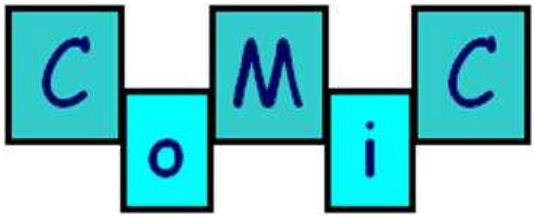
Authors:
Louis Vuurpijl, Stéphane Rossignol, Norbert Pfleger, Louis ten Bosch, Ralf Engel, Don Willems, Lou Boves

Date: 18th October 2004

# COMIC Deliverable 3.3

# Report on the T28 Human Factors Experiments with Simultaneous Coordinated Speech and Pen Input and Fusion




## Document history

| Date | Editor | Explanation | Status |
|---|---|---|---|
| Aug ’04 | Vuurpijl | set up first draft of t28specs.tex | (internal draft) |
| Aug 26 ’04 | Pfleger | added first turn-taking description | (internal draft) |
| Sep 10 | Vuurpijl *et al* | specification and design | final specs |
| Sep-Oct ’04 | KUN+DFKI | running experiments | |
| Sep 21 ’04 | Vuurpijl+Rossignol *et al* | first analyzes | draft v0.3 |
| Oct ’04 | KUN+DFKI | analyzes and report | version v0.4 |
| Oct 15 ’04 | KUN+DFKI | finishing report | version v0.5 |
| Oct 18 ’04 | all authors | finalizing report | final version |

# COMIC

*Information sheet issued with Public COMIC Document 3.3*

| | |
|---|---|
| *Title:* | Report on the T28 Human Factors Experiments with Simultaneous Coordinated Speech and Pen Input and Fusion |
| *Abstract:* | This document reports on the T28 experiments on multi-modal interaction in bathroom design |
| *Author:* | Louis Vuurpijl, Stéphane Rossignol, Norbert Pfleger, Louis ten Bosch, Ralf Engel, Don Willems, Lou Boves |
| *Reviewers:* | Lou Boves |
| *Project:* | COMIC |
| *Project number:* | IST-2001-32311 |
| *Date:* | 18th October 2004 |
| *Distribution list* | |
| *COMIC partners:* | All |
| *External COMIC:* | PO, Reviewers |
| | Public |
| *Key words:* | Human Factors experiments; multi-modal system design and evaluation |

# Contents

# Chapter 1

# Introduction

This document describes the T28 experiments on “simultaneous coordinated speech and pen input and fusion”, performed as a collaborative experiment by WP3 and WP4 of COMIC. The T28 experiments are part of a series of human factors experiments. In these experiments, several goals are being pursued:

1. exploration of multi-modal interaction patterns in situations of varying complexity;
2. development of increasingly more capable recognition technology, required to be able to reliably interpret and process these interaction patterns;
3. assessment of the usability of the developed technologies;
4. incorporation of the results in the evolving COMIC demonstrators.

This design process has proven to yield successful multi-modal input recognition systems, as reported in [2, 3, 4, 5, 8, 9]. The experiments described in this document are considered as an important step toward the development of the required input recognition technology, natural language processing (NLP) and fusion for the T30 [1] demonstrator and beyond.

## 1.1 The focus of the current research

The main goals of WP3 and WP4 are to develop usable, natural multi-modal interaction technology. The outcomes from the previous ”T16” human factors experiments showed that the appreciation of the systems is proportional to the recognition performance of the input modules. To enhance recognition performance in the T24 system, a system-driven dialog strategy helped to overcome limitations in recognition performance by (i) narrowing down the number of expected inputs, (ii) implementing a strict turn-taking protocol and (iii) requesting for simple, atomic, information items.

At the same time, users reported that they felt restricted by this interaction strategy. Therefore, in the design of the T28 research platform, the three design considerations have been adapted to construct a system that contains more capable recognition technologies, and consequently can provide a more natural, user-driven, interaction strategy. Furthermore, a new experimental design was developed that provides a means for performing interactive experiments with a working system while refraining from an operational dialog action manager. As a result, four modifications have been implemented in the current T28 research platform:

1. *new “isolated” interactive experimental design: focus on input technology and fusion.* Based on the experiences obtained from the design of the previous human factors experiments, we have decided that — given the resources — the current experiments could not make use of a dialog action manager that controls a dialog between the system and a user. Since the focus of WP3 and WP4 is targeted at recognition and fusion technologies, we have developed a new experimental interaction design that provides

subjects with a certain situation on the screen that has to be modified, thereby evoking specific user input. Upon receiving some user input in response to such a request for modifications, the system processes the input and generates a system response that is assessed by the user. In this way, multi-modal interactions, system responses, and user acceptance of the responses can be recorded and analyzed.

2. *compound multi-modal information: users will be allowed to input multiple information items in one turn.* An important finding from the previous human factors experiments [4] was that users found it particularly annoying that they had to wait for the system to come up with requests for small pieces of information. As opposed to such *atomic* information items, (like a single wall, a single size (of a wall or a window), or a single door), at T30 it will be possible for users to enter so-called *compound* information items. Compound information may for example contain the drawing of two walls (by pen), accompanied by their corresponding sizes (by speech), all in one turn.

3. *less constrained turn taking: end of turn detection is made more intelligent*. The strict system-driven turn taking protocol employed in T24 has been modified into a mechanism that allows users to enter information in a more natural way. Basically, this has resulted in a system in which the microphone (for entering speech) and tablet (for entering pen input) are open as long as the input recognizers and the Fusion module think that the user is still providing information. In this new mechanism, the user can continue providing information as long as (s)he wants, after the system has relinquished the floor. Only after it has become evident that the user no longer intends to produce input, the system will take over the turn, interpret the input, yield a response, and immediately after that will allow the user to enter information again.

4. *multi-modal specifications of spatial relations: users can move objects by entering relative or absolute spatial references.* In order to address the concerns of the reviewers that COMIC should investigate conversational treatment of spatial relations, in the current experiments tasks are included that invite subjects to use a combination of absolute and relative expressions to indicate spatial relations. We have implemented this issue by providing subjects with the option to modify parts of the screen state by moving windows and doors. Key to this option is the possibility for the user to employ absolute and relative spatial relations in an intuitive way to (re-)arrange objects.

In preparing the system for the T28 HF experiments, the pen input and speech recognition technologies, the natural language processing, and the multi-modal fusion modules have all undergone significant improvements to implement these modifications. The focus of the current research is in the assessment of the usability and effectiveness of these modifications. In the next section, the associated research questions are discussed in more detail.

## 1.2 Research goals

It has been the focus of the current study to assess the usability of the developed technologies through human factors studies and to explore whether the implemented computational models of (expected) multi-modal interaction patterns correspond to the actual behaviour of human subjects as observed in the experiments. More specific, the following research goals are pursued in this report.

### 1.2.1 How does the new experimental interaction design perform?

The experiments described in this report have been designed with the goal to evoke specific multi-modal interaction patterns in the context of compound objects, spatial relations and erasing gestures. This made it possible to conduct interactive experiments without requiring the development of novel dialog action management, fission, and user-interface technology. As will be described in detail in Chapter 5, this was implemented by providing subjects with a certain situation on the screen that had to be modified in a certain manner. The assumption that subjects understand what they have to do in response to a request for multi-modal information

will be validated. Moreover, an exploration of the resulting user input should indicate whether the goal of acquiring relevant interaction patterns was reached.

### 1.2.2 Do users learn from the interaction?

In the current study, system guidance is minimal. Subjects are merely presented with a sequence of stimuli, containing requests for information that trigger multi-modal responses. One particular human factor that is examined in this experiment is whether users can learn from the interaction. Since stimuli are grouped in four categories and the first group of stimuli requests relatively simple information, the assumption is that by presenting one half of the subjects with stimuli in sequential order and the other half with stimuli in random order, it can be observed whether the former subjects make less errors (use less turns) and therefore, whether experience with increasingly complex (starting with simple stimuli) multi-modal utterances benefits the interaction. This result would be an important finding, as COMIC strives for natural interactions with novice users. If users are not able to interact multi-modally without (small) guidance, given the state of the art in recognition technology, this means that an appropriate instruction (explaining what they can or should do) or a brief training session is required.

### 1.2.3 What multi-modal interaction patterns do users employ?

The new concepts of compound objects and modifying situations by means of expressing spatial relations will result in new observations of multi-modal interaction patterns. The goal of the current study is to investigate how users enter compound information (for more detailed research questions, see Chapter 2). Furthermore, in Chapter 4, detailed research questions are listed with respect to how users spatially refer to objects using pen and or speech.

### 1.2.4 How are the new technologies performing?

In order to provide a working system that is able to recognize compound information and spatial layout arrangements, certain assumptions have been made with respect to the expected multi-modal handwriting, drawing, gesturing, and speech utterances. We will analyze the loggings to identify cases that have not been foreseen and that therefore require further improvements of the system. This holds in particular for pen input, where compound objects and spatial moving gestures will be discovered that have not been observed until now.
Furthermore, the analysis will assess whether the new turn taking protocol and the user-driven interaction did indeed elicit a substantial amount of multi-modal user input. In particular, the suitability of the timing parameters employed by the different modules will be examined.

### 1.2.5 Comparison with the T16 system

As part of the assessment on how users rate the system, a similar questionnaire as the one that was used in the T16 experiments was used. By comparing the ratings between both systems, a comparison can be made between their performance, usability, acceptance and pleasantness. A truly objective comparison between both systems is difficult to make, as the experimental design and the task subjects have to perform differ significantly. In the T16 system, users follow a system driven dialog to enter all information required for phase1 of the COMIC demonstrator (i.e., shape and dimensions of bathroom, window, and door). In the current experiment, only sub-dialogs are contained, that request (compound) parts of a bathroom or that request modifications using spatial relations.

The same holds for an objective evaluation of the performance of the system components: since the repertoires of gestures and utterances in the present experiment is substantially larger than what could be entered in

the T16 system, error rates cannot be compared directly. Yet, it is interesting to look at the recognition performance in some detail, at least to the extent that this performance can be expected to have an impact on the subjective evaluations.

## 1.3 Organization of this Document

In the next chapter, Chapter 2, we will discuss how user initiative and natural interaction can be improved by providing the option to enter multiple information items in one turn. Furthermore, the research questions with respect to the specification of compound information will be explained in more detail. In Chapter 3, we will explain the new turn taking mechanism and associated changes to the input recognition, NLP and FUSION modules. It will be explained why we expect that the new turn taking protocol will increase the amount of user initiative and why that should lead to improved user satisfaction as well. In Chapter 4, the topic of spatial relations between bathroom objects will be discussed. An overview of computational models of "move specifications" will be provided. In Chapter 5, the experimental design and an overview over the stimuli that will provoke certain multi-modal interaction patterns will be given. In Chapter 6, the results of the experiment are presented. A summary of all findings and suggestions for research toward T36 will be given in Chapter 7. In the appendices A and B, the instructions presented to subjects before the experiment started and the questionnaire that was completed directly after the experiment are respectively provided.

# Chapter 2

# Multi-modal Specifications of Compound Information

As mentioned in the introduction, it is expected that when users are given the possibility to enter compound information, they will will indeed use this freedom in their interaction. Eventually, this should enable the long term goal pursued by COMIC: natural interactions. For pen input, we have christened this interaction paradigm as *write anything, anywhere* [4, 6]. If speech is added as an extra modality, the user would be allowed to say or write anything, at any time (s)he wants. Since the underlying turn taking protocol is still half duplex, users can only enter information when the system has relinquished the floor, and the turn is given to the user. However, once the user has the floor, (s)he can continue entering information, until the system detects an end-of-turn. This can imply that there is no more information to provide, or that the user is not sure what additional information is required. In the latter case, the system should take the turn to provide guidance in the form of a request for specific information, or with more general context dependent help.

## 2.1 Compound Information

Within the domain of bathroom design, we have identified a number of atomic objects: a wall, a door, a window, a size, and deictic gestures. Compound information within this domain can consist of any (meaningful) combination of these atomic objects. They are part of the elaborate COMIC ontology on bathroom design and are relevant for the first phase (phase1) of the COMIC demonstrators. Note that compound information is already being processed in COMIC in case of sizes. A size may consist of two smaller information items, like in “3 meter”, where “3” is the measure and “meter” the unit. Similarly, it must be noted that the FUSION module has been capable of combining certain compound multi-modal information like $\{$wall, wall_length$\}$ and $\{$window, width, height$\}$ since before T24. However, due to the relatively limited recognition capabilities of the input recognizers and in particular due to the strict system-driven dialog (which requested solely for atomic information), this has not been available and tested in the COMIC system yet.

Table 2.1 below depicts the T24 versus T30 capabilities of the system for processing compound information items. Compound walls comprise two or more walls, which can be input via pen only, e.g., by drawing several straight lines, or combinations of ’L’-shapes, ’U’-shapes or rectangular shapes.

