	Okay so I think I'll visit China Town Saturday afternoon *Calendar displayed and updated*
S:	Is there any, uh, good restaurant around this area? (draws large circle including two sites) *Map scrolled to show restaurant*
S:	Well, no, I was looking...that way (draws arrow) *Map scrolled*
S:	So what do we have here? (circles restaurant) *Restaurant selected and name displayed*
S:	Give me information about this *Textual description displayed*
S:	Okay, I will have dinner there *Calendar updated*

Table 1: Example Interaction

Because we were interested in collecting naturally occurring data which may include phenomena not currently handled by the system, we designed a Wizard of Oz (WOZ) experiment. In WOZ experiments, users believe they are interacting directly with an implemented system, but in actuality a human "wizard" intercepts the user's commands and causes the system to produce the appropriate output. Subjects were asked to plan activities during and after a hypothetical business trip to Toronto. They planned places to

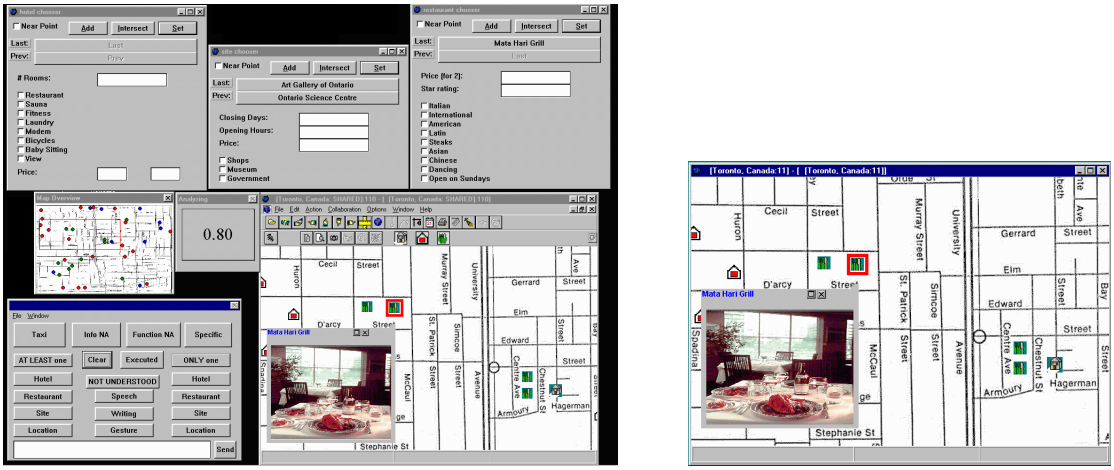

Figure 1: The Wizard Interface (left) and the Subject Interface (right)

stay, sights to see, and places to dine using speech, writing, and pen-based gestures. To first provide experience using each modality in isolation, subjects planned two half days using speech only and pen only respectively. Subjects then planned two half-days using any combination of these modalities they wished. For all tasks, the subjects were given only superficial instruction on the capabilities of the system. The tasks together took an average of approximately 35 minutes. Figure 1 depicts the subject interface and wizard interface, and Table 1 illustrates the type of interaction one finds in the data.

Entities of interest (hotels, restaurants, and tourist sites) were represented as displayed icons. Initially, the screen showed only the map; icons were subsequently displayed in response to questions about or requests to see entities of some type (e.g., "What restaurants are there in this area?"). When a particular entity was referred to (e.g., "Tell me about the museum"), its icon would become *selected* – indicated by highlighting – and previously selected icons would become unselected. Thus, at any given time during the interaction, the screen would usually show some number of icons, with one or more of these possibly highlighted.

We collected data resulting from sessions with 13 subjects. We transcribed 10 of the sessions as a training set, and the final three as a test set. The training and test sets contained 171 and 62 referring expressions respectively.

