*NICE project (IST-2001-35293)*

# Natural Interactive Communication for Edutainment

# NICE Deliverable D2.2b

# Collection and analysis of multimodal speech and gesture data in the first fairy-tale prototype

*22 April 2004*

*Authors*

*Joakim Gustafson, Johan Boye, Linda Bell, Mats Wirén*

*TeliaSonera AB*

*Jean-Claude Martin, Stéphanie Buisine, Sarkis Abrilian*

*LIMSI-CNRS, Orsay, France*

| | |
|---|---|
| **Project ref. no.** | IST-2001-35293 |
| **Project acronym** | NICE |
| **Deliverable status** | |
| **Contractual date of delivery** | |
| **Actual date of delivery** | 22 April 2004 |
| **Deliverable number** | D2.2b |
| **Deliverable title** | Collection and analysis of multimodal speech and gesture data in the first fairy-tale protype |
| **Nature** | Report |
| **Status & version** | |
| **Number of pages** | |
| **WP contributing to the deliverable** | Wp2 |
| **WP / Task responsible** | |
| **Editor** | |
| **Author(s)** | Joakim Gustafson, Johan Boye, Linda Bell, Mats Wirén, Jean-Claude Martin, Stéphanie Buisine, Sarkis Abrilian |
| **EC Project Officer** | Mats Ljungqvist |
| **Keywords** | |
| **Abstract (for dissemination)** | |

# Table of contents

# 1 Introduction

## 1.1 Scope of this report

This report presents an analysis of a corpus of multimodal speech and gesture data collected with the first prototype of the NICE fairy-tale system. Ten children have used the system, upon which they have been interviewed and asked to fill in a questionnaire. The transcribed user input data, together with the questionnaires, constitute the basis for this analysis.

## 1.2 Structure

The report is structured as follows. Section 2 describes the data collection method and some quantitative measures of the corpus. Section 3 discusses turn-taking, section 4 various adaptation phenomena, and section 5 discusses the implications for the second NICE fairy-tale prototype. Section 6, finally, gives an analysis of the gestural input.

# 2 Data collection and corpus

A corpus of spontaneous spoken human-computer dialogues between Cloddy Hans and ten children between the ages of 11 and 15 was collected at the Telecom museum in Stockholm. The corpus consists of sound files, multimodal input files and log files from the different system components. The sound files where transcribed and tagged, for details see section 2.3.

## 2.1 Method

Users were random visitors who were recruited by the experimental leader while exploring other parts of the Telecom museum. Before being presented to Cloddy Hans, users were asked to fill out a questionnaire with some basic background data. These included their age and gender as well as previous experience with computers in general and animated agents and speech technology in particular. Once seated in front of the large screen, users were given a brief introduction to the system by the experimental leader, who explained how Cloddy Hans could be addressed and made sure the cordless microphone and gyro mouse were in order. Users were also informed that they were going to be recorded, and that their sound files would later be transcribed and analyzed with the purpose of improving the system's performance. Subsequently, they were presented with the task at hand through a pre-recorded set of instructions, accompanied by three still pictures (see Figure 1). For a more thorough presentation of the system set-up, see Deliverable 7.2b.

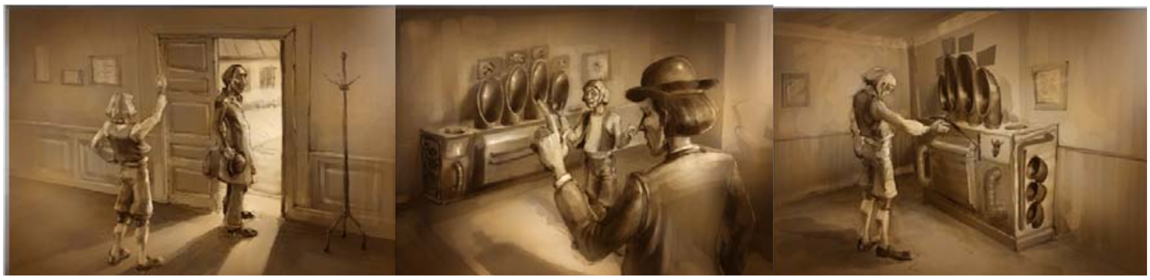

*Figure 1. The illustrations shown to the subjects during the verbal scene description.*

## 2.2 The subjects

The 10 users were between 11 and 15 years old, their ages and genders are shown in Figure 2 together with their answers to the background questions. As can be seen all subjects reported that they had much experience using computers for accessing the Internet, playing games or both.

| **subject** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **gender** | female | female | male | female | female | male | male | male | female | female |
| **age** | 14 | 14 | 15 | 14 | 14 | 15 | 14 | 12 | 11 | 11 |
| **Internet** | much | much | much | much | much | much | much | much | little | little |
| **Mail** | much | much | much | little | much | much | much | much | little | no |
| **Computer Games** | little | little | much | much | much | much | much | much | much | much |
| **Animated Agents** | no | no | - | much | no | little | little | little | little | no |
| **Voice Command Sys.** | no | no | little | little | little | no | little | no | no | no |
| **HC Andersen** | no | yes | yes | yes | no | no | no | no | no | no |

*Figure 2. An overview of the subjects and their background data.*

## 2.3 Spoken language input

Speech data was recorded with a speech recognizer with a bi-gram grammar trained on the task (see further deliverable 7.2b). All speech files were then manually transcribed and analyzed by linguistically trained labellers. Filled pauses were transcribed using the following six tags:

- Mmm
- Mhm
- Ehh
- Öhh
- Ehm
- Öhm

Furthermore, unintelligible speech, self-directed speech, laughter and words truncations were given separate tags to improve the accuracy of the analysis.

## 2.4 Corpus statistics

The corpus consists of 569 user utterances and 176 user gestures[1], varying from 36 to 89 utterances per subject, and from 1 to 45 gestures per subject. Figure 3 shows the number of utterances and gestures broken down for each subject.