From this table, it may be deduced that only atomic objects of the same class can be combined to form a compound object. However, as Figure 2.1 depicted below shows, far more complex hierarchical combinations may be possible. A particularly interesting topic to be researched is how users enter compound information while using pen and speech simultaneously. For example, for compound information containing walls and lengths, the latter information may be entered through pen and or speech.

| Class | T24 objects (all atomic) | T30 objects (mostly compound) |
|---|---|---|
| Wall | Wall | Wall* |
| Door | Door | Door |
| Window | Window | Window |
| Size | Size | Size* |
| Deictic | Tapping, encircling, erase | Tapping*, encircling, erase*, arrow, crossing* |

**Table 2.1:** Atomic objects, processed at T24, versus the objects (mostly compound) that the system can process at T30. The asterisk ('*') means compound.

The most complex compound is a bathroom. Each bathroom consists of at least four walls (in COMIC, we will restrict the complexity of a bathroom to exactly four walls). Each wall may contain a door and/or a window. Doors, walls and windows contain coordinates (not depicted here) that specify their location. Walls and windows contain sizes. Doors contain hinges and opening directions.

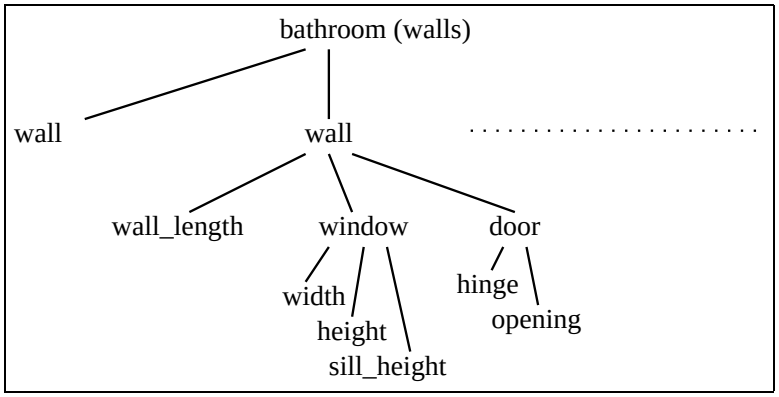



**Figure 2.1:** Simplified hierarchical organization of bathroom objects from the COMIC ontology. Each node in the tree is a compound object. Leaves from the tree are atomic objects.

Compound objects can be composed of several atomic objects or various hierarchical compounds. Suppose, for example, the user wants to draw each of the four walls individually in one turn. The resulting compound information would be composed of four atomic walls. Note that there are many alternatives to entering the outline of a bathroom. The complexity of this domain is combinatorial and becomes even more complex if it is considered that some walls may be accompanied by their corresponding lengths and some walls not. It is the goal of the current study to explore the typical way in which users generate compound information and to cover the most prominent part of this problem space through this exploration.

## 2.2 Challenges and research questions for the input and fusion components

This section briefly describes the technological challenges that compound multi-modal information pose on the underlying technologies. Furthermore, the accompanying research questions are listed below.

- Challenges for pen input interpretation

  By being able to process compound information, the developments in pen input recognition technology tackle one of the major goals in COMIC: to be able to distinguish different kinds of information based

on the available pen trajectories. In particular, the challenge in the current experiments is to distinguish handwriting from walls and to separate different pieces of handwriting or walls from each other. Based on an analysis of the recorded pen input, the following research questions will be pursued:

- What is the pen input repertoire for compound objects?
- How do users draw multiple walls and corresponding lengths?
- When inputing multiple information items, do users have a preferred order in which these are entered?
- What is the performance of the pen input recognizer for compound objects?
- Are the implemented computational models sufficient or can problematic cases be identified?

- Challenges for speech recognition

  The recognition of compound input, such as *"This length <+pointing gesture> is 2 meter [pause] and this length <+pointing gesture> is 3 meter fifty"*, demands specific requirements in the ASR software and in the language modeling. As a starting point for the recognition software in the ASR module, the open-source ASR software from Cambridge is used, which is basically able to recognize atomic utterances. For multiple reasons, its original architecture is not well equipped to manage recognition of long stretches of speech in an on-line environment. Any pause within an utterance will in principle be interpreted as an event that flags end-of-speech, triggering the back-trace and the entire decoding of the uttered speech signal. During the COMIC project and also for T30, the ASR module (the ASR wrapper and the underlying HTK code) has therefore been adapted on an ongoing basis to meet the specific requirements posed by the various experiments, such as the ability to recognize longer, compound utterances.

  The recognition of compounded input is a particular challenge for the ASR module. As mentioned before, the underlying HTK code is not well designed for truly continuous speech recognition. This fact limits the recognition of compound inputs, especially in the case where compounding is associated with long within-utterance pauses and filled pauses. Therefore, the leading questions are the following:

  - How can compound input be encoded by using modifications in the LM?
  - How well is the garbage modeling able to bridge acoustic realizations of compound inputs?
  - How well does the turn-taking protocol cohere with compounding input?

- Challenges for NLP

  Utterances can now contain several independent semantic entities, e.g., several wall lengths, and not just one as in previous COMIC systems. As a consequence, some kind of input segmentation had to be integrated in the NLP module including an enhanced scoring function which is able to cope with alternative segmentations.

- Challenges for FUSION

  Entering compound information permits the user to employ a wide variety of interaction patterns in order to convey complex information. Therefore, we had to extend and update the rule base of FUSION so that it will be able to cope with virtually all interaction patterns. The focus with respect on the analysis of the results will lie on the following two questions:

  - How do users coordinate unimodal contributions in order to input multi-modal compound information?
  - Are there any interaction patterns that are not covered by the FUSION rules?

# Chapter 3

# Turn-Taking and End-Of-Turn Detection in T28/T30

The general idea of the new turn-taking protocol is to give the user more freedom during the interaction with the system. Key to this approach is that both channels – the microphone and the tablet – are open until FUSION decides that the user has finished her turn. This enables the user to enter so called compound objects, which consist of multiple *atomic* objects. Besides this, the new turn-taking protocol permits a less restricted begin of user contributions. The user is no longer forced to start to speak or write in a short time window after the opening of the input channels. The realization and monitoring of this turn-taking protocol is managed by FUSION, which is responsible to identify the end of a user turn. All input events perceived during a turn will be integrated into a single multi-modal representation of the analyzed user intention.

On an abstract level the new turn-taking protocol is partitioned into two states: the *user-turn* – when the user is supposed to speak and/or write – and the *system-turn* – when the system has the floor –. In case the turn-taking protocol is in the state user-turn, the input channels are opened and the user can input information; otherwise the input channels are closed. These two states were already presented in the T24 turn-taking protocol. However, the new turn-taking protocol does not apply static time-outs defining the time frame within which the user can speak, but rather performs an *end-of-turn* detection based on the actual contributions of the user. The key idea is that the recognizers inform FUSION immediately when they detect some input signal. Such an input signal triggers in FUSION the expectation that it will receive most probably a recognition result from a particular channel in the near future. As long as there are open expectations about incoming recognition results, FUSION will never consider a turn as being finished. However, when all expectations are closed by a corresponding recognition result – which can be empty in case of a false alarm – FUSION employs a short time-out to ensure that the user does not want to enter additional information and eventually sends the integrated semantic representation of a turn to the next module in the processing pipeline.

## 3.1 Changes in the system components

Besides the obvious changes in FUSION this new turn-taking protocol requires several changes in other system components. In the following sections we will briefly describe the major requirements and changes to various system components.

### Interfaces between Modules

*New pool:* `turntaking.management`

Publisher: FUSION, FISSION (T30++), *GO-Button* (T28) / DAM (T30++)

Subscriber: ASR, PII, SAV

This pool monitors the current state of the conversation i.e. who is having the turn. In case the channels are open this pool will contain a message `userTurn` stating that the user is now able to enter commands. This message is send by FISSION (T30) after the system output is finished for T28 pressing the *Go-button* will trigger this message. If FUSION detects an end-of-turn it sends a message `systemTurn` to this pool. This message causes ASR and SAV (for T30 the savWrapper) to close their respective channels.

## Changes to PII

PII has four states: (i) channel-closed, (ii) idle, (iii) inputReceived, and (iv) recognizing/analyzing. Possible state transitions are:

channel-closed → idle
if a message *userTurn* appears on the pool `turntaking.management` SAV (or in T28 the savWrapper) opens the tablet and PII changes its state to idle (== expecting input; in contrast to T24 no message is sent to FUSION)

idle → inputReceived
if PII receives via the SAV (or the savWrapper) a message containing the first set of coordinates it changes its state to inputReceived and reports that to FUSION (i.e., by sending a message `<inputReceived/`). This indicates that it is certain that a recognition result will follow, so FUSION uses this message to delay a switch to the state *system turn*.

inputReceived → recognizing/analyzing
PII employs a timeout parameter that is set after each package of coordinates that is received from SAV. If a new package arrives, the parameter is reset. If the timeout passes, PII changes its state to `<recognizing/analyzing>` and reports that to FUSION.

recognizing/analyzing → idle
if PII finishes the processing of some user input it sends the computed analysis result to FUSION (or NLP in case of written input) and changes its state to *idle*

idle → channel-closed
if FUSION sends a *systemTurn* message to the pool `turntaking.management` SAV (savWrapper) closes the channel and PII goes into the channel-closed state.

## Changes to ASR

At a conceptual level, ASR has three states: (i) channel-closed, (ii) idle, (iii) recognizing/analyzing. (The actual software distinguishes more states, but this fact is not of primary relevance for the current discussion about the turntaking concepts.) Given this set of states, the possible state transitions are:

channel-closed → idle
if a message *userTurn* appears on the pool `turntaking.management` ASR opens the microphone and changes its state to idle (== expecting input; in contrast to T24 no message is sent to FUSION)

idle → recognizing/analyzing
if ASR detects (via the speech detection) an input signal, it changes its state to recognizing/analyzing and reports that to FUSION (i.e., by sending a message `<inputReceived/>`)

recognizing/analyzing → idle
if ASR finishes the processing of some user input, it sends the computed analysis result to FUSION (via NLP) and changes its state back to idle

idle → channel-closed
if FUSION sends a *systemTurn* message to the pool `turntaking.management` ASR closes the microphone and goes into the channel-closed state.

Due to the rigid and constrained architecture of the HTK/HVite recognizer, the modification of the ASR module necessary to make it compatible with the T30 turntaking protocol have a substantial impact on the flexibility of the ASR module to cope with user inputs.

### Changes to FUSION

For T28 FUSION will apply a combination of a pattern recognition and a time-out mechanism to identify the end of a user turn. During a user turn FUSION will collect all recognized user actions and integrate them (if possible) into a single multi-modal representation, until it detects the end of the turn. The end-of-turn detection will ensure that a user turn is only closed if there is no pending recognition result in the pipe-line. Therefore, it is important that FUSION gets informed in case one of the recognizers is faced with some input. FUSION will only detect the end of a turn in case the user enters at least a single command. However, to avoid deadlocks in the dialog, there will be a second (less prioritized) time-out that will cause FUSION to output a message stating a FUSION_TIMEOUT. In this case the T30 or T36 DAM can take appropriate actions to recover the dialog.

## 3.2 Research Questions

A smooth and immediate exchange of turn is most essential for a natural and pleasant conversation. Therefore it is very important for the further development of the COMIC system to design a turn-taking protocol that is accepted by the prospective end users. The natural full-duplex interaction, of course, would be the solution. However, there are several technical restrictions that forced us to stick to our previous half-duplex approach. But as the results of the T24 evaluation clearly show the need for a faster and more flexible end-of-turn detection, we extended this approach to fit to the turn-taking protocol explained above.

The outcome of the T28 experiments will be used to further enhance this turn-taking protocol and to test its general applicability. To this end we extracted from the recorded log-files of the interactions reliable timing information. Questions we want to answer are:

- How well do the timing parameters of each module perform?
- How long must FUSION wait after it received a recognition result until it can consider a turn as being finished?
- Are there any prominent interaction patterns that can be identified already during the interaction?
- How long must FUSION wait until it can be sure that the user will not enter any input and thus trigger a timeout?

# Chapter 4

# Multi-modal Specifications of Spatial Relations

In order to address the issue of spatial relations in COMIC, we designed an experiment in which subjects had to modify the position of objects like doors and windows. In T28, this task is implemented by requesting subjects to move existing objects, i.e. windows and doors. Objects can be moved through speech and/or pen. Each move operation is encoded by the FUSION component, after merging the pen and speech hypotheses. In T28, the object to be moved is known in advance, as this is contained in the expectation. Therefore, its identification can be added by the FUSION component:

```
<object type="Move_Bathroom_Part">
  <slot name="has_object">
    <object type="Window_Shape">        /* or, alternatively, "Door_Shape" */
      <slot name="id">                  /* the <id> is always 0 */
        <value type="integer">0</value> /* as there is only 1 door and window */
      </slot>
    </object>
  </slot>
  ....
  <slot name="has_reference_point">     /* for absolute moves (see below) */
  ....
  <slot name="has_reference_object">    /* for relative moves (see below) */
  ....
  <slot name="has_distance">            /* for relative moves (see below) */
  ....
  <slot name="has_relation">            /* for stepwise moves (see below) */
  ....
</object>
```

Depending on the way the user expresses the move operation, some or all of the slots indicated above may be required. The different stimuli will trigger subjects to employ many different speech and pen input expressions, resulting in various manners to perform the move operations. One of the goals of the current experiments is to find out what multi-modal interaction patterns actually are actually employed, and to adapt the capabilities of our modules accordingly. For the T28 functionality, we have anticipated the following move operations:

1. **absolute** moves, e.g., "Move the <`object`> <`here`>". The location <`here`> is encoded as a so-called <`Absolute_Location`>, indicating the position $(x, y)$ on the screen. Absolute locations can only be entered through pen. As it is known (because of the design of the stimulus) which object has to be moved, speech is not required to disambiguate between door or window. *However, at T30 this*

*is absolutely required. If in a later stage more than one door and window are present, other spatial references are required to disambiguate.* Absolute moves are encoded by FUSION via a slot called <slot name="has_reference_point">:

```
<slot name="has_reference_point">
  <object type="Absolute_Location">
    <slot name="has_x_position">
      <value type="integer">205</value>
    </slot>
    <slot name="has_y_position">
      <value type="integer">359</value>
    </slot>
  </object>
</slot>
```

2. **distance** moves, e.g., "Move the <object> <distance> to the <direction>". These move operations require a direction and a distance.