Results

Table 2 summarizes the distribution of referring expressions within information-seeking commands. (Commands to manipulate the environment, such as to scroll the screen or close a window, were not included.) Training and test data figures are shown without and within parentheses respectively. Listed on the vertical axis are the types of referential form used. The symbol ϕ denotes "empty" referring expressions corresponding to phonetically unrealized arguments to commands (e.g., the command "Information", when information is requested for a particular hotel). Full NPs are noun phrases for which interpretation does not require reference to context (e.g., "The Royal Ontario Museum"), whereas definite NPs (signalled by the determiners "the", "this", "these", or "that', with a head noun and possibly a locative) are reduced noun phrases that do (e.g., "the museum"). The horizontal axis includes two distinctions. First, we distinguish between cases in which an object was gestured to (e.g., by pointing or circling) at the time the command was issued, and cases with no such gesture. Second, we coded the cognitive status of the referents by distinguishing between selected entities, which correspond closely to Gundel et al.'s *in focus*, and unselected but visible entities, which correspond to the status *activated*. As we will see below, further distinctions proved to be unnecessary.[2]

Despite the difficulties in designing accurate reference resolution algorithms for human-human conversational data, it turned out that all of the HCI training data could be captured by a simple decision list procedure:

1. If an object is gestured to, choose that object.
2. Otherwise, if the currently selected object meets all semantic type constraints imposed by the referring expression (i.e., "the museum" requires a museum referent; bare forms such as "it" and "that" are compatible with any object), choose that object.
3. Otherwise, if there is a visible object that is semantically compatible, then choose that object (this happened three times; in each case there was only one suitable object).
4. Otherwise, a full NP (such as a proper name) was used that uniquely identified the referent.

When applied to the blind test data, this algorithm also handled all 62 referring expressions correctly.[3] Strikingly, and

[2]The table only includes cases of reference to objects. There was only one case not included, in which the subject used "here" to refer to the area represented by the entire map.

[3]When we began this project, we envisioned our data analysis leading to a more complex set of rules than the intuited rules used in the system at the time. Ironically, the resulting rules are actually *simpler*, yet do in fact exhibit greater coverage.

Form	No Gesture		Simultaneous Gesture		Total
	Unselected	Selected	Unselected	Selected	
ϕ		20 (3)	13 (3)	2 (0)	35 (6)
"it"/"they"		7 (4)		2 (0)	9 (4)
"here"			6 (0)	2 (0)	8 (0)
"there"		12 (1)		2 (0)	14 (1)
"this"		3 (2)	8 (1)	6 (1)	17 (4)
"that"		2 (2)	0 (2)		2 (4)
def NP	2 (2)	2 (0)	6 (5)	6 (1)	16 (8)
def NP *locative*	1 (2)				1 (2)
Full NP	21 (17)	35 (13)	9 (3)	4 (0)	69 (33)
TOTAL	24 (21)	81 (25)	42 (14)	24 (2)	171 (62)

Table 2: Distribution of Referring Expressions (Speech and Gesture)

in contrast to what is required for interpreting reference in HHI, this perfect accuracy was achieved despite the fact that the algorithm makes no reference to the *form* of the referring expression – pronouns, demonstratives, and lexical NPs (modulo *semantic* constraints) are all treated identically.

Cognitive Status in Multimodal Systems

Naturally, these results raise the question of why the data can be captured by a small set of rules that are insensitive to referential form. Have we stumbled onto a superior algorithm, or is some other factor at work? The answer lies within the distribution of the data in Table 2 – the central and simplifying aspect of which is that in no case was a referent that was not *in focus* (i.e., unselected) referred to with a pronoun or demonstrative without a disambiguating gesture. Instead, lexical noun phrases were used (21 full NPs and 3 definite NPs in the training data, and 17 full NPs and 4 definite NPs in the test data), and in all cases the content of this noun phrase constrained the choice to one possible referent.

