Children's convergence in referring expressions to graphical objects in a speech-enabled computer game

Linda Bell[1] and Joakim Gustafson[2]
[1]TeliaSonera R&D, Sweden
[2]Department of Speech, Music and Hearing, KTH, Sweden

linda.bell@teliasonera.com, joakim.gustafson@speech.kth.se

Abstract

This paper describes an empirical study of children's spontaneous interactions with an animated character in a speech-enabled computer game. More specifically, it deals with convergence of referring expressions. 49 children were invited to play the game, which was initiated by a collaborative "put-that-there" task. In order to solve this task, the children had to refer to both physical objects and icons in a 3D environment. For physical objects, which were mostly referred to using straight-forward noun phrases, lexical convergence took place in 90% of all cases. In the case of the icons, the children were more innovative and spontaneously referred to them in many different ways. Even after being prompted by the system, lexical convergence took place for only 50% of the icons. In the cases where convergence did take place, the effect of the system's prompts were quite local, and the children quickly resorted to their original way of referring when naming new icons in later tasks.

1. Introduction

In recent years, we have seen an increased interest to use speech technology in domains that capture the interest of children and young users. Examples include speech-enabled computer games, robot pets and toys. However, some of the best practices which have evolved in the design of conventional task-oriented spoken dialogue systems must be revised when systems for children are created:

- The objective of human-computer dialogue for children is not merely to perform a predefined task in as few turns as possible. Instead, from the point of view of an immersed computer gamer, longer might be considered better.
- Applications for quiet office environments are less interesting for children, thus systems must be robust enough to handle real-life settings.
- Children's linguistic adaptations follow partly different patterns than previous research has shown for adults who engage in human-computer interaction.

In this paper, we focus on the last issue. In a study of a corpus of spontaneous child-computer interaction, we examine two types of referring expressions to see if and how children converge to the lexical patterns suggested by the animated agent who embodies this particular spoken dialogue system. We have analyzed a corpus of 6,000 utterances in which we have focused on lexical convergence in the child-computer dialogues. Two different types of graphical objects, physical objects and icons in a 3D environment, were compared from the point of view of referring expressions and lexical convergence. The first object type could be denoted by a noun phrase, while the second type of object required a more complicated adjectival phrase.

2. Background

For designers of natural language interfaces, it is a well-known problem that one object can be referred to in numerous ways. This difficulty has been called *the vocabulary problem* [1]. Brennan and Clark [2] found that the likelihood for two people to use the same term for a common object was no more than 10%. However, they also found that two people talking about the same thing often come to use the same terms, something known as *lexical entrainment*. In the development of spoken dialogue systems, the question of how to be able to predict user's lexical choices becomes a critical one, since these systems cannot be expected to handle unlimited input. Brennan [3] points to the fact that while people are as likely to adopt the terms of a spoken dialogue system as they are to adopt the terms of a human interlocutor, they may be doing if for a different reason. In spoken dialogue systems, people's *lexical convergence* to the computer is more of a uni-directional process. Most systems are not able to negotiate, and humans are often aware of this.

Recently, methods for measuring degrees of convergence at different linguistic levels in dialogue have been proposed. Reitter et al [4] discuss whether syntactic priming is a phenomenon that is limited to smaller number of rules or constructions, or whether it is more wide-spread. In their examination of the Switchboard and Map Task corpora, they find that syntactic priming effects can be found even when a greater number of syntactic rules are taken into account. However, the authors point to the fact that the task-orientedness of the dialogues may have contributed to the alignment of the speakers. Building on the method developed by Reitter et al [4], Ward and Litman [5] examine a corpus of tutoring dialogues, in which they focus on convergence of lexical and acoustic/phonetic features.

Byron argues that animated agents that are situated in an environment have to be able to handle referring expressions to its physical setting, including spatial and deictic references [6]. A previous study on a multimodal dialogue system in the apartment domain indicated that users coordinate their referring expressions to that of the animated agent [7].

It has previously been shown that children adapt their language to spoken dialogue systems, a convergence that has been shown at several linguistic levels, such as response latencies [8], prosodic range and amplitude [9], as well as acoustic and phonetic features and lexical patterns during error resolution [10]. In this study, we perform a closer examination of different types of lexical convergence in child-machine interaction, and study the effects of implicit vs. explicit system prompts.