   Directions are encoded via a slot called <slot name="has_relation"> and distances via a slot called <slot name="has_distance">. As an example, consider the case where the user says: "Move the window 80 cm to the left", which is encoded as:

```
<slot name="has_distance">
  <object type="Size">
    <slot name="has_value">
      <value type="float">80</value>
    </slot>
    <slot name="has_measure">
      <value type="string">cm</value>
    </slot>
  </object>
</slot>
<slot name="has_relation">
  <value type="string">left</value>
</slot>
```

3. **stepwise** moves, e.g., "Move the <object> <direction>".

   Stepwise moves always indicate that the object must be moved along the wall it is contained in. For T28, a step is determined as 20% of the wall length. For example, "Move the window up" (for a wall with length 3 meters) is equal to "Move the window 60 cm up". The given example is encoded by FUSION via a simple:

```
<slot name="has_relation">
  <value type="string">top</value> /* possible values are left,right,top,bottom */
</slot>
```

4. **relative** moves, e.g., "Move the <object> to the <direction> of the <reference_object>". Relative moves can only be entered through speech and always contain a reference object (encoded via a slot called <slot name="has_reference_object"> and a direction. As an example, the user says: "Move the window below the door", encoded as:

```
<slot name="has_reference_object">
  <object type="Door_Shape"/>
</slot>
<slot name="has_relation">
  <value type="string">left</value> /* possible values are left,right,top,bottom */
</slot>
```

5. **relative distance** moves, e.g., "Move the <object> <distance> to the <direction> of the <reference_object>". Relative distance moves can only be entered through speech and always

contain a reference object (encoded via a slot called <slot name="has_reference_object"> and a direction. As an example, the user says: “Move the window 80 cm right to the door”, encoded as:

```
<slot name="has_distance">
  <object type="Size">
    <slot name="has_value">
      <value type="float">80</value>
    </slot>
    <slot name="has_measure">
      <value type="string">cm</value>
    </slot>
  </object>
</slot>
<slot name="has_reference_object">
  <object type="Door_Shape"/>
</slot>
<slot name="has_relation">
  <value type="string">right</value> /* possible values are left,right,top,bottom */
</slot>
```

## 4.1 Recognizing Move-Operations

By examining the combination of slots that are filled by FUSION, a simple heuristic can be implemented to distinguish between the cases. Table 4.1 shows that each of the move operations has a unique combination of slots:

| move | ref point | ref object | distance | relation |
|---|---|---|---|---|
| abs | X | | | |
| dist | | | X | X |
| step | | | | X |
| rel | | X | | X |
| reldist | | X | X | X |
| example | X | X | X | |

**Table 4.1:** The five allowed move operations. For a discussion on the sixth row, see the text below

The problem is that subjects can use more expressions than the ones defined above. As an example, the user may tap a location while asking to move a window: “This window is <tapping gesture> 80 centimeters from the door”. It appears that this example does not fit in one of the five possible move operations. And therefore, it will be rejected, unless the gesture information reveals the (missing) directions. So, unless the user marks the target position rather than the source.... A not-so-uncommon other problematic example is where the user taps twice: “this window is <source tapping> 80 centimeters <target tapping> from the door”.

Recognition of gestures in the T28 research platform is performed by a program called savWrap, (see Section 5.2) which interprets all outputs from FUSION. As this program lacks the intelligence of a dialog manager, cases not covered by the heuristics displayed in Table 4.1 are rejected and stored for further examination.

## 4.2 Challenges and research questions

This section briefly describes the technological challenges that spatial relations pose on the underlying technologies and lists the accompanying research questions.

- Challenges for Pen Interpretation

  To process moving gestures is one of the goals in COMIC. The developments in pen input recognition technology tackle this goals. In particular, the challenge in T28 is to distinguish between deictic gestures, such as tapping and encircling, and arrow gestures. The analysis of the results will be focused on an inventory of *spatial moving gestures* and a comparison of the detected categories with the models specified in this chapter.

- Challenges for Speech Recognition

  The recognition of spatial relations does not yield specifically new challenges for ASR - the LM covers the user utterances that were expected to be most useful and most frequent during the experiments. However, in connection to the structure of the language model, the challenges for ASR are imposed by the way in which users may produce the utterances - if many hesitations occur, the ASR is not able to make correct distinctions between a true end-of-speech event (after which a complete back-trace may follow) and an intra-utterance pause. This directly involves the interpretation of utterances such as *"Move this window to the left [pause] of this door"*, in which two deictic gestures occur. If the pause is beyond the threshold for end-of-speech detection, the recognizer will perform a full decoding (using the full grammar) on only a part of the utterance that the user actually intends to provide to the system, viz. 'move this window to the left'. If this sub-phrase happens to be grammatical, a decoding can be done that is correct from the ASR point of view, but entirely incorrect given the sub-phrases to come. This effect forms one of the key issues for turn taking, compound recognition, both for ASR and NLP and dialog management.

  To a large extent, these issues originate from the specific architecture of the core ASR recognizer. This architecture can only me modified marginally within the constraints of the COMIC project.

- Challenges for NLP

  Besides providing an update of the knowledge bases to deal with the new utterances for move and erase operations, the ontology had to be extended so that the referring expressions, like *this door*, *here* etc. can be represented in terms of the ontology. A proper representation of referring expressions is essential, as referential expressions are not resolved directly by NLP, but by FUSION.

- Challenges for FUSION

  For FUSION we had to address two issues that emerge from the frequent use of referring expressions in the context of spatial relations (e. g., "move **the door** left to **the window**" or "move **this door** [pointing gesture] left to **the window**"): (i) a robust handling of deictic gestures and (ii) resolution of referring expressions by accessing the discourse context. Whereas the first point required only a careful revision of the existing integration rules, the second point requires an extension of the expectation mechanism, as the T28 system does not comprise a dialog history. To this end we had to model all objects comprised by a stimulus in the expectation.

  We are mainly interested on the multi-modal constructions that the subjects employ (e.g., in what cases do they use multi-modal commands, when do they use unimodal ones and the respective timing). Can we find some adaptation effects?

- Research questions
  - What are natural interaction patterns when users try to move objects?
  - Which pen repertoire and which speech (words and syntax) are employed?
  - Are there any particular prominent phrases?
  - Are there any particular cases that are not covered by the models listed in Table 4.1?
  - Do subjects prefer to select target locations by speech (e. g., "...left to the door") or do they prefer pointing gestures?

# Chapter 5

# Stimuli and Expected User Behavior

This chapter describes the design and set up of the T28 experiments. Furthermore, an elaborate discussion of all stimuli is presented. For each stimulus, a description is given indicating why it was designed and what multi-modal utterances can be expected.

## 5.1 Examples of the T28 Stimuli and User Responses

In the T28 experiments, users are asked to modify a certain situation that is depicted on the screen. By designing these initial situations in a certain manner, while requesting particular modifications to be made, it is possible to invite users to enter multi-modal information in a natural fashion. As an example, consider the initial situation (called “stimulus”) depicted in Figure 5.1. The subject is requested to move the window below the door. Each stimulus is designed such that the requested target situation does not collide with the initial situation, as would have been the case if the door were located near the bottom wall.

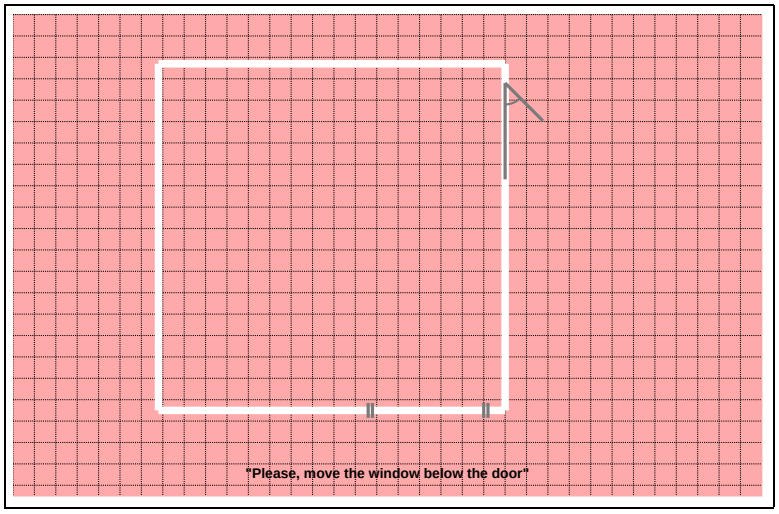



**Figure 5.1:** Example stimulus in which the user is requested to modify the current screen state by moving the window. It is expected that such stimuli will trigger users to employ pen and speech to specify spatial relations.

What will happen when the user has to fulfill this request is the following. First, the user will form a notion (mental model) of the wanted situation. Subsequently, (s)he will consider the options that are available to perform the modification. For example, the user can say “Move the window below the door”; the user can say “Move the window here”, while tapping on the target position; the user can generate an arrow-shaped gesture (starting at the source position of the window and ending at the wanted target position), etcetera. After this plan has been made by the user, (s)he will perform the requested operation using pen and/or speech.

After interpreting this multi-modal input produced by the user, the system will beautify[1], the modified situation, or, alternatively, reject the input in case it could not be interpreted with sufficient confidence. In the latter case, the user must retry the modification. In the former case, a human experimenter will acknowledge, or, alternatively, reject the beautification. In case the experimenter judged the system response as correct, a new stimulus will be presented to the subject. In case the system apparently made a mistake, the subject will have to retry.

As another example, consider Figure 5.2 below. In this example, the user is requested to complete the outline of the bathroom (depicted are two walls and corresponding lengths). Furthermore, the user is requested to enter the length of each wall. Here, the user could specify the requested information by drawing the upper horizontal wall, while saying “This is wall is 3 meters” subsequently drawing the second wall while saying its length. However, quite some other scenarios can be imagined, all expected to contain compound information entered in one turn.

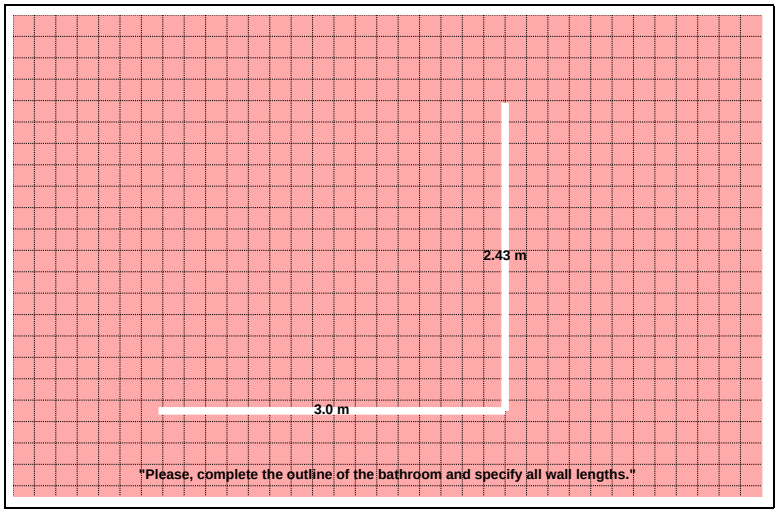



**Figure 5.2:** Example stimulus in which the user is requested to modify the current screen state by completing the outline and entering the corresponding wall lengths. It is expected that such stimuli will trigger users to employ pen and speech to specify compound information.

Note that in this particular example, a reasoning module would already have obtained the complete specification of the room, assuming that it is rectangular. Since in our system only rectangular rooms are supported, the information (shape and size) for two perpendicular walls suffices for computing the complete outline of the bathroom. However, this task can still explore multi-modal human behavior.

All stimuli could be specified through a control file using a simple syntax. For example, the following two descriptions specify the stimuli depicted in Figures 5.1 and 5.2:

[1] As explained in previous reports, beautification is the act of the system to render information on the screen in order to show the user what has been recognized. For example, recognized walls are rendered as straight lines, sizes are rendered as ascii text in a certain font, etcetera. See, e.g., Figure 5.2 for an example of beautified walls and corresponding wall lengths.

```
<stimulus>
"Please, move the window
   below the door"
wall 0 20 200 200 200
wall 1 200 20 200 200
wall 2 20 20 200 20
wall 3 20 20 20 200
window 0 0 130 200 190 200
door 0 1 200 30 200 80 North East
</stimulus>
```

```
<stimulus>
"Please, complete the outline of the bathroom
   and specify all wall lengths."
wall 0 20 200 200 200
length 0 0 "3.0 m"
wall 1 200 40 200 200
length 1 1 "2.43 m"
</stimulus>
```

In total 30 of such stimuli were designed, each requesting for compound information, for re-arranging objects through spatial relations, or erasing one or more objects visible on the screen.