While this is not a property one finds in normal modes of HHI, there is in fact a consistent explanation when one considers the role that the interface plays in these interactions. As we described in Section 2, speakers engaged in HHI must take into account their own beliefs about the hearer's (inaccessible) beliefs concerning the cognitive status of the referent, so that they can choose an appropriate expression to refer to it. The data compiled in the previous section suggests that when conversing with our multimodal interface, speakers inferred their beliefs about the computer's discourse state only from what was explicitly indicated by the (readily accessible) visual display. The display marked two cognitive statuses: selected (in focus) and unselected but visible (activated). As a graphically supplied indication of discourse state, selection is almost certainly a stronger indicator of salience than any linguistic marking afforded by language alone (e.g., placement of a noun phrase in subject or topicalized position), and thus it is unsurprising that reduced expressions are commonly used to refer to selected entities without a disambiguating gesture. Unselected referents, on the other hand, while perhaps carrying different degrees of salience in terms of the properties of the evolving discourse (for instance, some may have been mentioned previously in salient grammatical positions, whereas others appeared in less salient positions or were not previously mentioned at all), remain indistinguishable from each other with respect to their appearance on the screen. Thus, in accommodating the interface's conveyance of cognitive status, speakers could only distinguish between unselected referents by either accompanying their referential expression with a disambiguating gesture, or by choosing a fuller, uniquely-specifying definite description, both of which have the effect of greatly simplifying the interpretation process.

The data also show the effect of Grice's Maxim of Quantity in a speaker's choice of referential expression; in particular, subjects often violated the maxim in a manner which resulted in discourse that human hearers would find to be unnaturally redundant. There were 35 cases in the training data in which the selected (and thus most salient) entity was referred to using a full noun phrase, and 24 cases in which a reference to the selected entity included a gesture; in each case a pronoun unaccompanied by gesture would have sufficed. These two scenarios even overlapped in four cases, in which the selected entity was referred to with a full, unambiguous noun phrase *and* an accompanying disambiguating gesture. An analogous situation in HHI would be one in which a referent is already the topic of discussion (and thus highly salient), where the speaker nonetheless uses a full unambiguous NP *and* simultaneous gesture to refer to it again. Such a referential act would violate conversational principles to the extent that it might confuse listeners or cause them to draw unwanted implicatures. Speakers appear to be far less convinced of a computer's ability to understand natural language, however, and are thus inclined to sacrifice some degree of conversational coherence in an effort to reduce ambiguity. While perhaps not completely natural, one can see this as a fortuitous property upon which computational algorithms can (and do) capitalize in the near term. On the other hand, as users become more confident in the interpretative abilities of HCI systems, one might find an accompanying decline in the amount of redundancy employed.

The Speech-Only Experiment

Recall that in normal modes of HHI, different types of referring expression signal different cognitive statuses so they can, metaphorically speaking, "point" to different places in the hearer's mental model of the discourse. In multimodal

Form	Unselected	Selected	Total
ϕ		20 (13)	20 (13)
"it"/"they"		11 (3)	11 (3)
"there"		6 (0)	6 (0)
"this"		1 (0)	1 (0)
"that"		3 (1)	3 (1)
def NP	14 (0)	12 (3)	26 (3)
def NP *locative*	27 (11)		27 (11)
Full NP	47 (31)	32 (15)	79 (46)
TOTAL	88 (42)	85 (35)	173 (77)

Table 3: Distribution of Referring Expressions (Speech)

HCI, the need to metaphorically point is supplanted by the ability to physically point to objects on the screen, and in our experiments this conversion of modes was total.

This naturally raises the question of what behavior one finds in a speech-only setting, in which gesture is unavailable. One possible outcome is that the reference data becomes more ambiguous, and thus harder to handle, because speakers revert back to a reliance on ambiguous referring expressions to single out referents. The analysis provided in the previous section predicts a different outcome, however: Since the salience of referents within the same cognitive status are undifferentiable with respect to the interface display, speakers will use more descriptive, unambiguous noun phrases in place of reduced, ambiguous ones.

This is in fact what we find upon analyzing the data from the speech-only task, summarized in Table 3. Again, subjects used bare pronominal and demonstrative forms to refer only to selected entities. Without gesture to disambiguate reference to unselected entities, subjects used lexical NPs with uniquely specifying modifiers (such as a locative restricting reference to a single object, e.g., "The hotel on Chestnut Street") much more frequently. (In comparison, a locative modifier was used with a referential expression only three times in the entire corpus of multimodal data.)