3. The child-machine corpus

3.1. The NICE system

The goal of the NICE system was to build a speech-enabled computer game, consisting of a 3D world with fairy-tale characters that were inspired by the works of author H.C. Andersen. The system was intended for children who would use spoken dialogue as the main vehicle for story progression. The arguments for adding speech to computer games include enabling users to engage in social discourse and negotiation, as well as making it possible to refer to past events and objects currently not visible on the screen. Most commonly, however, users will refer to what is currently on the screen. In our collaborative "put-that-there" task, children refer to two types of graphical objects, which are the focus of this study.

All children were set-up with a headset microphone and were informed that they could speak with the system's animated helper character Cloddy Hans using spontaneous speech. They were also given a mouse with which they could point at, but not directly manipulate, objects. In the initial scene with Cloddy Hans, the children were introduced to the animated agent as he greeted them in the author H.C. Andersen's study. In the study, there is a shelf with fairy-tale objects, including a sword, a magical lamp and a diamond, see Figure 1. On the other side of the study stands a fairy-tale machine, which will construct a fairy-tale as soon as it is full of objects. The objects are to be put in machine-slots with different functions that are indicated by visual icons placed above them, see Figure 1. The task at hand is thus for the user to ask Cloddy Hans to pick up different objects and place them in the appropriate slots in the machine, according to how they will be used in the fairy-tale they are building. In the process, the animated agent tries to get the children to agree with him on how to use two different types of referring expressions to graphical objects in the 3D world:

(1) The physical objects on the shelf, Here, Cloddy Hans uses noun phrases such as "*the emerald*" or "*the knife*"
(2) The slots in the fairy-tale machine, each of which is labeled with a visual icon. Here, Cloddy Hans suggests using adjective phrases like "*valuable*" or "*magical*".

Figure 1 *(Left) Cloddy Hans and the shelf with objects (Right) The slots in the fairy-tale machine from left to right;* ***useful, dangerous, magical, valuable***

3.2. The NICE corpus

The NICE corpus has been previously described in Bell et al [11]. The corpus consists of Swedish spoken dialogues between children and embodied fairytale characters in a computer game scenario. Children interacted with the system in supervised mode at different stages of development, some of them only with the helper character Cloddy Hans and others in later stages also with a second gate keeper character, Karen. For the purposes of the current study, we focused on the dialogues with the Cloddy Hans character and could thus use data from all collections. A total of 49 children between the ages of 9 and 12 interacted with the animated agent, and the dialogues typically went on for 20-40 minutes. The total number of dialogue turns examined in this study was about 6,000 and around half of these were user utterances. The number of physical objects placed in the fairy-tale machine ranged from one to five, with an average of three. All in all, the number of physical objects discussed in the child-machine dialogues was 169. These objects were referred to by Cloddy Hans 550 times and by our subjects 625 times. The equivalent figures for the slots in the fairy-tale machine were 135, with 246 references from the system and 273 from the children, respectively. An excerpt from a typical interaction, where Cloddy Hans attempts to influence a 9-year old boy's way of referring to the slots, can be seen in Table 1.

Table 1. *Example from the NICE corpus, translated from Swedish. Tags given for reference type.*

User	Eh put that lamp in *the pipe with the skull*	Concrete
Cloddy	Do you want me to put eh the lamp in *dangerous*	Function
User	Nah actually put the lamp in eh put that lamp in *that hole eh pipe where the jewels are*	Concrete
Cloddy	Do you want me to put the lamp in *valuable*	Function
User	Yeah and put it in *that jewelery thing*	Function+ slot
Cloddy	I will put the lamp in *valuable*	Function
User	Okay put the lamp in *valuable* oops did you do that already?	Function **convergence**
User	Put the sword in *the pipe with the skull*	Concrete

4. Data analysis and results

4.1. Objects on the shelf

In 87% of all cases, an object was first verbally referred to by the user. In the remaining cases, the children failed to select a new object, in which case Cloddy took the initiative by suggesting that a certain object should be picked up.