## 5.2 T28 system architecture

In this section, the architecture of the T28 research platform is briefly described. This platform is used for performing the T28 human factors experiments on input technology, NLP, and FUSION. Therefore, other COMIC modules (e.g., dialog action management, fission, speech synthesis, Visoft user interface) are not contained in this platform. However, in order to be able to design interactive experiments, some kind of system feedback is required. This consideration has lead to the decision to implement a user interface via which pen input can be acquired and graphical system output can be displayed.

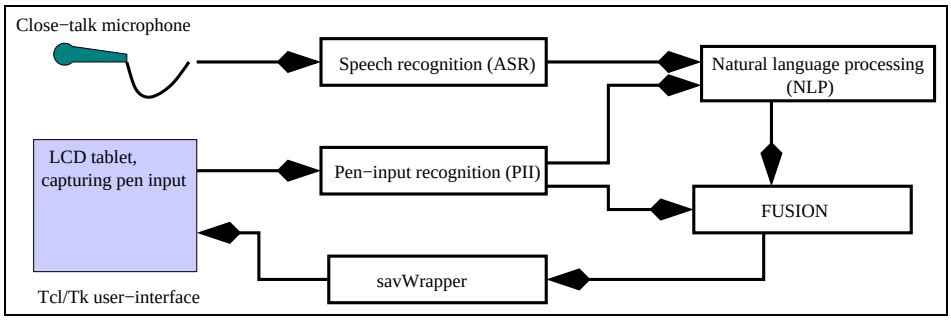



**Figure 5.3:** T28 research platform. For explanation see text below.

The system comprises speech and pen input recognition modules, a natural language processing module, a fusion module, a wrapper program that transforms output from fusion to beautification commands and a Tcl/Tk graphical user-interface that displays information and renders pen trajectories. The latter two programs emulate the Visoft user interface module, called SAV (service application Visoft). All modules implement the new turn-taking protocol, which is described in Chapter 3.

1. *Tcl/Tk graphical user-interface* for rendering stimuli, pen trajectories and beautifications received via savWrap. This program outputs pen coordinates to the same pool that the T30 SAV employs for sending information to the pen input interpretor (PII).

2. *savWrap*, a program that transforms messages received from FUSION to commands that are interpreted by the Tcl/Tk program.

3. *PII*, the pen input interpreter, which recognizes drawings and sketches, deictic, pointing and moving gestures, erasing gestures, and numerical strings. The hypotheses concerned with the latter category are output to the NLP module, whereas outputs concerned with the other categories are directly sent to FUSION. Because of the complexity of compound pen input, an important functionality of the PII is to distinguish handwriting from drawings and multiple drawn objects or texts from each other.

4. *ASR*, the automatic speech recognizer, which recognizes speech and outputs its hypotheses to the NLP modules. On command from outside the ASR, the ASR module opens the microphone and listens to the input signal. By using a language model, acoustic models, the incoming signal is matched against word string hypotheses. As a result, the output of the ASR consists of a so-called word lattice that can be represented in the form of an N-best list. The length of the N-best list as applied in the current COMIC system equals 10. The lattice and N-best list are available after end-of-speech detection. The ASR has been adapted to produce 'early recognition results' *during* the time the user speaks. This makes it in principle possible to reduce response latencies by the entire system. Apart from the hypotheses, several scores (such as confidence scores) are provided.
5. *NLP* The task of the natural language processing module is to analyze the output of the speech recognizer (ASR) and the hand writing recognizer. In the case of ASR the output consists not only of the best hypothesis but a scored n-best list is provided. The result of the analysis process is a list of semantic hypotheses describing the user intention in an abstract representation. The abstract representation is based on the system wide ontology. Several alternative results are possible, since the speech recognizer produces an n-best list and the content of the utterances may be ambiguous, leading to different representations.
6. *FUSION* The task of the Fusion module is to combine and integrate the output of the different modalities into a single representation describing the user intention. During processing the module receives different hypotheses containing analyzed output from the gesture recognition module and the natural language processing module. Each input structure has to be interpreted with respect to the context provided by the other modality. Additionally, it is important to take the current dialog state into account. One important task of the FUSION component is related to the synchronization of the output of the different analyzers and the detection of the end of a user turn.

## 5.3 Set-up of the experiment

### Subjects

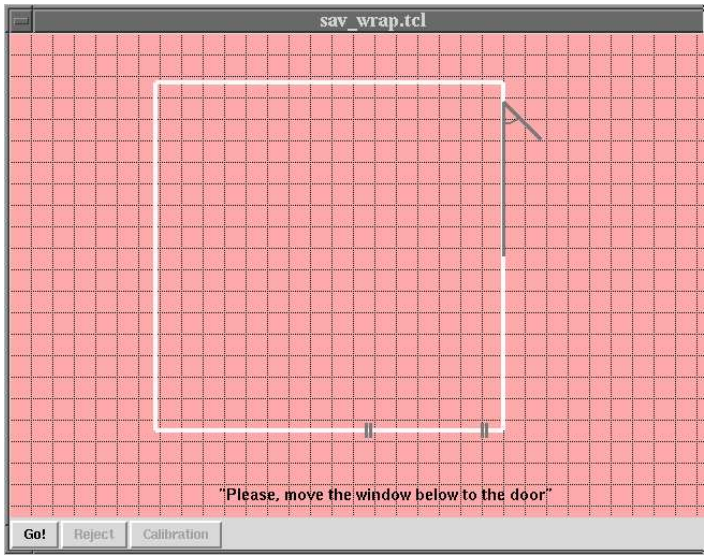
In total 40 subjects participated to the experiment (20 at the NICI in Nijmegen, and 20 at the DFKI in Saarbrücken). Most subjects were students or researchers, did have very little experience with interactive dialog systems and nearly no experience with speech or pen based systems. Before each experiment started, subjects had to carefully read the instructions (see Appendix A) and after the experiment was finished, they had to fill in a questionnaire (see Appendix B).

### Stimuli

Each subject had to complete 30 stimuli. A description of each stimulus is given in the next section. Each group of 20 subjects was divided into two separate groups. The first groups (subjects 1 to 10 at the DFKI and subjects 21 to 30 at the NICI) were presented with stimuli in sequential order. The first set of stimuli comprised requests for atomic information and are therefore considered as less complex than compound information or information associated with the specification of spatial relations. The second groups (subjects 11 to 20 at the DFKI and subjects 31 to 40 at the NICI) were presented with stimuli in random order. This between-group design made it possible to assess whether subjects learn from the interaction. An analysis on number of turns should be able to reveal whether learning effects result in a more efficient interaction. Below, references to *the first condition* or *the sequential condition* both refer to the groups of subjects who performed the tasks in sequential order. It should be noted that due to a programming error, 18 subjects did not see the first stimulus. Therefore, in total 1182 (22*30 plus 18*29) requests for information were presented to the subjects. Stimuli were divided into four groups: (i) atomic (n=8), (ii) compound (n=8), (iii) spatial (n=10) and (iv) erasing stimuli (n=4). Section 5.4 describes each stimulus in detail.

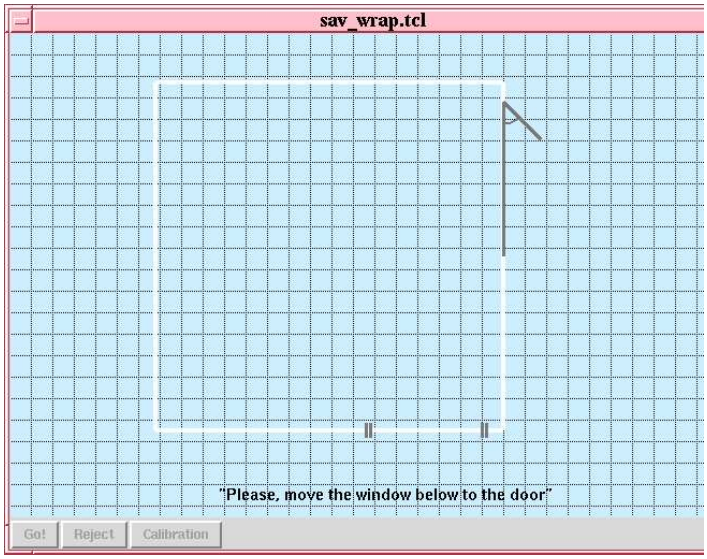
## The user interface

The user interface (see the figures below) contained a drawing canvas for displaying information and rendering ink, and three control buttons below the canvas. Each stimulus represents a certain beautified version of a part of a bathroom blueprint. Individual stimuli can contain a partly completed bathroom. As an example of the progress of a user handling a stimulus, consider the following scenario. In the initial situation, the background of the drawing canvas was light-red, indicating that no user input was allowed: the system has the turn. During this time, the subject could inspect the stimulus and determine a plan to provide the system with a response:



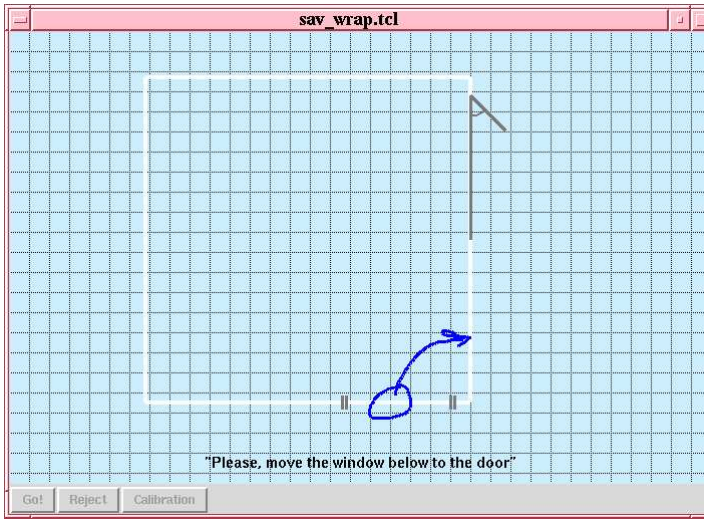

The T28 user interface implemented by savWrap. In this situation, the system has the turn and the user must press “Go!” to start the user turn. Subsequently, subjects could use the left-most button (“Go!”) to indicate that they were ready to start entering information: the user has the turn and the background of the drawing canvas becomes light-blue.

**Figure 5.4:** Initial situation



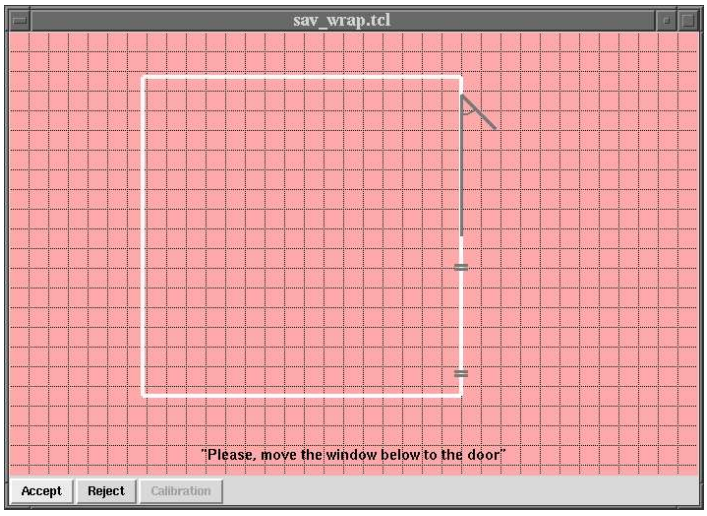

The user has pressed “Go!” and (s)he is allowed to enter information. The background has turned light blue to indicate that user input is allowed.

**Figure 5.5:** User turn




The user has specified the move operation through a pen gesture. Upon deciding that the turn has finished, the system will take the floor.

**Figure 5.6:** End-of-turn




The system has processed the information and rendered the beautification. User input is impossible as the system has the floor. After entering information (this moment is decided by the turn taking timeout parameters employed by FUSION), the background turned red again and the system has the floor. The system will close the input channels (indicated by the red background) and render the response.

**Figure 5.7:** System response through beautification

Now, the subject will be able to reject the beautified information or accept it after which the next stimulus would be displayed. In case the subject would reject the beautification, this will result in re-displaying the current stimulus.

## The role of the experimenter

During all experiments, an experimenter was present. The tasks of the experimenter were:

- to perform the SNR calibration. The right-most control button was clicked by the experimenter after the subject acknowledged that (s)he understood all instructions and was ready to start the experiment. This button triggered the calibration of the computation of the signal to noise ratio required for a proper begin- and end-of-speech detection. The experimenter ensured that this calibration was successful.
- to explain the experiment and make sure that the subjects understood the instructions well. Users could

reply to a stimulus by using pen and/or speech. It was explained to each subject that clicking the side button of the pen would result in so-called erasing gestures, which was rendered on the drawing canvas via a thick light-gray pen trajectory. For each stimulus, all user input (pen and speech signals) was recorded during the time frame in which the channels were open.