Thus, the speech-only setting resulted not in more ambiguous forms of reference, but in less efficient reference than in the multimodal case.[4] This result provides further evidence, therefore, that speakers are accommodating what they perceive to be the system's beliefs concerning the cognitive status of referents from their prominence on the display, and tailoring their referring expressions to those. This fact also provides a potential explanation for why Gundel et al. found that definite NPs co-occurred with all cognitive statuses in their linguistic study, unlike the other forms, as described in Section 2 – such NPs may have been required to distinguish between several referents holding the same cognitive status in a given context, regardless of where on their Givenness Hierarchy the status lies.

[4]A similar result was found by Oviatt and Kuhn (1998), who point out that the need to use longer referring expressions can result in a greater number of other types of processing difficulties, such as recognition and parsing errors.

Previous Work

Space precludes a detailed summary of previous work on reference in multimodal systems, but generally speaking, much of this work has proposed methods for reference resolution without focusing on the special properties of multimodal discourse with respect to modeling discourse state and its relation to form of referring expression. A study that nonetheless warrants further discussion is due to Huls et al. (1995), who describe data from interactions with a system using a keyboard to type natural language expressions and a mouse to simulate pointing gestures. To model discourse state, they utilize Alshawi's (1987) framework, in which *context factors* (CFs) are assigned significance weights and a decay function according to which the weights decrease over time. Significance weights and decay functions are represented together via a list of the form [w_1,...,w_n,0], in which w_1 is an initial significance weight which is then decayed in accordance with the remainder of the list. Four "linguistic CFs" and three "perceptual CFs" were encoded. Linguistic CFs include weights for being in a major constituent position ([3,2,1,0]), the subject position ([2,1,0], in addition to the major constituent weight), a nested position ([1,0]), and expressing a relation ([3,2,1,0]). Perceptual CFs include whether the object is visible ([1,...,1,0]), selected ([2,...,2,0]), and indicated by a simultaneous pointing gesture ([30,1,0]).

As in our system, all referring expressions are resolved in the same manner, regardless of the type of referential form: The system simply chooses the most salient entity that meets all semantic constraints imposed by the command and the expression itself (e.g., the referent of "the file" in "close the file" must be an entity that is a file and can be closed). After developing their algorithm using several hundred constructed sentences, Huls et al. tested their framework on a set of user commands containing 125 referring expressions drawn from interactions with 5 subjects, and compared it against two baselines: selecting the most recent compatible referent, and a pencil-and-paper simulation of a focus-based algorithm derived from Grosz and Sidner (1986). They found that all 125 referring expressions were correctly resolved with their approach, 124 were resolved correctly with the Grosz and Sidner simulation, and 119 were resolved correctly with the simple recency-based strategy.

Huls et al. were thus also able to achieve perfect performance using a strategy that does not account for the differences in constraints on cognitive status imposed by different types of referring expressions. They do not, however, use this as a basis to take a deeper look into the nature of multimodal reference, given that this property of the algorithm is obviously untenable for resolving reference in normal HHI. Instead, they promote this simplification as an advantage of their algorithm, and in particular as an improvement over other methods (e.g., the algorithm derived from Grosz and Sidner) which rely on more complex sets of rules.[5] Using

[5]It should be noted that this is almost certainly an unfair comparison, as these other methods were originally developed for monomodal (i.e., speech or text only) HHI, which no doubt requires this greater complexity.

this reasoning, we could argue that the greater simplicity of our rule set renders it superior to the Huls et al. method. In actuality, however, the fact that our approach and each of those tested by Huls et al. all obtained very high accuracy supports the thesis of this paper, specifically, that the ability to achieve high accuracy is due to special properties of HCI, and not to the superior adequacy of any particular algorithm.