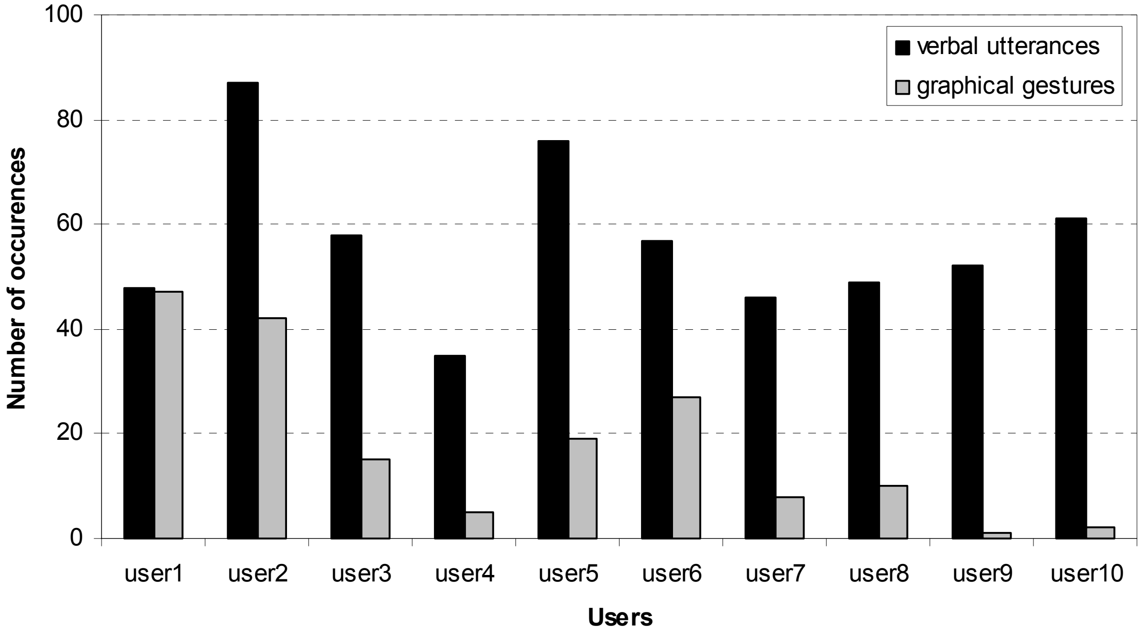

*Figure 3. Number of utterances and gestures (for definition of graphical gestures, see Section 5)*

[1] In Section 3, we discuss the precise meaning of the term ‘utterance’. Section 6 explains the term ‘gesture’ in detail.

The number of words per utterance varied from 1 to 18 words, with a total average number of words per utterance on 3.6. The average number of words per utterance varied per user, from 2.1 to 5.6. Figure 4 shows the average number of words, broken down for each subject.

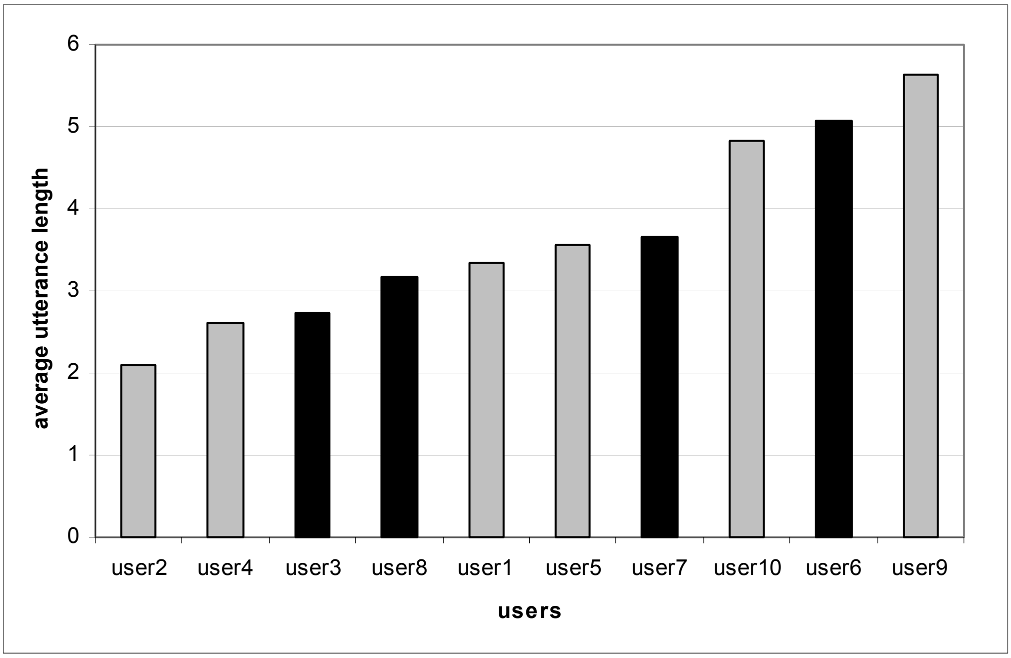

*Figure 4. Average utterance length (male subjects represented by black bars, female by grey bars)*

As can be seen the average utterance lengths are quite short. This is partly explained by the dialogue design choice to often make Cloddy ask for clarification before performing the actions the users asked for. As can be seen in Figure 4 there is no difference in utterance length between girls and boys.

## 2.5 Dialogue example

The following excerpt from the dialogue with user 9, an 11-year old girl, is provided for illustration. For the convenience of the reader, all transcribed interactions in this report have been translated to English. Actions are written inside square brackets [], and graphical gestures are written inside curly brackets {}. In order to make the transcript more readable, a lot of small physical actions performed by Cloddy Hans have been omitted (like nods, blinks, shrugs, eye and eyebrow movements, etc). Such small gestures are generated automatically as a result of various events from the speech recognizer (start-of-speech, rejected input, speech-too-early, and timeout. See further deliverable 7.2b). Instansesof overlapped speech between the subject and Cloddy are shown as comments in italics inside parentheses, and truncated words are marked with asterisks.

When the dialogue starts, Cloddy Hans is seen standing next to a shelf with objects (a sword, a diamond, a magic wand, small figurines portraying a prince, a princess, Thumbelina etc.). These objects can be dropped into various slots in HC Andersen's fairy-tale machine which is seen in the background. For a more thorough scenario description, see deliverable 7.2b.

In Figure 5 below, the utterances, gestures and actions have been grouped into *turns* (dialogue contributions). The difficulties associated with carrying out this grouping in a systematic way are discussed in the next section. An ellipsis (…) means that a part of the dialogue has been left out.

| **User9** | **Cloddy Hans** |
|---|---|
| What is your name? | My name is Cloddy Hans, actually. What is your name? |
| Linnea | How old are you? |
| *... a number of turns later* | |
| Let’s take that sword. | Do you mean the sword? |
| Then there will be some action!<br>Don’t you agree? | OK.<br>[Picks up the sword]<br>OK, done. |
| Come on, let’s hurry and put it in the machine! | [The camera angle is changed, and Cloddy Hans is seen from a distance]<br>[Cloddy Hans starts walking towards the machine]<br>What do you want me to do?<br>[The camera angle is changed, and Cloddy Hans is seen standing next to the machine]<br>Should I drop it in the slot for dangerous things? |
| Yes do that; that sounds fine. | OK.<br>[Cloddy Hans turns around and faces the machine]<br>[He drops the sword in the “dangerous” slot. The sign above the slot trembles a little]<br>[Cloddy Hans turns around and faces the camera] |
| Can we look for more things? | OK.<br>[The camera angle changes again, and Cloddy Hans starts walking towards the shelf with objects]<br>[The camera angle changes, and Cloddy Hans is seen standing next to the shelf]<br>Can you point at the diamond? |
| {Clicks on the diamond} | Should I pick up the diamond? |
| Yes, do that. | OK.<br>[Picks up the diamond]<br>That went fine! |
| *... a number of turns later* | |
| [The camera is now showing a close-up on the figurines in the shelf] | |
| We could take one of these couples.<br>{Clicks on the figurine portraying the prince}<br>We could put them in too. | Shall I pick up the prince? |
| Yes, let’s do that. | OK<br>[The camera now shows Cloddy Hans picking up the little figurine. Then he faces the camera]<br>What do you want me to do? |
| The prince is rather rich so he could …(inaudible) | Do you think so? |
| (inaudible) | Shall I put the prince in the slot for good persons? |
| Yes, that’s fine. | … |