- to ensure that each subject pressed the correct button (left="Accept" or middle="Reject") after the system did interpret his/her input. In the case the subject wanted to give up or did not notice that the system response (via beautifications) is wrong, the subject was asked, by the experimenters, to retry. In certain occasions, the system could not recognize any user input (e.g., no input was recorded, or no meaningful interpretation could be computed). In such cases, the "Accept" button would be inactive and the only option for subjects was to retry with the same stimulus by pressing "Reject".

  The left-most button was used to control the progress of the experiment. By clicking that button, subjects could indicate that they were ready to proceed to a next stimulus.

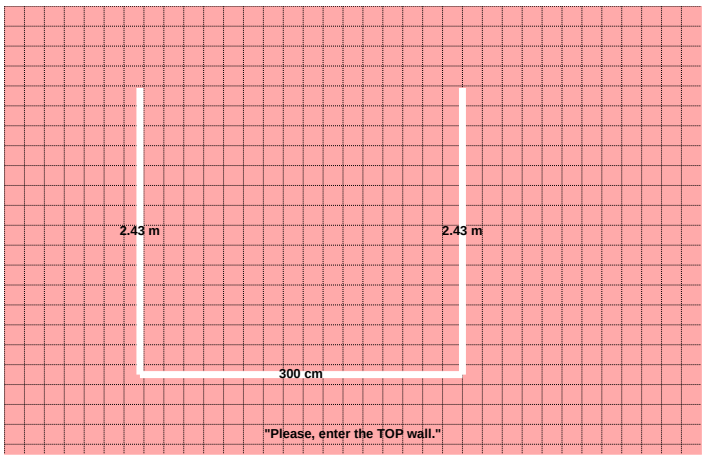
  By supervising the correct button usage, the experimenter ensured that the proper transitions between stimuli were performed.

## 5.4 List of Stimuli

In this section, 4 categories of stimuli are presented: (i) atomic (n=8), (ii) compound (n=8), (iii) spatial (n=10) and (iv) erasing stimuli (n=4). Below, for each stimulus, a brief description is given as well as some particular comments recorded by the experimenters during the experiment.

### 5.4.1 Atomic Stimuli

Atomic stimuli are accompanied by relatively simple prompts, indicating what kind of information is requested and where this information belongs to. Note that the way in which prompts are given may provoke users to employ similar utterances. This issue is well-known from speech-based systems. In the sequential condition, these atomic stimuli are presented to subjects before the more complex stimuli (e.g., containing compound objects or doors and windows to be moved) are shown. This makes it possible to examine what inputs are generated in cases that the subject is unaware about how this could be done. In former human factors experiments (pilot experiments, T16 HF experiments [4], T24 Evaluation), subjects were always shown a blueprint that they had to copy.




The first stimulus is very simple: subjects are requested to draw one single wall. A wall can be inputed only using the pen. Some subjects tried to enter the wall using speech. They said, for instance: "please draw the top wall". This was not expected. Subjects have been influenced by the written request (by the word "top"). This has not been noticed during the former experiments, for which the requests were much less directive: "please draw a wall" at most.

**Figure 5.8:** Atomic wall.

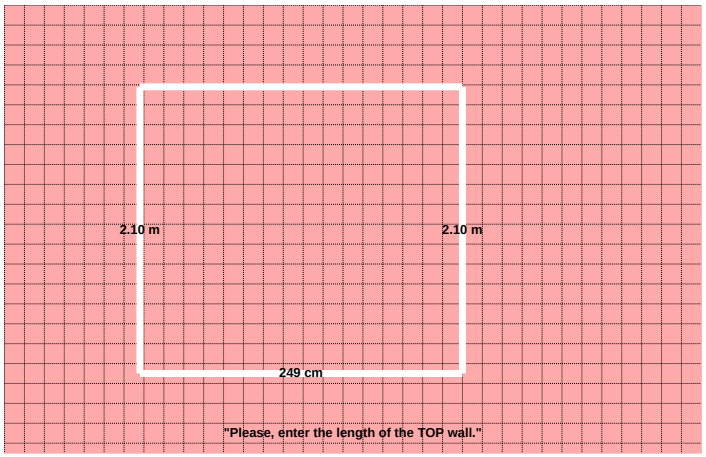



**Figure 5.9:** Atomic size.

This stimulus makes it possible to examine how subjects enter a length. Do they prefer to use one of the modalities (pen or speech) or not? Most of the subjects copy the length of the bottom wall.

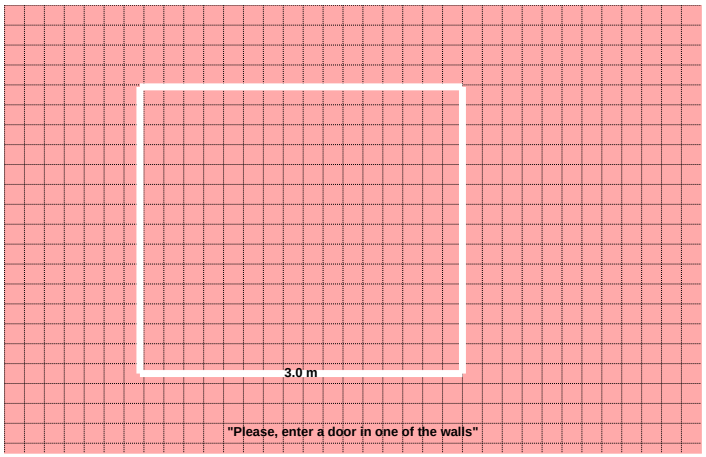



**Figure 5.10:** Atomic door.

In the first condition, subjects did not see a beautified door yet. As this was not the case in the former experiments, when the subjects had to “copy” a blueprint, we expect to get more insight in how users enter a door.

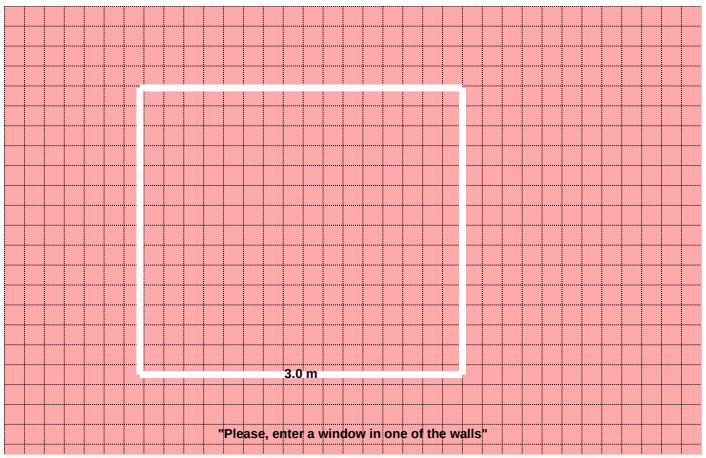



**Figure 5.11:** Atomic window.

In the first condition, subjects did not see a beautified window yet. This allows us to get more insight in how users enter a window when they are unaware of how this could be done.

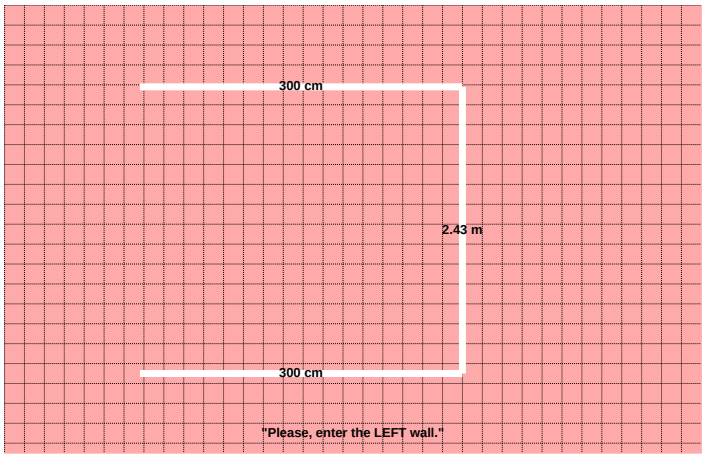



**Figure 5.12:** Atomic wall.

The remarks expressed for the stimulus 5.8, where the top wall had to be entered, are also relevant here.

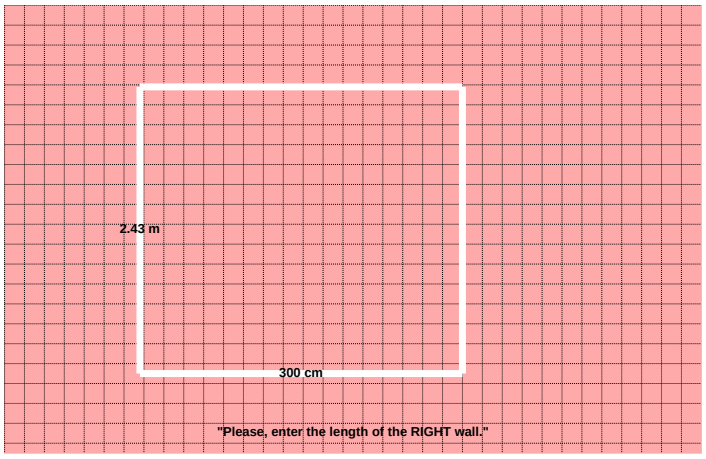



**Figure 5.13:** Atomic size.

Same comments as for 5.9 hold here.

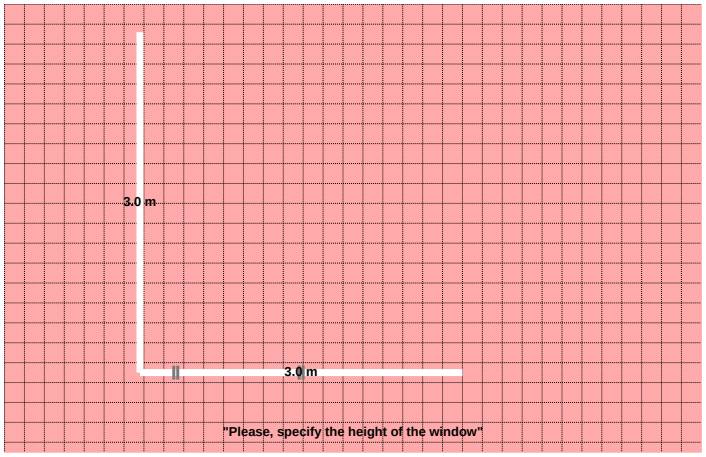
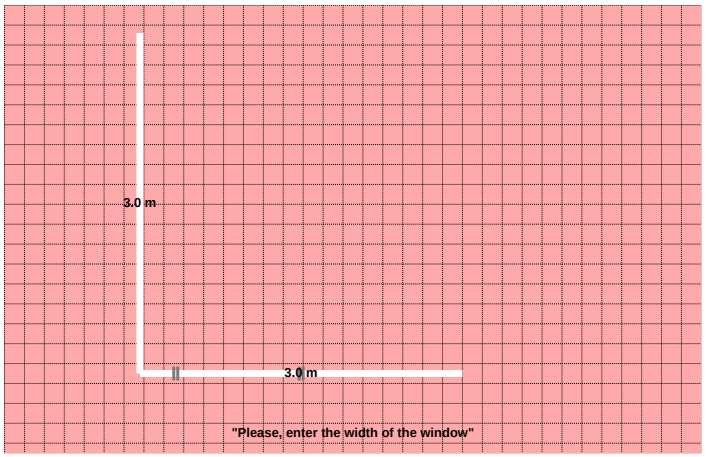



**Figure 5.14:** Atomic size.

Compared to stimuli 5.9 and 5.13, where subjects could deduce the requested size, here, subjects had to “invent” a size. More natural (with hesitations) input is expected, for pen as well as for speech.

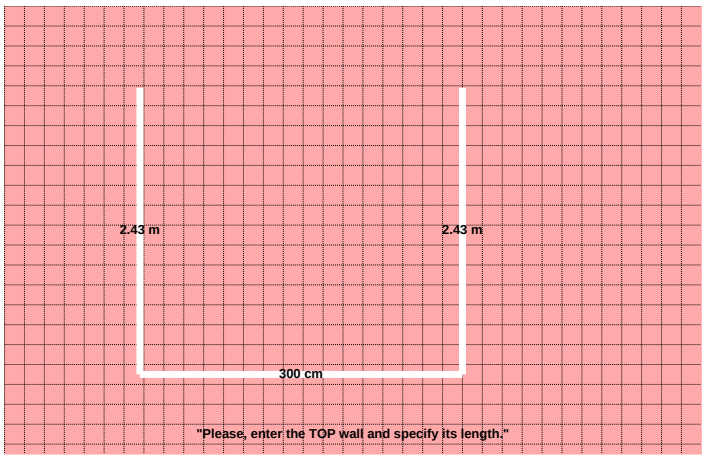



Same comments as previous stimulus.

