Reference Marking in Children's Computer-Directed Speech: An Integrated Analysis of Discourse and Gestures

Simona Montanari[+], *Serdar Yildirim*[*], *Elaine Andersen*[+], *Shrikanth Narayanan*[*+]

Integrated Media Systems Center, Speech Analysis and Interpretation Lab
[+]Department of Linguistics, [*]Department of Electrical Engineering
Viterbi School of Engineering, University of Southern California, Los Angeles
[montanar, eanderse]@usc.edu, [yildirim, shri]@sipi.usc.edu

Abstract

Understanding the fine details of children's speech and their gestural characteristics helps, among other things, in creating natural computer interfaces. We analyze reference marking in young children's computer-directed speech using audio-video data from 3- to 6-year-old children engaged in a series of age-appropriate computer tasks, using a Wizard of Oz technique. Along with speech transcriptions and acoustic information, discourse (referential devices, conversational repairs) and gestural characteristics (hand/head movement type) were annotated in a synchronized multi-layer system. The results point to the developmental variability and the multimodal nature of the speech young children produce while interacting with the computer agent, suggesting that interfaces addressed to this age group should be specifically designed to integrate multisensory information as well as to adjust to the child's specific needs and interactional style.

1. Introduction

Young children are one of the primary potential beneficiaries of computers that use conversational interfaces because even preschoolers enjoy computer games and computer instructional materials, while they lack the fine motor and literacy skills that are needed to in most computer situations [1]. Current automatic speech recognition (ASR) and natural language processing (NLP) systems are neither designed nor successful with children, especially in our targeted age group of preschool and early elementary-school aged children. A fundamental problem for these systems is the great acoustic mismatch between children speech and the models used for recognition. Previous studies have shown that children's speech exhibits higher pitch and formant frequencies, and longer segmental duration than that in adult speech [2][3]. Also, preschool children are still developing linguistic and communicative competence: they are in the process of learning the linguistic characteristics of their language as well as the rules that govern language use, such as how to express meaning in specific types of interactions [4]. This makes their speech different from adults in several important ways.

While there is a wealth of linguistic and psycholinguistic literature on the discourse skills of young children, relatively little work has been done to investigate children's interactions with computer agents. Similarly, while there is recent and on-going work on developing speech technology - especially reading applications - for school-age children, there has been very little focus on the younger preliterate age group. Applications for the younger group require a more flexible, interactive dialogue setting that allows for the interpretation of highly variable verbal and non-verbal information.

It is the aim of this study to carry out an integrated analysis of the verbal (discourse) and non-verbal (gestural) characteristics of children's computer-directed speech and to examine, in particular, how children establish reference while interacting with a machine. It has been shown repeatedly that young children have trouble taking into account the informational needs of the listener and that appropriate reference marking is not established until the late elementary school years [5]. In particular, young children have been shown to inappropriately alternate full noun phrases and pronouns to talk about new and previously mentioned referents, with the consequence that their speech is not only referentially implicit but also lacking in overall cohesion and coherence [6]. It has also been shown that conversational interactions are inherently multimodal [7] and that multimodal streams provide *complementary* in addition to converging redundant information [8]. Thus it seems plausible to ask whether young children, in establishing reference, will compensate for their impoverished verbal skills by making greater use of other modalities, for example gestures. Our audio-video data, collected using a Wizard of Oz paradigm in which children are engaged in a series of cognitive tasks while interacting with a machine, are analyzed with a synchronized multi-layer tool that allows for the simultaneous annotation and analysis of both discourse and gestural information. With the aid of this information-integrating tool, we examine young

children's ability to make use of referential devices in the production of coherent and cohesive discourse; we further explore the types of communicative non-verbal behaviors (i.e. gestures) that occur in this situation; and we investigate whether and how gestural use and reference marking are correlated.

2. Method

2.1. The Data

The corpus of data analyzed in this work is part of a larger database that is being collected to investigate child-machine spoken-language interaction. (for details, see http://sail.usc.edu/chimp). For the present study, data from ten girls and five boys aged between 3 and 6 years of age are analyzed (three 3-year-olds, and four 4-, 5-,and 6-year-olds respectively). All children come from white upper-middle class families and are native speakers of English. All participants had had some previous experience with computers prior to the experiment.