Nonetheless, this is not to say that these results will extend to any other multimodal HCI system; indeed, the complexity of reference behavior one finds can vary with interface design choices, domain, and task complexity. As a result, the optimal reference resolution strategy will likely also vary on a per-system basis. Other previous systems that use more complex methods for resolving reference include CUBRICON (Neal *et al.* 1988), which uses a focus space model (Sidner 1983), and ALFRESCO (Zancanaro, Stock, & Strapparava 1997), which uses a revision of the centering framework (Grosz, Joshi, & Weinstein 1995). Neither work provides a quantitative evaluation of their algorithm, nor do we have the means to determine the extent to which a simpler method, perhaps coupled with interface choices designed specifically to reduce the complexity of reference, would have provided as good or better results.

Conclusions

We have presented an analysis of a corpus of referring expressions collected from multimodal interactions which suggests that, in dramatic contrast to human-human interaction, the interpretation of referring expressions can be computed with very high accuracy using a model which pairs a highly impoverished notion of discourse state with a simple set of rules that are insensitive to the type of referring expression used. This is contrary to previous research on purely linguistic reference, in which the differences between such forms have been demonstrated to be crucial for understanding. We attributed this result to the implicit manner in which the interface conveys the system's beliefs about the operative discourse state, to which users tailor their choice of referring expression. This result therefore demonstrates one way in which a computer interface can shape the language behavior of users, a fact which can be exploited to turn ordinarily difficult interpretation problems into tractable ones.

Acknowledgements

This work was supported by National Science Foundation STIMULATE Grant IIS-9619126. This work would not have been possible without the contributions of Adam Cheyer, Luc Julia, Jean-Claude Martin, Jerry Hobbs, John Bear, and Wayne Chambliss.

References

Alshawi, H. 1987. *Memory and Context for Language Interpretation*. Cambridge University Press.

Chafe, W. L. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Li, C. N., ed., *Subject and Topic*. New York: Academic Press. 25–55.

Cheyer, A., and Julia, L. 1995. Multimodal maps: An agent-based approach. In *Proceedings of CMC95*, 103–113.

Grice, H. P. 1975. Logic and conversation. In Cole, P., and Morgan, J., eds., *Speech Acts*. New York, New York: Academic Press. 41–58.

Grosz, B., and Sidner, C. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3):175–204.

Grosz, B. J.; Joshi, A. K.; and Weinstein, S. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics* 21(2).

Gundel, J. K.; Hedberg, N.; and Zacharski, R. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2):274–307.

Huls, C.; Bos, E.; and Classen, W. 1995. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics* 21(1):59–79.

Julia, L., and Cheyer, A. 1997. Speech: a privileged modality. In *Proceedings of EUROSPEECH'97*, 103–113.

Lappin, S., and Leass, H. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4):535–561.

Martin, D.; Cheyer, A.; and Moran, D. 1999. The Open Agent Architecture: A framework for building distributed software systems. *Applied Artificial Intelligence* 13(1-2):92–128.

Neal, J. G.; Dobes, Z.; Bettinger, K. E.; and Byoun, J. S. 1988. Multi-modal references in human-computer dialogue. In *Proceedings of the 7th National Conference on Artificial Intelligence (AAAI-88)*, 819–823.

Oviatt, S., and Kuhn, K. 1998. Referential features and linguistic indirection in multimodal language. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-98)*.

Oviatt, S. 1996. Multimodal interfaces for dynamic interactive maps. In *Proceedings of CHI96*, 95–105.

Passonneau, R. 1989. Getting at discourse referents. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL-89)*, 51–59.

Prince, E. 1981. Toward a taxonomy of given-new information. In Cole, P., ed., *Radical Pragmatics*. New York, New York: Academic Press. 223–255.

Prince, E. 1992. The ZPG letter: Subjects, definiteness, and information-status. In Thompson, S., and Mann, W., eds., *Discourse Description: Diverse Analyses of a Fundraising Text*. Philadelphia/Amsterdam: John Benjamins B.V. 295–325.

Sidner, C. 1983. Focusing in the comprehension of definite anaphora. In Brady, M., and Berwick, R., eds., *Computational Models of Discourse*. MIT Press. 267–330.

Zancanaro, M.; Stock, O.; and Strapparava, C. 1997. Multimodal interaction for information access: exploiting cohesion. *Computational Intelligence* 13(7):439–464.