**Figure 5.15:** Atomic size.

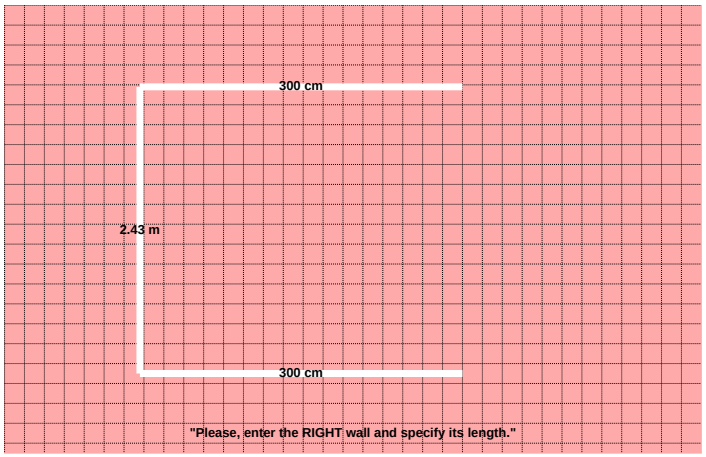
### 5.4.2 Compound Stimuli

Below, eight stimuli are presented that request compound information comprising one or more walls and one or more lengths. It is expected that walls are entered using the pen. Both pen and speech can be used to enter lengths. If users employ both, the question is in what order pen and speech are used. Do subjects first draw a wall and subsequently specify its length through speech, or vice versa.



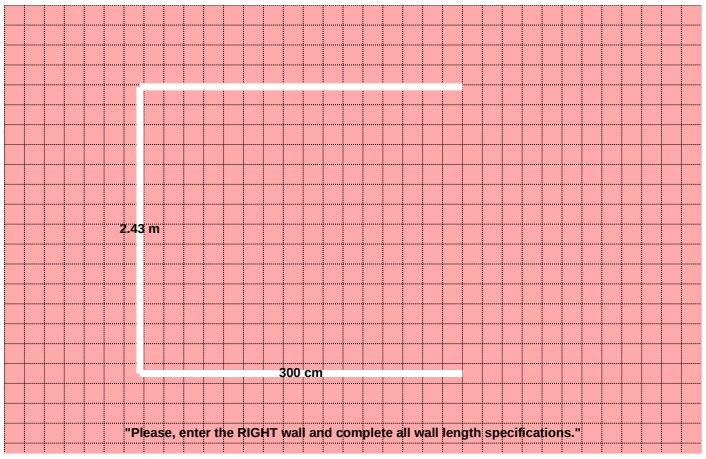

How do the subjects enter these two objects? It is expected that they provide the wall using pen, but they can enter the length by writing, speaking or using both modalities together. If they use speech to provide the length, do they draw the wall and then speak, or do they draw and speak at the same time?

**Figure 5.16:** Two atomic objects: a wall and a length.



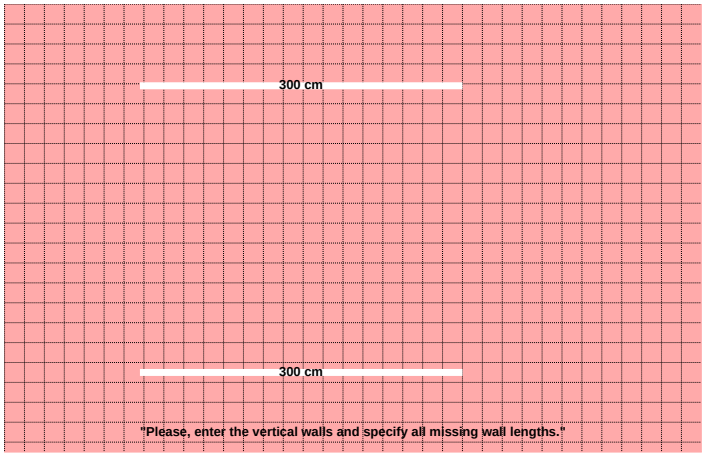

See 5.16.

**Figure 5.17:** Two atomic objects: a wall and a length.



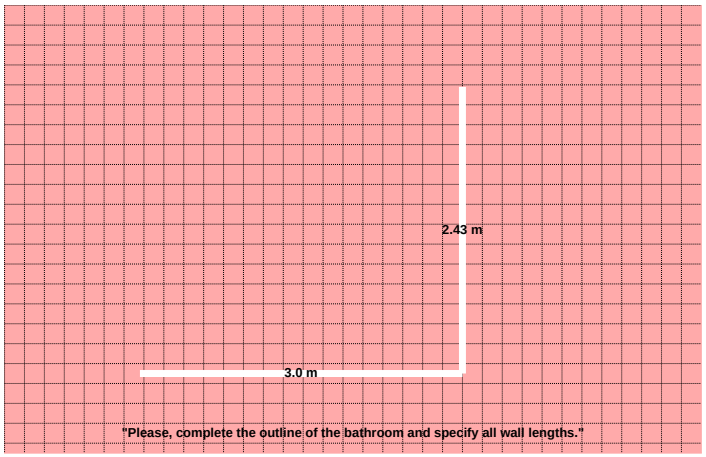

Which modality use the subjects for providing the lengths? Do the subjects enter first the wall, and then the lengths (in that case, do they prefer to enter first the length of the wall they just provided or not?); or the length of the top wall first, and then the wall and its length? Do most of the subjects take into account the lengths already present on the screen?

**Figure 5.18:** Three atomic objects: a wall and two lengths.



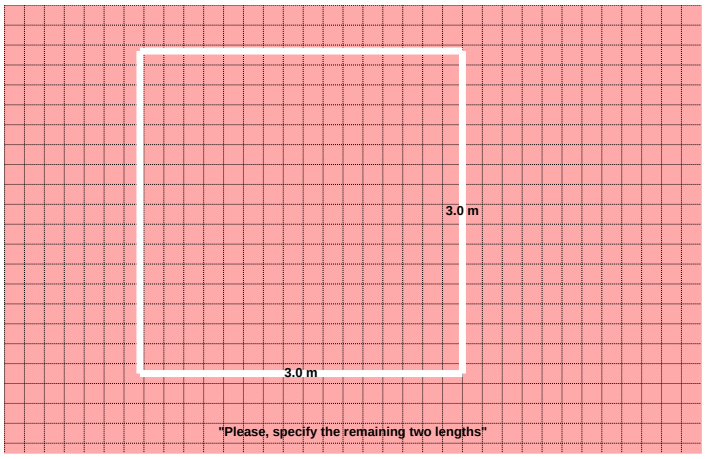

In which order do the subjects enter these four objects? Which modality do they use for providing the lengths? Note that in this stimulus, the subjects have to figure out the size of the two vertical walls.

**Figure 5.19:** Four atomic objects: two walls and two lengths.



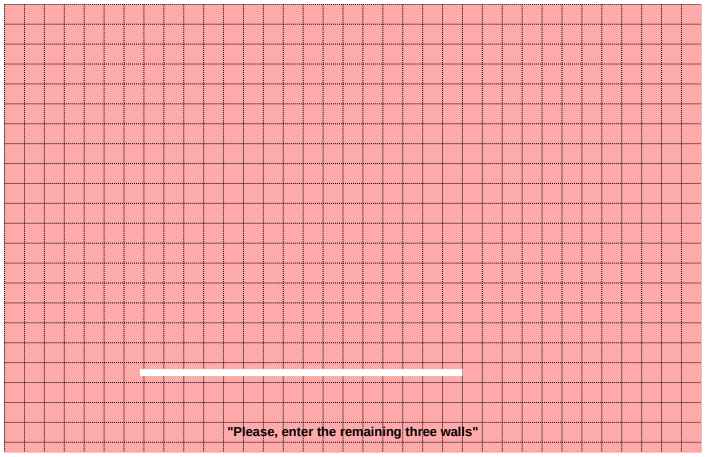

Again, the order in which subjects enter information can be assessed. Furthermore, the question is whether subjects take the lengths already present on the screen into account.

**Figure 5.20:** Four atomic objects: two walls and two lengths.



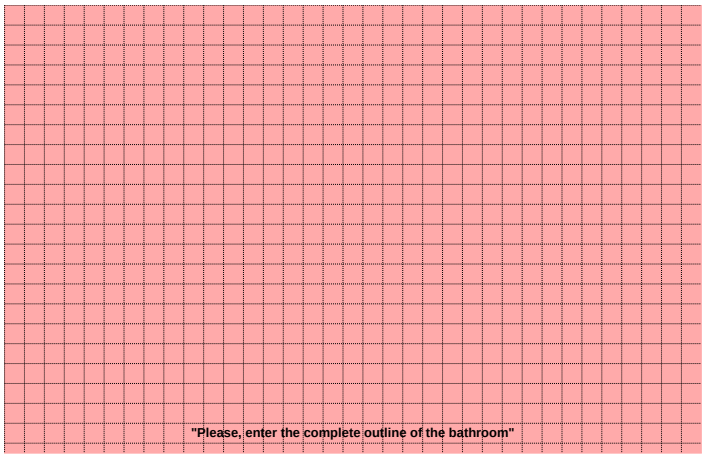

See stimulus 5.20.

**Figure 5.21:** Two atomic objects: two lengths.




How do the subjects draw these three walls: producing one stream (a U-shape one), or two, three (one for each wall) or more streams?

**Figure 5.22:** Three atomic objects: three walls.

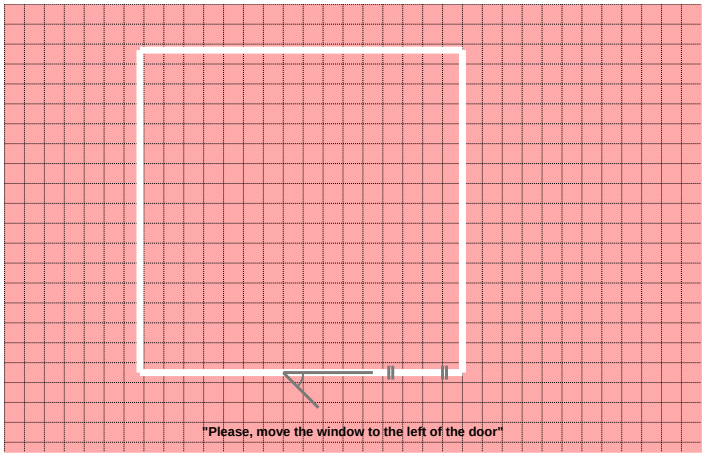



How do the subjects draw these four walls: producing one stream (a rectangular), or two, three or four (one for each wall) or more streams?

**Figure 5.23:** Four atomic objects: four walls.

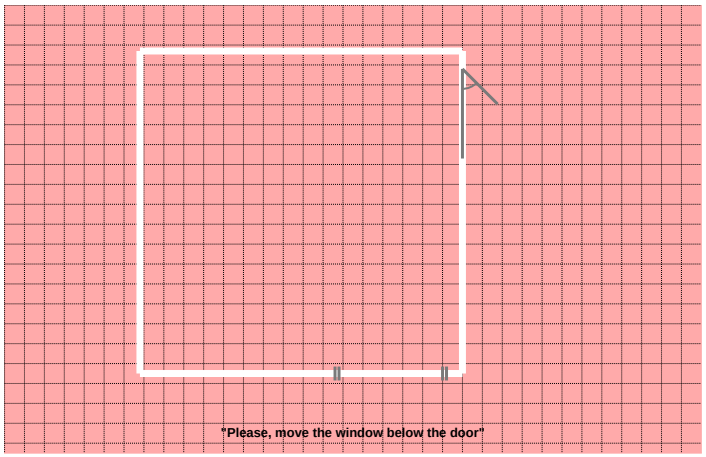
### 5.4.3 Moving Stimuli

The next ten stimuli request input from the subjects that is concerned with moving doors and windows. As explained in Chapter 4, a number of categories can be distinguished, each representing a certain expected manner in which subjects can move objects. During the experiments, the experimenters observed several cases that did not match our expectations. For example, some subjects tried first to erase the object to move and then to redraw it, which was not allowed. Notice that in the instructions, the experimenter explicitly said that, during one turn, the subject was able to erase or to write, but could not do both of them.



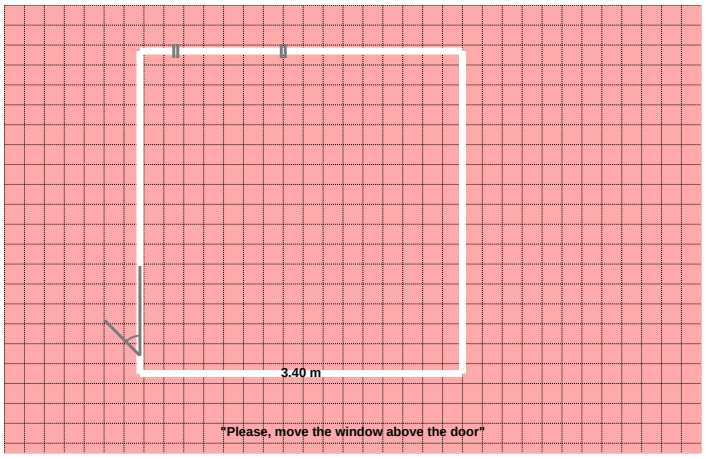

Absolute, relative, or relative distance moves are expected here.