Following Ward and Litman [5], we wanted to identify objects for which there was only one lexical choice available. These objects were then removed, as no synonyms meant lexical convergence was not possible. Instead of using WordNet or a dictionary to find synonyms, which would make sense for adult subjects, we decided to remove those objects for which not a single of our children subjects came up with an alternative noun to the one in the system's lexicon. As can be seen in Table 2 below, this meant taking out five of the objects for which the children had not mentioned any additional lexical items apart from the one used by the system.

Table 2 *Number of different noun phrases used for each of the objects on the shelf. The objects on the top row were not given any synonyms and were subsequently removed.*

Axe, hammer, key, knife, sword	1
Sack, magical wand	2
Poison flask, ruby	3
Magical book, diamond	4
Magical lamp, emerald	5

As can be seen in Table 2, there was little disagreement on how to refer to the objects on the shelf. Without being prompted, all our subjects used noun phrases to describe these objects. Figure 5 below shows that in around 75% of all

cases, the children's preferred way of referring to the objects coincided with the system's first choice. Here we had removed the five objects without synonyms, and counted only objects that both dialogue parties had verbally described. In the cases where there were several variants, these were often based on a small number of different head nouns, sometimes modified with an adjective ("the diamond", "the jewel", "the white diamond", "the white stone"). After Cloddy Hans had referred to the object, we found a failure to converge to the system in no more than 10% of all cases. This can also be seen in Figure 5 below.

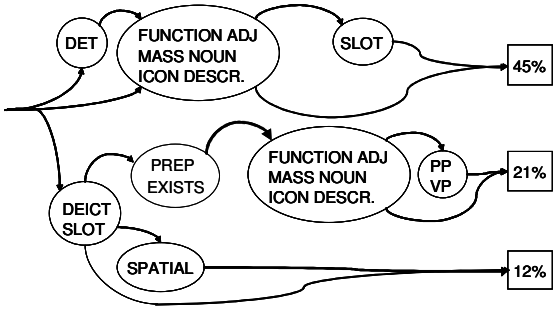
4.2. Slots in the fairy-tale machine

There was greater variation in the children's references to the slots with the icons in the fairy-tale machine. Indeed, our subjects were imaginative and sometimes appeared to want to say everything but the target word intended, even when Cloddy Hans explicitly prompted them: ("*Could we call it **dangerous**?*").

Table 3 *Number of different constructions used for the slots in the fairy-tale machine. Deictic expressions ("this one") were grouped into one category*

Useful	Dangerous	Magical	Valuable
45	16	28	42

On the surface level, the variation in the referring expressions used for the slots seems overwhelming, and would present a serious challenge for a spoken dialogue system to handle. However, an in-depth analysis showed that 78% of the expressions can be described according to the following syntactic constructions:



DETERMINER	det / *definite article*
DEICTIC	det där / *that one*
SLOT	röret / *the pipe*
PREP	vid, med / *at, by*
EXISTS	där det är / *where there is*
MASS NOUN	verktyg / *tools*
FUNCTION ADJECTIVE	magiskt / *magical*
ICON DESCRIPTION	en hammare och en nyckel *a hammer and a key*
VERB **P**HRASE	är (bild) / *is (picture)*
PREPOSTIONAL **P**HRASE	på (bild) (över) / *on (picture) (above)*
SPATIAL	Längst bort till höger/ *on the far right*

Figure 2 A *syntactic description of the children's references to slots in the fairy-tale machine*

The network in Figure 2 results in sentences like "*Put it in the slot with a hammer and key on picture above*", "*Put it the one on the far right*" or "*Put the axe in magical*". These ways of constructing references can be categorized into the main types: *Visual*, *Functional* and *Anaphorical*, which can be further divided as follows:

- **Visual** references
  - **Concrete** descriptions ("*the skeleton hole*", "*the slot where there is a picture of a hammer and a key*")
  - **Spatial** description ("*the one on the far right*", "*the next to last pipe*")
- **Functional** references
  - **Functional** adjective/noun ("*valuable*", "*tools*")
  - **Function+slot** Adjective noun compounds and phrases ("*the magical slot*", "*the death pipe*")
- **Anaphorical** references
  - **Deictic** combined with click ("*here*", "*in this one*")
  - **Discourse** reference ("*where we put the magic wand*")

These six sub-categories were used to tag the children's references to the slots in the fairy-tale machine, while solving the collaborative "put-that-there"-task. Utterances labelled as "before" were the ones where the subjects had not yet heard Cloddy's reference to the slot, while lexical convergence could be measured following the references labelled "after".