*Figure 5. A dialogue excerpt from a dialogue between Cloddy and an eleven-year-old girl.*

# 3 Turn-taking

The standard model of turn-taking in many simpler spoken-dialogue applications is based on the assumption that system utterances and user utterances will proceed in alternation. Hence, for such systems no distinction between "utterance" and "turn" is necessary; the two concepts can be equated. By contrast, in the NICE fairy-tale system, a "turn" is typically a complex object made up from an arbitrary number of smaller building blocks: utterances, graphical gestures, and animated actions. These building blocks may occur consecutively or overlapping in time. Adding to the complexity is also the fact that turns may themselves be overlapping; the user might talk, for instance, when Cloddy Hans is moving from one place to another as the result of a previous user request.

In our parlance, we will define a *user utterance* to be a piece of spoken input returned as a single result from the speech recognizer. The recognizer's end-of-speech detector was set on 500ms, which means that a user utterance consists of the words from the moment the user starts speaking, up until the moment he pauses for 500ms or more. A *user turn* consists of a sequence of such user utterances together with all accompanying gestures, or (in the case of gesture-only input) only a sequence of gestures. A *system turn* consists of a sequence of synthesized system utterances, together with all the accompanying animations as well as performing physical actions. It is not always obvious how to group utterances and gestures into turns, as discussed in Section 3.1.

The 569 user utterances, 176 user gestures, 560 system utterances and 380 system animation scripts[2] in the corpus were grouped into 515 user turns and 551 Cloddy-Hans turns. 439 of the user turns were verbal-only, 11 gestural-only, and 65 were multimodal. Figure 6 below shows the breakdown for individual users.

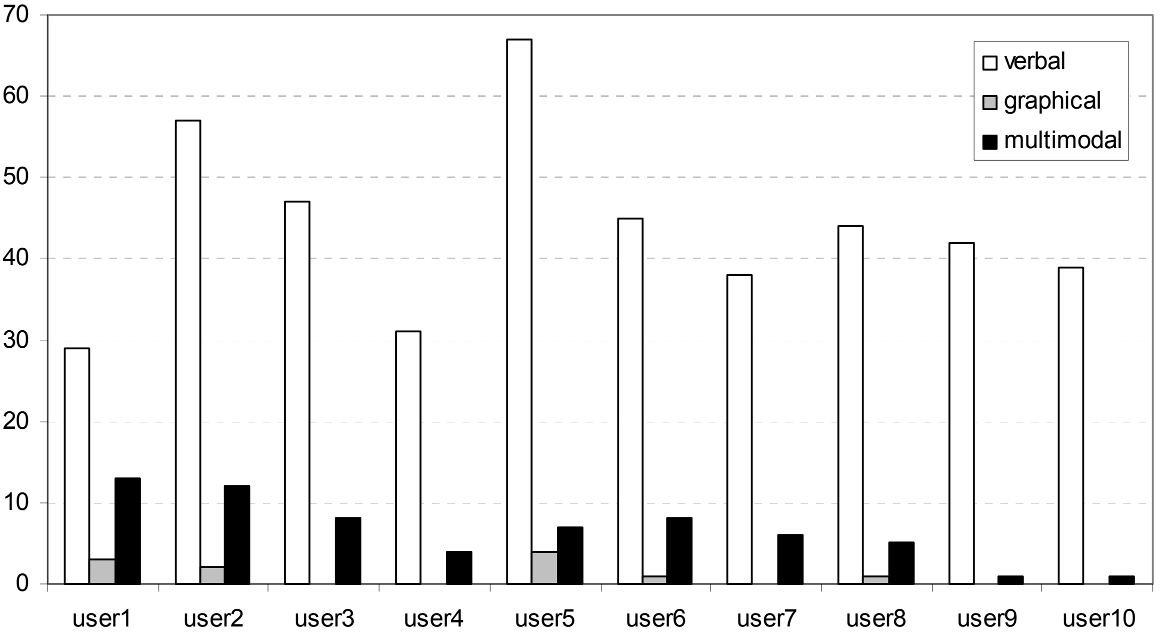

*Figure 6. Number of user turns per subject.*

[2] An animation script was a sequence of animation instructions sent to the animation system. Each such script corresponded to a physical action performed by Cloddy Hans, either "go to", "pick up", put down" or "point at". In addition to these scripts, a huge number of animation instructions for communicative gestures were also sent to the animation system, as well as instructions for changing the camera angle.

The turn segmentation underlying the above diagram was done post-mortem using a human expert. Further analysis of this data will reveal whether our previously developed algorithm for real-time turn segmentation, outlined in Bell et al (2001), can be adapted to this domain.[3]

Figure 7 below shows the number of utterances per turn, broken down for individual users. Note that some users have an utterance-per-turn ratio below 1; this is of course due to the fact that some turns were purely gestural. As can be seen the number of multi-utterance turns is quite user dependent ranging from 0.98 to 1.5 utterances per turn.

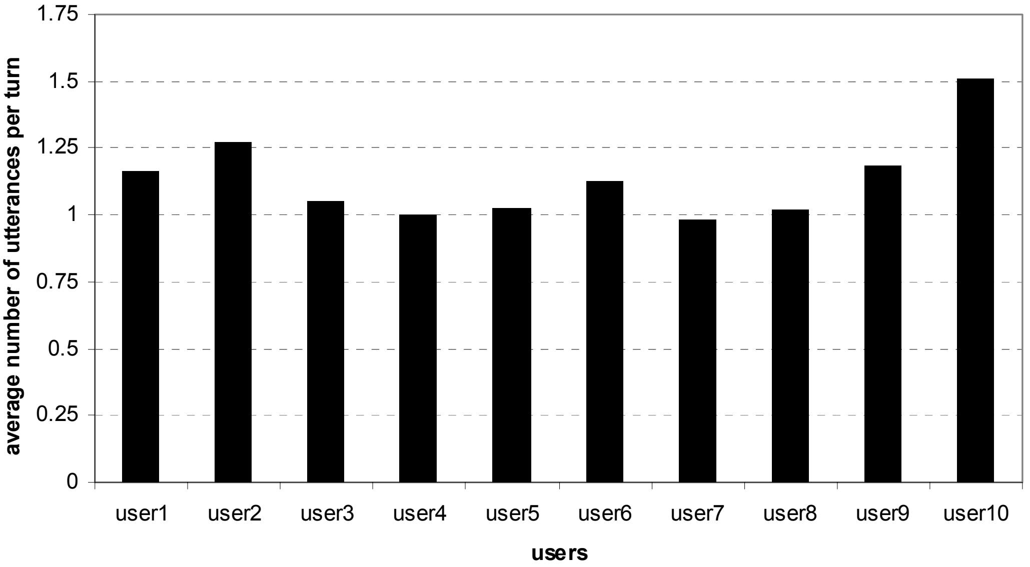

*Figure 7. The average number of utterances per turn for the individual users.*

Figure 8 below shows the distribution of various types of utterances. We distinguish between the following categories