2.2. Experimental Setup

The experiment is conducted in a Wizard of Oz (WoZ) paradigm, where a hidden human agent manipulates a computer's behavior. This procedure enables careful control of experimental parameters including speech interpretation and machine response patterns, thus allowing for the collection of a variety of child-computer interactions. High quality audio recordings of the child-computer interactions are collected using a directional desk microphone. The experiments are simulataneously recorded using two Sony TRV330 digital cameras, one focused on the child's face from the front and the other capturing the child and the computer from the side.

2.3. Protocol

Each session takes approximately 30 minutes beginning with a warm-up briefing by the experimenter. This is followed by a briefing by the computer agent that parallels the human briefing. Next comes the experimental battery, which includes a set of five tasks comprised of pattern recognition, sorting and category membership. Following the five tasks, subjects are debriefed by the computer agent, and then once again by the experimenter in an analogous format.

2.4. Data Transcription and Annotation

The data from each session are organized according to section of session (i.e. initial human-child briefing, subsequent computer-child briefing, computer-child games, computer-child debriefing, and human-child debriefing). The recordings are transcribed by a native speaker of English, using a modified version of the CHILDES format [9], and they are further double-checked by a second native speaker of English. Next, the transcriptions are imported, utterance-by-utterance, into the PRAAT tool [10] to allow for a matching of the transcribed material with their acoustic counterpart. Three tiers were constructed for this purpose: (1) child, (2) interviewer, and (3) background. These three tiers are further imported into our multi-layer annotation tool to encode further verbal and non-verbal information. Our multi-track annotation board was constructed using the ANVIL tool kit [11]. Along with the speech transcriptions and acoustic information (pitch contour and intensity information), our system encodes and analyzes, in a synchronized multi-layer manner, discourse information, such as referential devices (e.g., full noun phrases vs. pronouns), conversational repairs (e.g., repetitions, clarifications, corrections), speech acts (e.g., opening, providing information, acknowledging) and pacing strategies (e.g., topic termination, anticipated response, topic shift), as well as gestural information, such as hand/head movements and movement type (e.g., touching, pointing, nodding, head shakes).

3. Results and Discussion

3.1. The Referential Analysis

The goal of our first analysis was to examine the children's use of referential devices while interacting with the computer agent in order to investigate the extent to which young preschoolers' discourse is referentially clear, and thus intelligible, in this sort of situation. For this purpose, we calculated, for each age group, the percentage of full forms (numerals, common and proper names) and of pro-forms (pronouns, deictic adverbs, deictic numerals, and deictic pronouns) over the total number of forms employed. Because a quantitative analysis does not necessarily reveal whether the children are employing referential devices appropriately, we further examined *qualitatively* the extent to which children appropriately employed full forms to introduce new referents while using pro-forms to maintain reference to previously mentioned entities. The results of the quantitative analysis are reported in Fig.1.

It appears that with age children rely more on full forms rather than pro-forms while interacting with the computer agent. This means that the 5- and 6-year-olds tend to be more explicit in their formulations than their younger peers, producing referentially clearer, and thus more coherent and cohesive, stretches of discourse. As a result, they are more successful than younger children in their interactions with a talking machine. For instance, older children appropriately alternate full noun phrases and pronouns to introduce new or talk about previously mentioned referents, thus producing referentially explicit stretches of discourse (for example, they describe a new picture as ***a boy*** *is rolling* ***a ball*** rather than *he is*