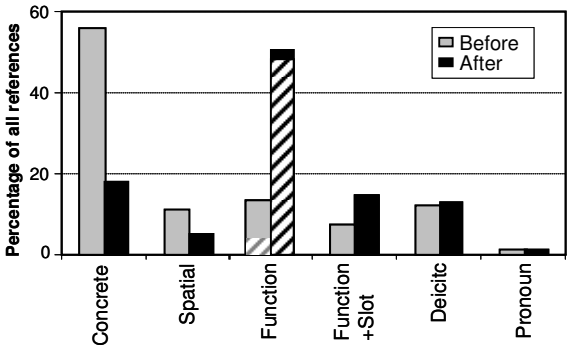


Figure 3 *Distribution of references to the slots in the different categories, before and after being mentioned by Cloddy Hans*

Figure 3 shows that the children preferred to refer to the slots by means of visual references, most often a concrete description such as "*the skeleton hole*". By explicitly prompting them with a functional reference ("*dangerous*"), more than half of the subjects converged to this way of referring in their subsequent reference to the same slot. The striped pattern on the function bars in Figure 3 indicates the share of these where the child used exactly the same word as Cloddy Hans. When the first object had been successfully placed in the appropriate slot, the same task would be repeated up to five times.

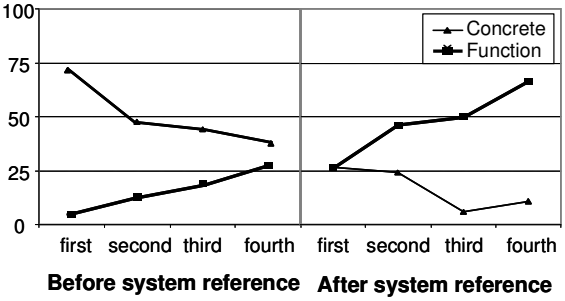


Figure 4 *Percentage of referring expressions labeled as "concrete" vs. "function" in the children's first four tasks.*

As can be seen in Figure 4, the convergence effect for the slots in the fairy-tale machine appeared to be local, and most subjects resorted to their preferred way of referring as they moved on to the second, third and fourth slots. Although the children often failed to match Cloddy Hans' precise way of referring to the slot, there is a trend towards more and more functional references as the dialogue progresses. However, even after three explicit prompts with functional references, no more than 25% of the children verbally converged to this way of referring to a new slot.

4.3. Lexical convergence effects for graphical objects

When referring expressions in each of the two groups (objects and slots) were compared, we could observe a difference in convergence effects. As can be seen in Figure 5, we failed to obtain lexical convergence for only 10% of the physical objects but for more than half of the slots with icons. 40% of the convergence failures for slots can be explained by the fact that Cloddy Hans guided the children through the process by explicitly prompting them and allowed them to say "yes" to complete the task: (User: "*Put the axe in **the skeleton hole***", Cloddy: "*Do you want me to put the axe in **dangerous***", User: "*yes*").

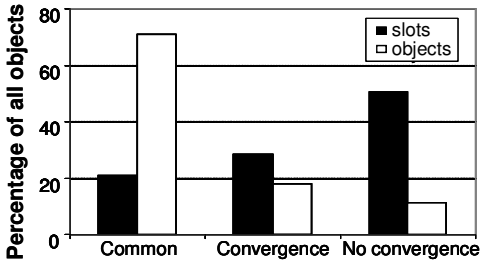


Figure 5 *'Common' refers to objects for which the user and system spontaneously used the same term. All objects which both parties referred explicitly to were included.*

5. Discussion

This study has shown that children's referring expressions to physical objects converge towards the system's choice of words. When referring to the slots with icons, children adopted the lexical items previously used by the system in half of the cases. Moreover, they fail to generalize and come up with a new referring expression of the same type as they refer to new icons. One reason for this might be that it is easier for the children to describe what they see than to come up with a referring expression that describes the meaning of a new icon. However, if this is the case you might question whether the children had understood the purpose of the game, which was to sort the objects according to their role in the fairy-tale that is created once the machine is full.