**Figure 5.24:** Move a window.



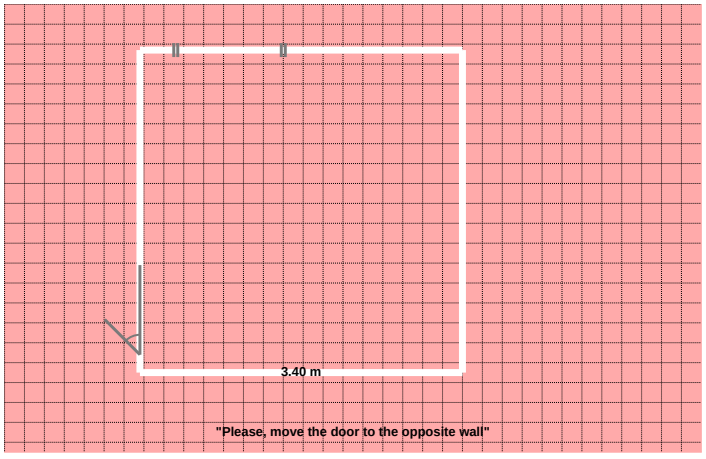

Absolute, relative, or relative distance moves are expected.

**Figure 5.25:** Move a window



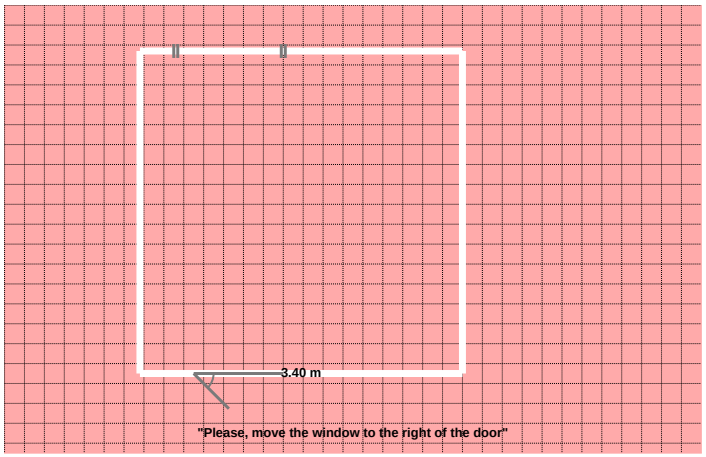

Absolute, relative, or relative distance moves are expected. Some of the subjects misunderstood “above the door”: they superimpose the door and the window. But they had no problem to understand “below the door” (see stimulus 5.25).

**Figure 5.26:** Move a window:



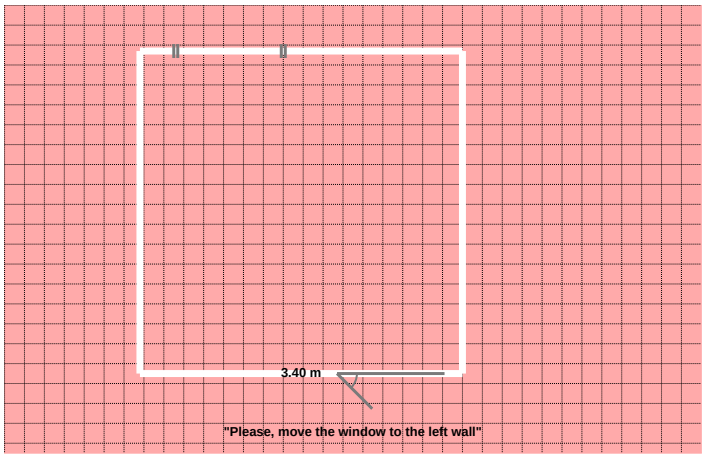

Absolute moves are expected.

**Figure 5.27:** Move a door:



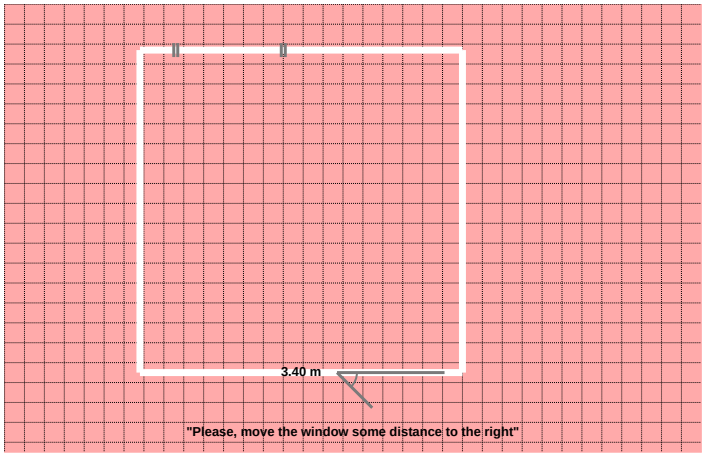

Absolute, relative, or relative distance moves are expected.

**Figure 5.28:** Move a window:



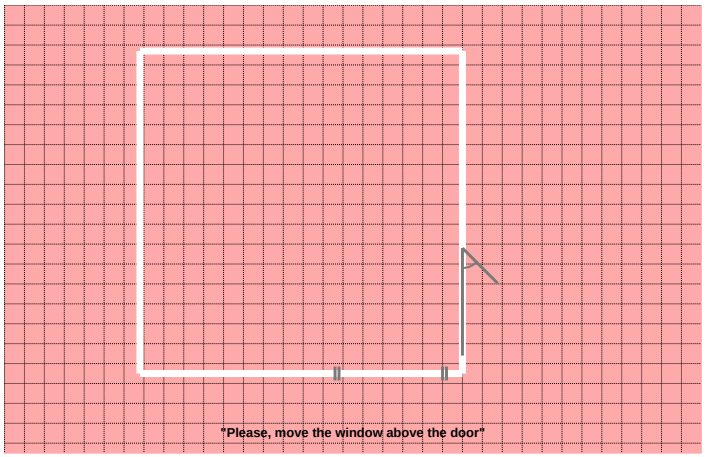

Absolute moves are expected here.

**Figure 5.29:** Move a window:



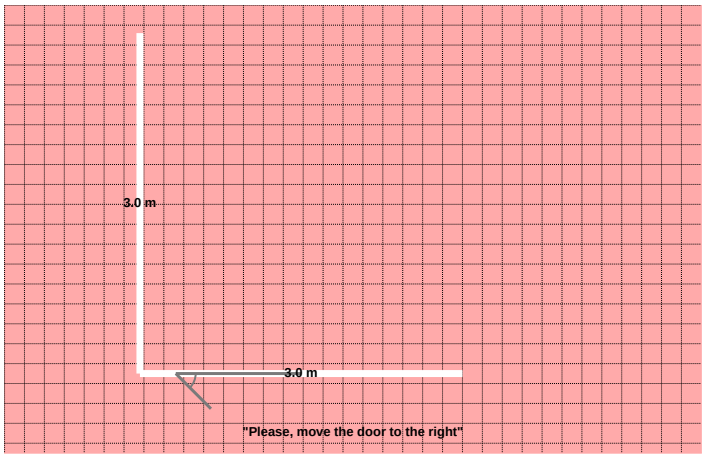

Absolute, distance or stepwise moves are expected.

**Figure 5.30:** Move a window:



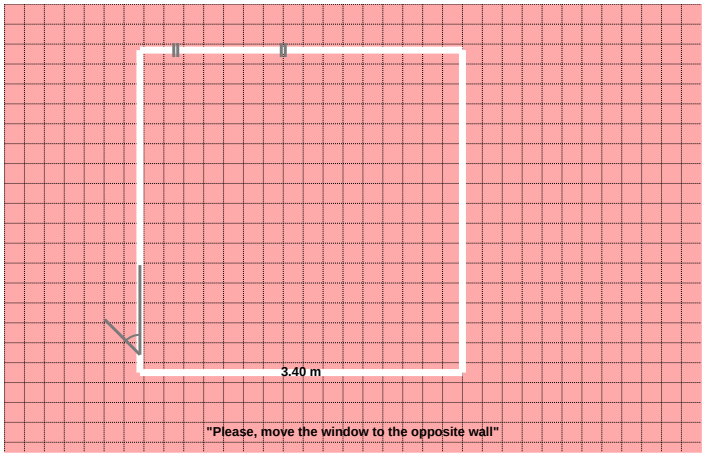

Absolute, relative, or relative distance moves are expected.

**Figure 5.31:** Move a window:




Absolute, distance or stepwise moves are expected.

**Figure 5.32:** Move a door:

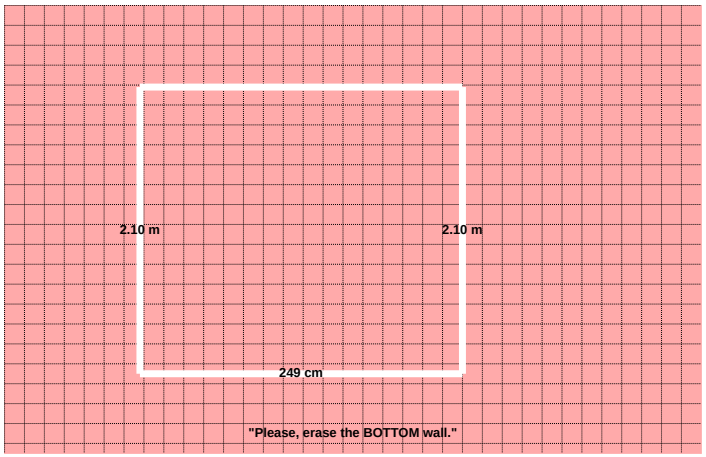



Absolute moves are expected.

**Figure 5.33:** Move a window:

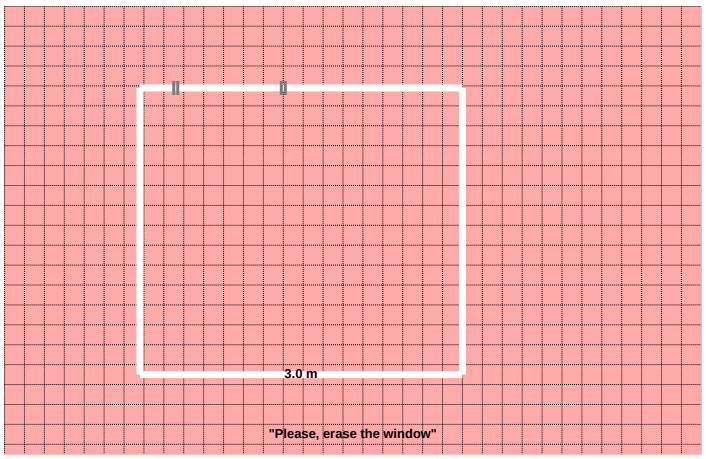
### 5.4.4 Erasing Stimuli

The final four stimuli request subjects to erase objects from the screen.




Do the subjects prefer to use pen or speech? When they use the pen, do they prefer to encircle the object to erase, or do they erase it like if they were using a rubber?

**Figure 5.34:** Erase a wall.




The window is beautified as four small segments. Some subjects erased only these four segments.

**Figure 5.35:** Erase a window.

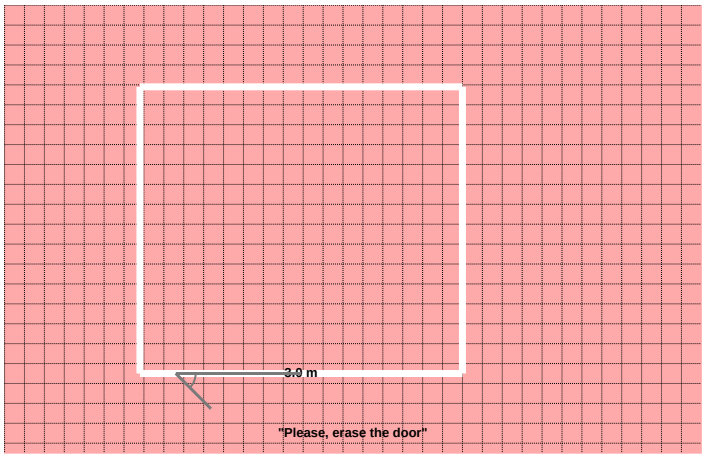



**Figure 5.36:** Erase a door.

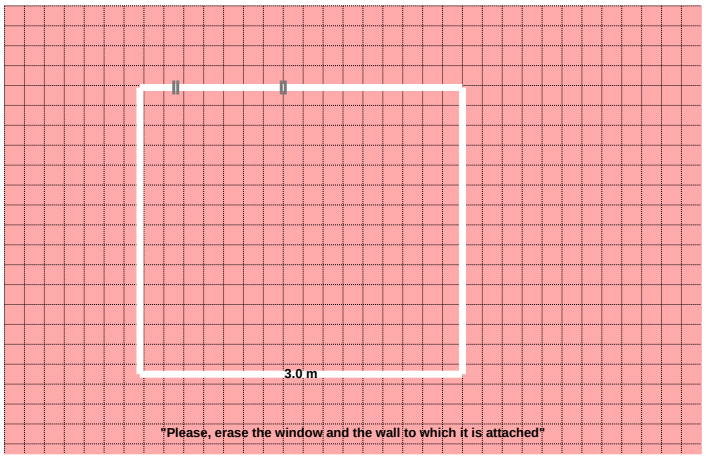



**Figure 5.37:** Erase two atomic objects: a wall and a window.

Here, it can be assessed how subjects erase multiple objects.