- ***Social,*** e.g. "How are you?"
- ***Meta utterances,*** e.g. "What can you do?"
- ***Yn,*** answers to yes-no questions, e.g. "Yes do that; that sounds fine."
- ***Task,*** e.g. "Take the hammer"

[3] As described in deliverable 7.2b, the system used in the data collection was semi-automatic, allowing a human operator to choose one of several system replies. The system replies were automatically generated by various sub-systems, but the real-time turn segmentation was done using human intervention.

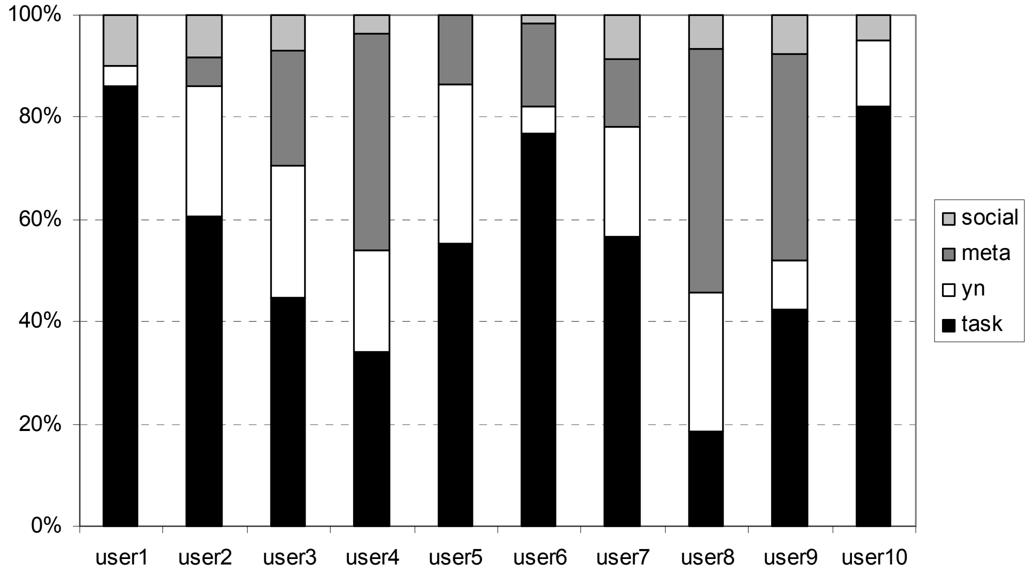

*Figure 8. Categorization of user utterances and their shares of all utterances per user.*

As can be seen in Figure 8, the utterance type distribution is highly user-dependent, with the share of task-oriented utterances ranging from 18% to 83%. Figure 9 shows an excerpt from a dialogue with one of the most task-oriented users, an eleven-year-old girl. Below, utterances within the same turn are separated by "…", and "*" denotes a truncation.

| User 10 | Cloddy Hans |
|---|---|
| How are you? | I'm fine, thanks for asking … how are you? |
| I'm fine…that book on the shelf…take it and put it in* {pointing at the valuable slot} | Do you want me to take the magical book? |
| And then you put it in the slot with a hat and a magical wand. | Do you want me to go to the shelf?<br>[walks over to the shelf] |
| Pick up the magical book* | Do you want me to pick up the magical book? *(overlapped with user utterance)* |
| Yes. | [picks up the book]<br>What do you want me to do? |
| I want you to put the book in the slot with…a hat and a magical wand. | [walks over to the machine]<br>What should I do now? |
| Put the book in the slot with hat and a magical wand. | Could we call it 'magical'? |
| Yes. | Mmm. |
| Put* *(overlapped with Cloddy's action)* | [Puts the book in the magical slot]<br>Yes that went well! … What do you want me to do? |
| Then you can…I want you to go to the shelf. | Ok. [walks over to the shelf]<br>What should I do now? |
| I want you to pick up the sack. | Could you point at the money bag? |
| {points at the money bag} | Is that the money bag?…Do you want me to pick up the bag? |
| Yes, pick up the bag. | [picks up the bag]<br>Yes that went well! |
| Then I want you to…Then I want you to go to the slots | Ok, [walks over to the machine] |

*Figure 9. A dialogue between Cloddy and a task-oriented eleven-year-old girl.*

A general tendency was that only few user utterances were of socializing nature (0-10%). This is different our previous dialogue corpora collected of human-animated character interactions (Gustafson and Bell 2001, Bell 2003). This is probably explained by the task-oriented nature of the scene, and by the fact that all users got a verbal description of the scene that introduced Cloddy Hans (which meant that they did not have to ask Cloddy about himself). The number of yn-answers (3-26%) can be seen, among other things, as an indicator of the degree of system initiative in the dialogues. The number of meta utterances (0-40) is an indication of how well the dialogue went, and how uncertain the users were. User 8 was an uncertain, and uncooperative twelve-year-old boy, who was very unwilling to take the initiative in the dialogue. Therefore only 18% of his utterances were task oriented. The first part of his dialogue with Cloddy is shown in Figure 10.

| **User 8** | **Cloddy Hans** |
|---|---|
| Hi. | Hi there. |
| Hi again. | What should I do? |
| What did you say? | What do you want me to do? |
| I don’t know, want do you want to do? | Should we go to the shelf? |
|  | [walks over to the shelf] What should I do now? |
| I don’t know. | Should I pick up the sword? |
| What, what did you say? {clicks on the sword} | Do you want me to pick up the sword? |
| {clicks on the sword} | Should I pick up the sword? |
| No. | What do you want me to do? |
| I don’t know, I thought you should tell m* me that. | *Gives an introduction to the machine plot, the machine and the shelf of objects* |
| Alright.. | *Introduces the task and that the objects have to be taken from the shelf and put into the slots* |
| Alright.. | Do you want me to pick up the sword? |
| Yes, do that! | [picks up the sword]<br>How amusing! |
| Yes, very funny (sarcastic prosody). | Do you want me to go to the machine? |
| Yes, you might as well do that | [walks over to the machine]<br>What should I do? |
| Well, I don’t know. | I could put the sword in the machine. |
| Are you a bit slow today?*(overlapped wi.th Cloddy)*<br>Yes you could do that | [puts the sword into the dangerous slot] |
| {clicks on lamp} | Do you want me to go back to the shelf? |
| Yes, LAUGH | Ok [walks over to the shelf]<br>What do you want me to do? |
| Well, you could pick up…the lamp or something. | Ok [picks up the magic lamp]<br>What should I do now? |
| Put it into the machine. | Do you want me to go to the machine? |
| Well I don’t care, but ok then. | [walks over to the machine} |