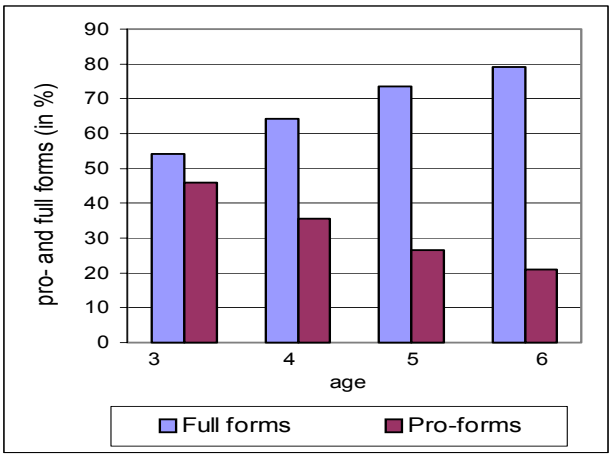

Figure 1: The mean percentage of pro- and full forms employed by the 3-, 4-, 5-, and 6-year-olds while interacting with the computer agent. The difference between the 3- and 5- and 4- and 6-year-olds' percentages is statistically significant at $p<.05$. The difference between the 3- and 6-year-olds' percentages is statistically significant at $p<.01$.

rolling it). Younger children, on the other hand, fail to take account of the informational needs of the computer agent, and use pronouns both to introduce and maintain reference to the task entities. This results in the machine's failure to identify and interpret noun-phrase antecedents or referents (compare the referentially unclear entity-introducing utterance ***he** uses **it*** with the referentially clear entity-introducing utterance ***the gardener uses the rake***). Similarly, younger children make extensive use of deictics (*here/there, this one, this/that*) rather than full noun phrases (*on top of the tree*) to describe full events, producing referentially implicit, and thus ambiguous, stretches of discourse. Deictics are often accompanied by pointing gestures; however, if the machine is not programmed to process non-verbal utterances, the interaction will necessarily fail. We will return to this point in the following section.

3.2. The Gestural Analysis

The goal of our second analysis was to examine the children's use of communicative gestures while interacting with the computer agent as well as to investigate whether there was a relationship between gestural use and reference marking. For this purpose, we calculated, for each age group, the percentage of purely verbal utterances, of purely gestural utterances, and of verbal utterances accompanied by gestures produced by the children over the total number of utterances employed; also, we calculated the percentage of touching and pointing gestures, and of nodding and head shakes over the total number of gestures produced. Finally, we examined *qualitatively* the communicative functions performed by these gestures as well as their contribution to reference marking. The results of the quantitative analysis are reported in Fig.2.

It appears that with age children rely more on purely

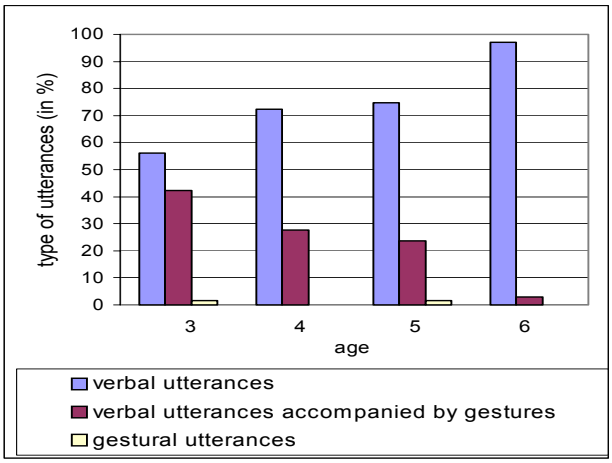

Figure 2: The mean percentage of utterance types (verbal, gestural, and verbal accompanied by gestures) employed by the 3-, 4-, 5- and 6-year-olds while interacting with the computer agent. The difference between the 3- and 6-, the 4- and 6-, and the 5- and 6-year-olds' percentages is statistically significant at $p<.05$

verbal utterances than on multimodal or merely gestural utterances. For instance, the majority of the 3- and some of the 4-year-olds describe the task entities or the location of the objects in the pictures with a combination of pro-forms and pointing gestures e.g., *he/this one/that* + pointing/ touching gesture to describe or introduce an entity; or *here/there* + pointing/touching gesture to refer to the location of an object), tacitly assuming that the machine can process both verbal and non-verbal utterances in its interpretation of the incoming speech. On the other hand, most of the 5- and 6-year-olds rely solely on words (*the blue cat or on the roof*) to make their answer explicit. Therefore, while older children display awareness of the nature of the communicative needs of the computer task and modify their speech accordingly, younger children fail to take account of the informational needs of the computer agent, producing multimodal turns that can not be interpreted solely thru the verbal modality. Because their gestures are almost exclusively deictic in nature, gestural use complements the children's implicit verbal answers performing a crucial referential function. Interestingly, we find a positive correlation between the percentage of verbal utterances and the percentage of full forms ($r(2)= 0.942$, $p<.05$), indicating that an increase in the number of verbal utterances is accompanied by an increase in the percentage of full forms. This means that the more the children rely on multimodal utterances, the more they make use of pro-forms such as deictic adverbs (*right here*) and deictic numerals (*this one*). On