In post-experimental interviews, most children reported that they found it fun and natural to use speech in the NICE game. They also said that they expected games to be like this in the future [11]. Nonetheless, many challenges remain before we can build collaborative conversational computer games which handle unconstrained spontaneous speech. By adding speech understanding to high quality animated characters in computer games, it is likely that we increase the children's expectations to an unrealistic level. However, carefully designing the system's personas to reflect limited understanding capabilities is one way to mitigate this effect. An advantageous feature of computer games is that the designer has full control of the 3D environment as well as the tasks at hand. This can be used to simplify the understanding of the users' verbal input, thus making speech-enabled computer games feasible.

Boye et al [12] argue that in order to interpret referring expressions in the computer game domain, the system has to be able to keep track of the visual context (all objects visible on the screen) as well as past events (all earlier actions related to objects). As long as speech is used to refer to physical, concrete objects in the 3D world, our results indicated that high levels of lexical convergence can be achieved. This can partly be explained by the fact that there exists a high degree of agreement among the children on how to refer to these objects, even before they have been mentioned by the system. If we are to add more complex objects such as icons, our results suggest that it is necessary for an interpretation system to have knowledge about these objects' functional, visual and spatial properties. Even when a system's understanding module makes use of these properties, nonetheless, the variability in the spoken input will make it more difficult to handle references to such objects. One way of solving this might be to change the animated agent's way of referring to the icons, so that it uses deictic expressions instead and encourages the subjects to do the same ("*Do you mean **this one*** (Cloddy Hans pointing at the icon)?" "*No*" "*Could you **point at** the one you meant?*").

6. Acknowledgements

This work was carried out within the EU-funded project NICE (IST-2001-3529, http://www.niceproject.com).

7. References

[1] Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. The vocabulary problem in human-system communication. Communications of the ACM 30 (11): 964-971, 1987.

[2] Brennan, S.E. and Clark, H.H. Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory and Cognition 22(6):1482-1493, 1996.

[3] Brennan, S.E. Lexical entrainment in spontaneous dialog. Proceedings of ISSD: 41-44, 1996.

[4] Reitter, D., Moore, J. and Keller, F. Priming of Syntactic Rules in Task-Oriented Dialogue and Spontaneous Conversation. In Ron Sun, ed., Proceedings of the 28th Annual Conference of the Cognitive Science Society, 685-690, 2006.

[5] Ward, A. and Litman, D.:Measuring Convergence and Priming in Tutorial Dialog. Technical report TR-07-148 2007.

[6] Donna K. Byron. Understanding referring expressions in situated language: Some challenges for real-world agents. In Proceedings of the First International Workshop on Language Understanding and Agents for the Real World., pp 80-87, 2003.

[7] Skantze, G. Coordination of referring expressions in multimodal human-computer dialogue. In Proceedings of ICSLP 2002 (pp. 553-556). Denver, Colorado, USA, 2002.

[8] Darves, C. and Oviatt, S. Adaptation of users' spoken dialogue patterns in a conversational interface. In Proc. of ICSLP, 2002

[9] Coulston, R., Oviatt, S. and Darves, C. Amplitude convergence in children's conversational speech with animated personas. In Proceedings of ICSLP, 2002.

[10] Bell, L. and Gustafson, J. Child and adult speaker adaptation during error resolution in a publicly available spoken dialogue system. In Proceedings of Eurospeech, 2003.

[11] Bell, L., Boye, J., Gustafson, J., Heldner, M., Lindström, A. and Wirén, M. The Swedish NICE Corpus – Spoken dialogues between children and embodied characters in a computer game scenario, In Proceedings of Interspeech, 2005.

[12] Boye, J., Mats Wiren, M., and Gustafson, J. Contextual Reasoning in Multimodal Dialogue Systems: Two Case Studies, Proceedings of Catalogue'04 , Barcelona, 2004.