*Figure 10. A dialogue between Cloddy and an uncooperative twelve-year-old boy.*

The 551 Cloddy-Hans turns were partitioned as follows:

- verbal-only (339 turns)
- verbal + action (134 turns)
- action-only (78 turns

54 Cloddy-Hans turns were not initiated by user input, but initiatives from the system that aimed at helping uncertain users. These can be further categorized as follows:

- plot description (4 instances)
- suggesting names for objects, e.g. "Could it be the axe?" (5 instances)
- suggesting places where to put an object, e.g. "Shall I put the axe in the slot for dangerous things?" (6 instances)
- suggesting an object to pick up, e.g. "Should I pick up the axe?" (7 instances)
- suggesting a location where to go, e.g. "Should I go to the machine?" (13 instances)
- indications of problems, e.g. "I don't understand what you want me to do" (19 instances)

## 3.1 Multimodal behaviour

Figure 11 below presents an overview how users made use of the gestural input channel. We distinguish between the following categories:

- ***Multimodal***
    - ***gestural reference + verbal accept***. The user accepts a suggestion from the system by saying "yes", or something to the same effect, and pointing at an object.
    - ***gestural reference + verbal correction***. The user verbally rejects a suggestion from the system, and points at another object instead.
    - ***gestural reference + verbal deictic pronoun***. The user says a deictic phrase (like "this one") and points at an object.
    - ***gestural reference + verbal redundant reference***. The user issues a request and backs it up with a pointing gesture (e.g. "take the knife" while pointing at the knife).
    - ***gestural reference + verbal contradicting reference***. The user gives contradictory information in the two input channels (e.g. "take the knife" while pointing at the axe).
- ***Unimodal***
    - ***Gestural reference***. The user points at an object and says nothing.
    - ***Gestural reference on demand***. The user points at an object (and says nothing) when Cloddy Hans has asked him to point at a certain object.
    - ***Doodle***. The user makes gestures with the pointing device without any *obvious* communicative intention.

Every multimodal or gestural dialogue contribution was assigned to exactly one category. Note that the typology suggested above is tentative, and can most likely be improved.

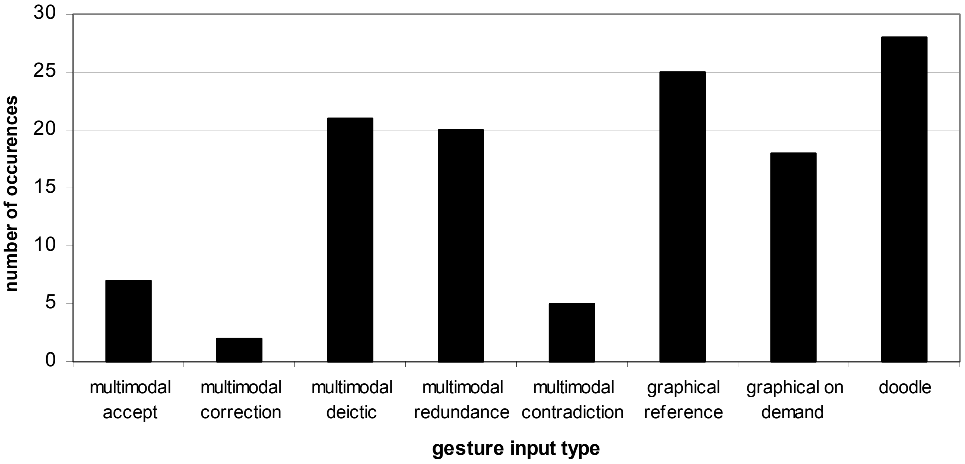

*Figure 11. Categorization of gestural and multimodal input*

One of the purposes of the initial scene was to see if it is possible to make the subjects use both verbal and gestural means of expressing themselves. Therefore all subjects were asked to point at a certain object a total of 19 times in the current corpus, and as can be seen in the figure, the subjects were very obedient and actually did a gestural reference in 18 times of these cases. The only subject who did not follow the request only did a total of 3 gestural references, out of which one was after a request from the system. This means that all subjects were compliant and gestured when Cloddy asked them to.

## 3.2 Problems associated with turn-taking

There are a number of known problems associated with asynchrony and turn-taking (see e.g. Thompson 1996, Traum and Heeman 1997, Nakano et al 1999, Boye et al 2000, and Bell et al 2001), some of which are the following:

**Segmentation problem -** Which utterances should be grouped together in the same turn?

**Multiple-click problem -** To refer to an object, the user clicks on it multiple times. The problem here is that the multiple clicks are to be considered as a single communicative act of referring.

**Doodling problem -** The user keeps making gestures with the pointing device without any obvious communicative intention.

**Own-communication problem -** The user talks to himself and not to the system.

**Asynchrony problem -** The user talks or clicks while the system is in the process of performing some action

Several of these issues (segmentation, doodling, asynchrony) were also observed in the preliminary 2D embodied agents Woz done at Limsi at the beginning of the NICE project (Buisine et al. 2003) with a similar but simple setting.

In the corpus discussed in this report, instances of the two latter problems are almost non-existent. The segmentation problem has already been discussed. As for multiple clicks and doodling, these phenomena seem to have large individual variations. The doodling phenomenon is indicated by Figure 12 below.

| | user1 | user2 | user3 | user4 | user5 | user6 | user7 | user8 | user9 | user10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Doodle** | 0<br>*(0%)* | 2<br>*(9%)* | 5<br>*(33%)* | 9<br>*(78%)* | 1<br>*(9%)* | 4<br>*(33%)* | 4<br>*(33%)* | 4<br>*(40%)* | 0<br>*(0%)* | 0<br>*(0%)* |

*Figure 12. The absolute number of doodle gestures per user and its share of all their gestures.*

At first sight, doodling in the present corpus seemed to be associated with some users, in particular user 4, a fourteen-year old girl. However, when analyzing that interaction it is apparent that all instances of doodling occur in the initial part of the dialogue. The user does not speak at all in the first 11 turns, even though the system desperately tries to help her (see Figure 13). In the interview with this user after the interaction, she told the experimental leader that she did not talk to the system at the beginning because she felt embarrassed talking, not because she did not know how to solve the task. However, once she started talking she thought the game was fun. After the session, her answer to the question "what did you like most about the game" she wrote "that he could talk back!". Her overall judgment was that the system as a whole was "good". The words this subject selected to describe Cloddy Hans included: *slow, robot-like, boring, does what I ask him to, helpful, slow, sweet*, and the words she selected to describe the game were: *fun, easy, too short, a new experience.* This means that, in her opinion, Cloddy's limited linguistic capabilities did not make the game boring. She enjoyed talking to him and wanted to continue at the end. She thought that the task was a little too easy, but would she have liked to enter the fairy-tale world.