the other hand, the more they rely on verbal utterances alone, (i.e. the more they become verbal), the more they tend to use full noun phrases, i.e. they tend to become more explicit in their formulations. These findings suggest that "talking computers" might be appropriate educational tools for early elementary school children, but that younger children might benefit from a combination of "talking" and "touch-screen computers." The results of our final analysis also points towards this direction: while touching gestures are most frequent among the 3-year-olds (almost 90%), the 4- and 5-year-olds show an increased preference for purely pointing gestures (33% and 50% respectively), and the 6-year-olds avoid touching and pointing gestures altogether. The majority of gestures employed by the older age group are exclusively in the form of nodding and head shakes which accompany *yes/no* utterances. Given that an explicit yes can sufficiently communicate one's intent irrespective of whether a nod accompanies it or not, there is no doubt that the older children in our experimental group have no trouble communicating with a machine that processes exclusively verbal utterances.

4. Conclusions

The results of our integrated analysis point to the complexity and great variability of the speech produced by young children while interacting with a computer agent. The participants have indeed been shown to establish reference not only through the verbal modality but also through the use of pointing and touching gestures serving the function of complementing and clarifying their implicit verbal utterances. At the same time, the children's way of marking reference has been found to vary according to age group: while preschoolers' discourse is verbally implicit but highly multimodal, early elementary school children are more explicit in their formulations and interact in a more unimodal fashion. Clearly, these findings have important implications for future research: interfaces addressed to children aged between 3 and 6 should not only be specifically designed to integrate multisensory information but they should also be programmed to adjust to the child's specific needs and interactional style. Given the benefits that children can obtain from computers that use conversational interfaces, it is important that more research be devoted to further investigating the features of child-computer spoken language interactions and to helping develop technologies that not only enable natural child-machine communication but also ensure that the child has a positive and successful experience with the system.

5. References

[1] Narayanan, S., Potamianos, A. 2002, "Creating conversational interfaces for children," IEEE Trans. Speech and Audio Processing, vol. 10, no. 2, pp. 65-78.

[2] Lee, S., Potamianos, A., and Narayanan, S. 1999, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am., vol. 105, pp. 1455-1468.

[3] Li, Q. and Russell, M. 2001. "Why is automatic recognition of children's speech difficult?" Eurospeech 2001, Scandinavia.

[4] Andersen, E. S. 1996 "A cross-cultural study of children's register knowledge." In D. Slobin, J. Gerhardt, A. Kyratzis and G. Jiansheng (eds.). Social Interaction, Social Context, and Language. Hillsdale, N.J.: Erlbaum, 125-142.

[5] Berman, R. and Slobin, D. I. (Eds.) 1994 Relating Events in Narrative: A Cross-linguistic Developmental Study. Hillsdale, N.J.: Lawrence Erlbaum.

[6] Bamberg, M. 1987. The Acquisition of Narratives: Learning to Use Language. Berlin, New York, Amsterdam: Mouton de Gruyter.

[7] McNeill, D. 1992. Hand and Mind: What Gestures Reveal About Thought. Chicago: University of Chicago Press.

[8] Oviatt, S.L. 1999. "Ten myths of multimodal interaction." Communications of the ACM, 42(11): 74-81.

[9] MacWhinney, B. 2000. The CHILDES Project: Tools for Analyzing Talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

[10] Boersma, P., and Weenink, D., "Praat Speech Processing Software," Institute of Phonetics Sciences of the University of Amsterdam. http://www.praat.org

[11] Kipp, M. 2001. "Anvil - A Generic Annotation Tool for Multimodal Dialogue" Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), pp. 1367-1370, Aalborg. http://www.dfki.de/ kipp/anvil/