| **Cloddy Hans** | **User4** |
|---|---|
| Hello | - |
| *Gives a verbal description of the plot* | - |
| What do you want me to do? | {clicks on money sack} |
| Do you want me to go to the shelf? | {clicks on money sack} |
| What did you say? | {clicks on useful slot} |
| Could you say that in another way? | {clicks on magic slot} |
| Do you hear me? | - |
| Don't you want to talk to me? | {clicks on money sack} |
| Hello? | {clicks on the machine} |
| Hello? | - |
| Do you want me to go to the shelf? | - |
| Are you able to talk? | Pick up the bag! |
| Do you want me to pick up the sack? | Hoho?*(overlapped with Cloddy)* Yes! {clicks on money sack} |
| Should I go to the shelf? | Yes! |
| [Walks over to the shelf] | Stop! |
| Should I take the sack? | Pick up the sack!*(overlapped with Cloddy's utterance)*<br>Yes! {clicks on money sack} |
| [picks up the money bag]<br>What do you want me to do now? | Put it in the slot with* |
| Do you want me to go to the machine? | Yes. |
| [walks over to the machine]<br>What do you want me to do? | Put the sack into the slot with the skull. |
| Could you repeat that? | Hoho? *(overlapped with Cloddy's utterance)*<br>Put the sack into the slot with the skull. |
| Could we call that dangerous? | Yes. |
| Do you want me to put it in dangerous? | Yes. |
| [puts the sach into the dangerous slot]<br>What do you want me to do? | Go to the shelf. |
| Do yo want me to go to the shelf? | Go then! |
| [walks to the shelf] | Pick up the book! |
| Is it this one? [points at the book] | Pick up the book! |
| Do you want me to pick up the book? | Pick up the sword! |
| Do you want me to pick up the book?<br>[picks up the book] | [OCM-he refuses, stupid] *(overlapped with Cloddy's action)*<br>LAUGH, go to the machine |

*Figure 13. A dialogue between Cloddy and an initially 'talk-shy' fourteen-year-old girl.*

# 4 Adaptation phenomena

In this section we will discuss how users changed their behavior during the course of the dialogue, as they realized how to communicate with Cloddy Hans in a successful way.

## 4.1 Lexical adaptation

It is a known fact that, in human-machine dialogues, users tend to pick up the words and syntactic constructs used by the system. Constructors of dialogue systems can take advantage of this *convergence* effect, as it is sometimes called, by letting the system be very explicit about how to refer to key concepts in the domain. This will increase the probability that the users will express themselves in a way the system will be able to understand.

In the scenario implemented in the first fairy-tale prototype, we had anticipated that users would have difficulties in naming the objects on the shelf, and the slots on the fairy-tale machine. Especially the second hypothesis turned out to be true. As the reader might recall (see deliverable 7.2, section 2), the fairy-tale machine had slots for "dangerous things", "valuables", "magic things" and "useful things". One strategy adopted by users was to refer to the slot by referring to the picture above it, as in the following dialogue excerpt:

**User 3** Drop the sword in the skull.

**Cloddy** Could we call that slot 'dangerous'?

**U** Yes.

…

**U** Take the knife.

…

**U** Drop it in the 'dangerous' slot.

Here the user immediately picked up the term 'dangerous' to refer to the slot in question, and ceased to use the word 'skull' (which was referring to the picture above). When the situation arises again (9 turns later) that the user wants to drop an object in the same slot, he uses the term 'dangerous', which was proposed by Cloddy Hans.

Some users directly confirmed that they had accepted the proposed term:

**User 1** Pick up the bag.

**Cloddy** Could it be the money sachet?

**U** Pick up the money sachet.

Other users were more resilient to this kind of priming. For instance, one user (user 6) kept referring to the slots by coordinates ("first slot on the right", "third slot on the left", etc.) despite Cloddy Hans's consistent use of the terms 'dangerous', 'useful', etc. On the other hand, all users picked up the terms "fairy-tale machine", "the shelf", which were used all the time by Cloddy Hans and users alike.

## 4.2 Adaptation to the task

We could observe that all users except one accepted the scenario, and tried to play the game to the best of their ability. Most users were a bit unsure of how to proceed at the beginning, but then they quickly grasped the overall structure (make Cloddy Hans go to the shelf and pick up an object, then make him go to the machine and drop it one of the slots). This learning phenomenon is matched by an improvement in speech recognition rates, as follows: Each dialogue can be divided into an *introductory phase*, where the user familiarizes himself with the system, usually with some socializing and meta utterances, followed by a *task phase*, where the user tries to make Cloddy Hans pick up objects and put them in the fairy-tale machine (the task phase has been assumed to begin when the user issues the first task-oriented instruction to Cloddy Hans, or starts asking questions about the task). Figure 14 below shows the speech recognition word error rates for the two dialogue phases, broken down per user.

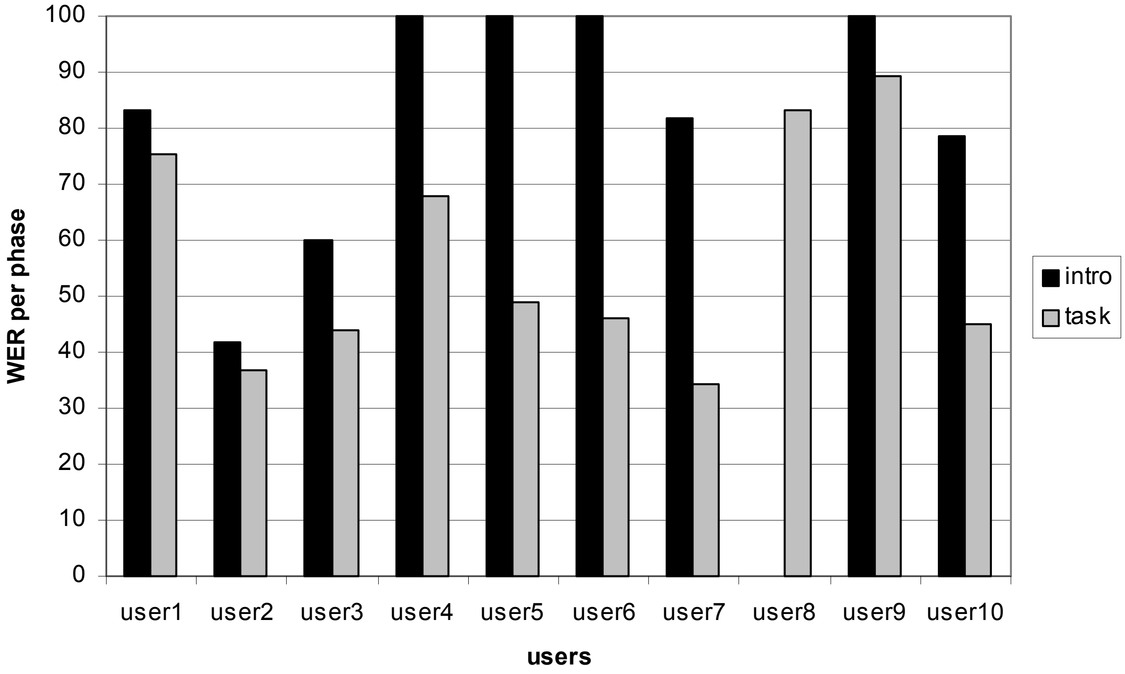

*Figure 14. Word error rates for different dialogue phases (the dialogue with user 8 had no intro phase)*

Of course it is difficult to draw any firm conclusions from such figures, as they rely on the success of a particular recognizer with a particular grammar. But if anything can be said on the basis on these numbers, it is surely this: Once the users have grasped the nature of the *task* at hand, then it is easier for them to understand what *linguistic patterns* they may use in order to communicate successfully with the system.

# 5 Discussion and pointer to 2$^{nd}$ fairy tale prototype

The first scene used in the first prototype was deliberately designed to be simple, task-oriented and repetitive. The real purpose is not to solve the task, but to engage in a collaborative grounding conversation, where the user and Cloddy Hans have to agree on what the fairy-tale objects can be used for, and how to refer to them. This process lets the players find out (by trial-and-error) how to adapt in order to make it easier for the Cloddy Hans to understand them, e.g. by using multimodal input in certain contexts. Hopefully, this will make the interaction smoother in the subsequent scene in the fairy-tale world (prototype 2), since the objects and characters that appear in it already have been grounded in the initial scene.

The current study has verified the fact that users adapt to the behavior of the system to a high degree. This phenomenon is of great help when constructing a dialogue system of this kind. We found that it is possible to make users use gestural references, and that users often adapt their verbal references to match the ones Cloddy uses. Users were able to grasp the way they have to divide a task into smaller units in order for Cloddy to perform them one by one.

The collected interactions also indicate that it is a good thing to have a task-oriented initial scenario that lets the users familiarize themselves with the new input possibilities (note that all users thought that using a speech-controlled game was “a new experience”). Some users had problems getting started, and needed to be guided by the system. Once users started using speech, they were able to solve the task and thought it was fun, but they wanted a more interesting and fun task, which is was we want to give them in the subsequent interactions in the fairy-tale world. Many users started by exploring the interface (clicking around to see what would happen, trying to say different things - mainly socializing and meta utterances). This entails that the system needs to be robust to such input, and gently but firmly get the game going, by giving suggestions on what to do. To start a speech enabled game with a simple straightforward task that is not part of the overall story of the game is therefore a good idea (cf the initial tutorial introductions to the available commands in computer games such as Harry Potter, Black&White and Star Wars-knights of the old republic).

Finally, the collected interactions showed that it is essential to have an advanced turn handling component. Such a component has to be able to cope with turns with

- multiple utterances (due to unfilled pauses)
- multimodal turns that consist of both utterances and gestural input,
- user utterances that are overlapped with system turns (where the system instantly must be able to distinguish between back-channeling utterances and barge-in utterances).

The turn handling component must interact with the internal planner of the animated character in order to handle user turns that appear while the character is in the process of carrying out a physical action. The last topic will become more critical to solve in the subsequent plots in the fairy-tale world, where actions will take considerably more time to carry out (e.g. “walk to the village” may take more than a minute. While Cloddy Hans is walking to the village, the user might want to talk to Cloddy (who should continue walking while engaging in dialogue). In some cases the user might see something that he wants Cloddy to pick up before he resumes his last action (going to the village). Yet another possibility is that the user might change his mind and asks Cloddy Hans to go to the mill instead (so Cloddy Hans will have to cancel the ongoing action and start a new one instead).

# 6 Gestural input analysis

## 6.1 Introduction

As mentioned in section 2, user tests on FTW version of PT1 were conducted at Telia with children and teenagers (6 girls, 4 boys, 11 to 15 years old). The gestural input device was a gyro mouse. The system included LIMSI's GR (Gesture Recognizer) and GI (Gesture Interpreter) modules, whose outputs have been logged by Telia. The present section describes the data LIMSI could extract from GR and GI log files.

## 6.2 Log files analysis

Twenty log files (2 per user), covering about 2 hours of gestural input, were analysed. LIMSI followed the same analysis process as for HCA version of PT1 (see deliverable D7.2a, section 2).

| User | Age | Gender | Duration of log files (min) | Number of GR frames | Number of GI frames |
|---|---|---|---|---|---|
| U1 | 14 | Fem | 14 | 94 | 95 |
| U2 | 14 | Fem | 20 | 78 | 79 |
| U3 | 15 | Male | 16 | 31 | 31 |
| U4 | 14 | Fem | 9 | 20 | 21 |
| U5 | 14 | Fem | 20 | 24 | 24 |
| U6 | 15 | Male | 11 | 33 | 34 |
| U7 | 14 | Male | 13 | 36 | 37 |
| U8 | 12 | Male | 17 | 22 | 24 |
| U9 | 11 | Fem | 1 | 2 | 2 |
| U10 | 11 | Fem | 2 | 3 | 3 |
| **TOTAL** | | | **123** | **343** | **350** |

*Figure 15. Data extracted from GR and GI log files.*

Figure 15 presents the individual and gestural data from Telia's users. As for HCA log analysis, the slight difference between the number of GR and GI frames indicates some system or log failure.

On an average, 2.8 GR frames were produced every minute (this rate is the same if U9 and U10, who have very short log files, are excluded from the data). The raw number of GR frames does not seem to be influenced by the age or by the gender of users.

Figure 16 presents the shapes of movements produced as $1^{st}$ best results in GR module. A large majority of gestures (75%) were points. The percentages of circles and lines are respectively 13% and 12%.

| **User** | **Age** | **Gender** | **Nb of points** | **Nb of circles** | **Nb of vertical lines** | **Nb of horizontal lines** | **Nb of diagonal-up lines** | **Nb of diagonal-down lines** | **Nb of other shapes** | **TOTAL** |
|---|---|---|---|---|---|---|---|---|---|---|
| U1 | 14 | Fem | 91 | 2 | 1 | 0 | 0 | 0 | 0 | **94** |
| U2 | 14 | Fem | 64 | 9 | 2 | 3 | 0 | 0 | 0 | **78** |
| U3 | 15 | Male | 19 | 6 | 1 | 4 | 0 | 1 | 0 | **31** |
| U4 | 14 | Fem | 12 | 4 | 2 | 2 | 0 | 0 | 0 | **20** |
| U5 | 14 | Fem | 21 | 1 | 1 | 1 | 0 | 0 | 0 | **24** |
| U6 | 15 | Male | 7 | 14 | 1 | 7 | 2 | 2 | 0 | **33** |
| U7 | 14 | Male | 29 | 3 | 3 | 1 | 0 | 0 | 0 | **36** |
| U8 | 12 | Male | 10 | 6 | 1 | 2 | 1 | 2 | 0 | **22** |
| U9 | 11 | Fem | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **2** |
| U10 | 11 | Fem | 3 | 0 | 0 | 0 | 0 | 0 | 0 | **3** |
| **TOTAL** | | | **258** | **45** | **12** | **20** | **3** | **5** | **0** | 343 |
| **%** | | | **75** | **13** | **3** | **6** | **1** | **1** | **0** | |

*Figure 16. Shapes of movements extracted from GR log files.*

Finally, we could list the 1st best objects logged in GI module (Figure 17).

| 1st best object | TOTAL | % |
|---|---|---|
| Dangerous Slot | 43 | 12 |
| Null | 38 | 11 |
| Gold Sack | 37 | 11 |
| Magic Book | 27 | 8 |
| Useful Slot | 27 | 8 |
| Object Shelf | 24 | 7 |
| Woz Machine | 20 | 6 |
| Magic Lamp | 18 | 5 |
| Magic Slot | 17 | 5 |
| Tummelisa Doll | 17 | 5 |
| Valuable Slot | 17 | 5 |
| Sword | 16 | 5 |
| Hammer | 12 | 3 |
| Poorgirl Doll | 7 | 2 |
| Good Slot | 6 | 2 |
| Jewel | 5 | 1 |
| Prince Doll | 5 | 1 |
| Princess Doll | 5 | 1 |
| Axe | 3 | 1 |
| Magic Wand | 3 | 1 |
| Knife | 2 | 1 |
| Poison | 1 | 0 |
| **TOTAL** | **350** | **100** |

*Figure 17. List of 1st best objects logged in GI.*

Since Telia's user tests did not use LIMSI's IF module because of substantial modifications and extensions to Telia's NLU after IF delivery and before user testing, LIMSI were not able to extract data about user's multimodal behaviour.

## 6.3 Comparison between HCA and FTW gestural data

LIMSI worked out the same kind of data from files logged in HCA and FTW prototypes; it is thus possible to make a few comparisons, although limited since the two settings and instructions were different (speech-only phase in HCA, scenario length...).. First of all, users in FTW gestured more than users in HCA study (F(1/22) = 10.76; p = .003 on the frequency of gestures). The mean frequency of gestures in FTW was 2.8 per minute whereas in HCA study there were only 1.1 gestures per minute (1.5 with the mouse and 0.8 with the tactile screen). This difference could be due either to the game scenario or to the input device. We assume that the scenario, conversational in HCA study and rather object-oriented in FTW, was of prior importance to the amount of gestural behaviour.

Concerning the shape of movements, the percentage of pointing gestures was higher in FTW (75%) than in HCA study (44%; F(1/22) = 5.44; p = .029). We may hypothesise that the input device is responsible for this effect. Since the gyro mouse is an unusual device and was used with a large display at a certain distance, pointing gestures may have been the easiest way to accurately select objects, especially small ones. Figure 18 shows a comparison of shapes of gestures according to the input device.

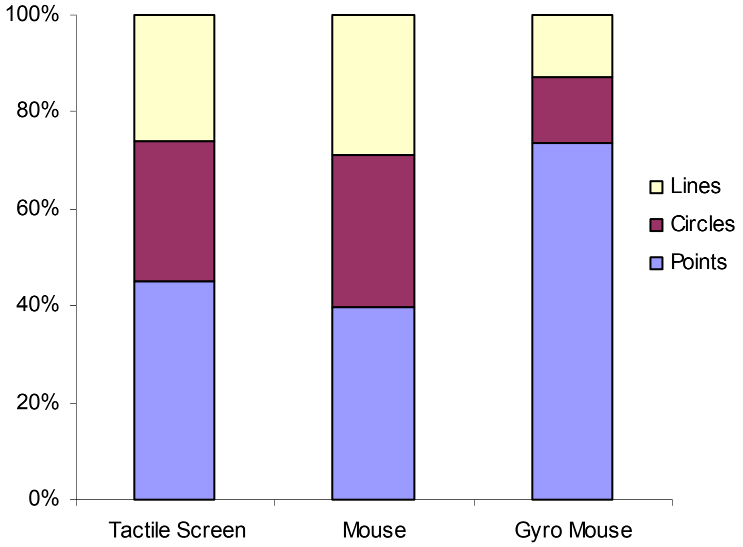

*Figure 18. Percentage of lines, circles and points as a function of the input device. function of the input device (tactile screen and mouse for HCA, gyro mouse for FTW).*

Finally, it is worth mentioning that the proportion of gestures to non-referrable objects was much lower in FTW (11%) than in HCA study (57%). This effect may be due to the object-oriented scenario of FTW, because users were prompted to refer to certain objects. Exploration must have been lower in this environment.

# 7 References

Gustafson, J, and Bell, L (2000) "Speech Technology on Trial: Experiences from the August System", *Journal of Natural Language Engineering: Special issue on Best Practice in Spoken Dialogue Systems*.

Bell, L (2003): "Linguistic adaptations in spoken human-computer dialogues – Empirical studies of user behavior", PhD thesis KTH.

Thompson, H. S. (1996). Why 'turn-taking' is the wrong way to analyse dialogue: Empirical and theoretical flaws. *Proc. 1996 international symposium on spoken dialogue.*

Traum D., and P. Heeman (1997). Utterance units in spoken dialogue. In Maier et al (eds.); *Processing in spoken language systems*, pp 125-140.

Nakano, M., Miyazaki, N., Hirasawa, J., Dohsaka, K., and T. Kawabata (1999). Understanding unsegmented user utterances in real-time spoken dialogue systems. *Proc. ACL'99*, pp. 200-207.

Boye, J., Hockey, B. and M. Rayner. Asynchronous dialogue management – two case studies. In *Proc. Götalog, 4th workshop on the pragmatics and semantics of dialogue*, pp. 51-55.

Bell, L., Boye, J. and J. Gustafson (2001). Real-time handling of fragmented utterances. In *Proc. NAACL workshop on adaptation in spoken dialogue systems*.

Buisine, S. and Martin, J.-C. (2003). Experimental Evaluation of Bi-directional Multimodal Interaction with Conversational Agents. *Proceedings of the the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'2003),* Zürich, Switzerland, September 1-5, IOS Press, 168-175. 1-58603-363-8.
http://www.limsi.fr/Individu/martin/research/articles/Interact03.pdf