# Chapter 6

# Analysis

In this chapter, we first present an analysis of the T28 human factors experiments from the perspective of each involved module (Sections 6.1 to 6.4). Subsequently, an analysis of the questionnaire (section 6.5) is presented. As the research goals of each technology involved may differ from the others, the focus of each subsection differs. For example, the focus of ASR and FUSION lies on the suitability and effectiveness of the new turn taking protocol, whereas pen input recognition focuses on the new pen input repertoire associated with compound and spatial input.

A number of analysis tools were developed for extracting relevant information from the loggings. Loggings contained all messages sent by all modules and typically had the following structure for each stimulus:

1. **expectation**, sent by savWrap, and containing the identification of the stimulus;
2. **turn-taking information**, sent by all modules and indicating whether speech input and/or pen input was detected;
3. **multi-modal input**, detected by ASR and/or savWrap;
4. **unimodal hypotheses**, produced by ASR and/or PII;
5. **processed language**, generated by NLP;
6. **multi-modal hypotheses**, generated by the FUSION module.

The next sections show observations and statistics extracted from these loggings. Based on the number of expectations, the number of turns can easily be determined. For each turn, the identification of the stimulus indicates the kind of information that was expected. For example, the first stimulus requested a single wall as input. By examining the loggings on the availability of a speech and/or pen signal, it could be recorded whether the user response was performed through a single modality or multi-modally. By visually/manually examining each response in detail, the information discussed in this chapter could be extracted.

## 6.1 Automatic Speech Recognition

During the development of the T28 (and T30) versions of the ASR module, the functional aspects of the ASR module have been adapted by extensive testing, in cooperation with FUSION. These adaptations were mainly based on results obtained with a development set of multi-modal stimuli, that were used to guarantee that the output generated by ASR conforms with the expectations of NLP. The tests, which had the form of a systematic check of a set of over 30 input stimuli covering various multi-modal combinations of pen, speech, have led to a systematic adaptation of the language model, as well as an update of the acoustical model for garbage. The language model (LM) adaptation was not entirely straightforward: since the ASR LM spans

the set of all expressions that can be decoded, the NLP component in FUSION should be able to handle these. On top of that, T28 allows multiple inputs to be presented by the user, which implies the introduction of loops in the LM. However, the introduction of these loops showed to imply several challenges for NLP, especially in the parsing of multiple length specifications. An input of ”2 meter 3”, which was unambiguous in earlier experiments, becomes ambiguous in T28, because it might refer to the specification of two lengths (’two meters [and] three [meters]’) of which a measure is missing (or not recognized) in the second length specification. Furthermore, the expansion of the LM increases the range of possible ASR errors, such as ”2 meters and 30 centimeters” being recognized as ”2 meters 10 30 centimeters”. Without any preventive measures, the second hypothesis evidently leads to NLP interpretation problems. These issues have been solved as much as possible by simultaneously constraining the LM and adaptation of the parsing component in the NLP module.

From the annotated loggings of the T28 experiment, a number of examples have been extracted and shown below that illustrate the issues ASR has to solve. For each utterance, first the name of the log file is specified (its name includes the date and moment of creation). Next, time-stamped early hypotheses are presented (lines starting with integer index), followed by the N-best list (lines starting with ’h’) based on the full word lattice. The numbers at the beginning of each line correspond to frame numbers. The initial frame has index 0; hundred frames correspond to one second. From a certain point, the hypothesis remains fixed and is (usually) identical to the first hypothesis (indexed ’h0’) as provided in the sorted 10- best list (indexed ’h0’-’h9’). As mentioned above, differences between the final early hypothesis and the first best from the lattice are due to subtle differences in acoustic content at the end of the acoustic recording, and may also due to approximations in the construction of the word lattice in the forward pass.

### Example 1: Impact of garbage modeling

The utterance was ’two meters forty three centimeters’

```
/tmp/rossignol/recognized_speech_english/log/wavelog_020_2004-09-
15T15:23:53.781\
0   \
50  FORK1 &lt;sil&gt; FORK2  \
100  FORK1 GARB FORK2  \
150  FORK1 GARB FORK2  \
200  FORK1 two meters forty three FORK2  \
250  FORK1 GARB forty three centimeters FORK2  \
300  FORK1 GARB forty three centimeters FORK2  \
h0 FORK1 GARB forty three centimeters FORK2
h1 FORK1 GARB forty three centimeters FORK2
h2 FORK1 GARB forty three centimeters FORK2
h3 FORK1 GARB forty three centimeters FORK2
h4 FORK1 GARB forty three centimeters FORK2
h5 FORK1 GARB forty three centimeters FORK2
h6 FORK1 GARB forty three centimeters FORK2
h7 FORK1 GARB forty three centimeters FORK2
h8 FORK1 two meters forty three centimeters FORK2
h9 FORK1 GARB forty three centimeters FORK2
```

This example shows that the garbage model (for this case) is cheaper than desired. The competition with the model for ’two meters’ that is evident in the N-best list, shows an undesirable effect of the garbage penalty for this particular example. The tuning of garbage is critical: while the average log-likelihood score of a word like ’two’ is in the order of 700-1000, a modification of the garbage penalty by a value of 2 has a clear impact on the decoding results. The example also shows that it might be advantageous for the NLP to also look at the early recognition results - something that is not done in the current system.

**Example 2: Compound ASR decoding may yield errors that can be solved by expectation-sensitive NLP**

The utterance for this example: ”Two meters and forty three centimeters.”

```
/tmp/rossignol/recognized_speech_english/log/wavelog_021_2004-09-
15T15:24:03.629\

0   \
50  FORK1 &lt;sil&gt; FORK2  \
100  FORK1 &lt;sil&gt; FORK2  \
150  FORK1 GARB FORK2  \
200  FORK1 two meters FORK2  \
250  FORK1 two meters ten FORK2  \
300  FORK1 two meters ten &lt;sil&gt; GARB FORK2  \
350  FORK1 two meters eighty one GARB FORK2  \
400  FORK1 two meters and forty three centimeters FORK2  \
450  FORK1 two meters and forty three centimeters FORK2  \
500  FORK1 two meters and forty three centimeters FORK2  \
h0 FORK1 two meters and forty three centimeters FORK2
h1 FORK1 two meters ten forty three centimeters FORK2
h2 FORK1 two meters and forty three centimeters FORK2
h3 FORK1 two meters and forty three centimeters FORK2
h4 FORK1 two meters ten forty three centimeters FORK2
h5 FORK1 two meters ten forty three centimeters FORK2
h6 FORK1 two meters ten forty three centimeters FORK2
h7 FORK1 two meters and forty three centimeters FORK2
h8 FORK1 two meters and forty three centimeters FORK2
h9 FORK1 two meters and forty three centimeters FORK2
```

The LM both allows an atomic utterance as well as a compound utterance. Hypothesis h0, which is in line with the last early hypothesis, is decoded according to one atomic hypothesis. Among the N-best candidates, however, a number of hypotheses support the compound decoding: h1 is to be interpreted as a specification of one length (’one meter 10’) followed by another length (’forty three centimeters’). The ASR LM does not default to any of these interpretations. In principle, such biases could be trained on a sufficiently large corpus. An expectation-sensitive NLP might be able to resolve this type of ambiguity.

**Example 3: Unambiguous decoding**

The utterance was: ”One meter”

```
/tmp/rossignol/recognized_speech_english/log/wavelog_022_2004-09-
15T15:24:38.034\
0   \
50  FORK1 &lt;sil&gt; FORK2  \
100  FORK1 one meter FORK2  \
150  FORK1 one meter FORK2  \
h0 FORK1 one meter FORK2
h1 FORK1 one meter FORK2
h2 FORK1 one meter FORK2
h3 FORK1 one meter FORK2
h4 FORK1 one meter FORK2
h5 FORK1 one meter FORK2
```

```
h6 FORK1 one meter nine FORK2
h7 FORK1 one meter eight FORK2
h8 FORK1 one meter FORK2
h9 FORK1 one meter FORK2
```

This decoding is straightforward and correct.

#### Example 4: Out of grammar

The utterance was: ”And the right wall”

```
/tmp/rossignol/recognized_speech_english/log/wavelog_029_2004-09-
15T15:26:42.499\
0   \
50  FORK1 there FORK2  \
100  FORK1 GARB FORK2  \
150  FORK1 GARB FORK2  \
200  FORK1 GARB FORK2  \
250  FORK1 GARB FORK2  \
300  FORK1 GARB FORK2  \
350  FORK1 GARB FORK2  \
400  FORK1 GARB &lt;sil&gt; GARB FORK2  \
450  FORK1 GARB &lt;sil&gt; GARB FORK2  \
500  FORK1 GARB &lt;sil&gt; GARB FORK2  \
550  FORK1 GARB &lt;sil&gt; GARB two meters three FORK2  \
600  FORK1 GARB &lt;sil&gt; GARB two meters three FORK2  \
650  FORK1 GARB &lt;sil&gt; GARB two meters three &lt;sil&gt; GARB
FORK2  \
h0 FORK1 GARB FORK2
h1 FORK1 GARB FORK2
h2 FORK1 GARB FORK2
h3 FORK1 GARB FORK2
h4 FORK1 there GARB FORK2
h5 FORK1 GARB FORK2
h6 FORK1 GARB FORK2
h7 FORK1 there GARB FORK2
h8 FORK1 GARB FORK2
h9 FORK1 GARB FORK2
```

In this case, there is a difference between the final early hypothesis and h0 - this difference is due to the way in which the forward pass builds the word lattice. The utterance was not in the grammar, and the decoding is not straightforward, but, given the non-grammaticality, correct.

#### Example 5: Out of grammar

The utterance was: ”Three meters both”

```
/tmp/rossignol/recognized_speech_english/log/wavelog_030_2004-09-
15T15:27:02.535\
0   \
50  FORK1 &lt;sil&gt; FORK2  \
100  FORK1 &lt;sil&gt; FORK2  \
150  FORK1 GARB FORK2  \
200  FORK1 three meters FORK2  \
```

```
250  FORK1 three meters four FORK2  \
300  FORK1 three meters four FORK2  \
h0 FORK1 three meters four FORK2
h1 FORK1 three meters GARB FORK2
h2 FORK1 three meters four FORK2
h3 FORK1 three meters GARB FORK2
h4 FORK1 three meters GARB FORK2
h5 FORK1 three meters four FORK2
h6 FORK1 three meters GARB FORK2
h7 FORK1 three meters GARB FORK2
h8 FORK1 three meters GARB FORK2
h9 FORK1 there three meters four FORK2
```

Also this utterance is not according to the grammar. Both the h0 as the h1 hypotheses are the closest possible within the grammar. In this case, the GARB model is too expensive to have the decoding 'three meters GARB' winning.

#### Example 6: Correct decoding of an erase command

The utterance was: "<breath> erase the bottom wall"

```
/tmp/rossignol/recognized_speech_english/log/wavelog_044_2004-09-
16T15:12:24.826\
0   \
50  FORK1 &lt;sil&gt; FORK2  \
100  FORK1 two meters FORK2  \
150  FORK1 erase the door FORK2  \
200  FORK1 erase the bottom wall FORK2  \
250  FORK1 erase the bottom wall FORK2  \
h0 FORK1 erase the bottom wall FORK2
h1 FORK1 erase the bottom wall FORK2
h2 FORK1 erase the bottom wall FORK2
h3 FORK1 GARB erase the bottom wall FORK2
h4 FORK1 erase the bottom wall FORK2
h5 FORK1 erase the bottom wall FORK2
h6 FORK1 erase the bottom wall here FORK2
h7 FORK1 erase the bottom wall FORK2
h8 FORK1 erase the bottom wall there FORK2
h9 FORK1 GARB erase the bottom wall FORK2
```

The utterance is in the grammar. The decoding is correct.

#### Example 7: ASR goes wrong due to garbage penalty setting

The utterance was: "please move the window to the left wall <breath>"

```
/tmp/rossignol/recognized_speech_english/log/wavelog_046_2004-09-
16T16:39:50.448\
0   \
50  FORK1 &lt;sil&gt; FORK2  \
100  FORK1 two meters FORK2  \
150  FORK1 erase the door there FORK2  \
200  FORK1 erase the window GARB FORK2  \
250  FORK1 erase the window GARB FORK2  \
```

```
300  FORK1 erase the window GARB FORK2  \
350  FORK1 erase the window GARB FORK2  \
h0 FORK1 erase the window GARB FORK2
h1 FORK1 erase the window GARB FORK2
h2 FORK1 erase the window GARB FORK2
h3 FORK1 erase the window GARB FORK2
h4 FORK1 erase the window GARB FORK2
h5 FORK1 erase the window GARB FORK2
h6 FORK1 erase the window GARB GARB FORK2
h7 FORK1 two meters move the window to the left wall GARB FORK2
h8 FORK1 erase the window GARB there GARB FORK2
h9 FORK1 erase the window GARB FORK2
```

Although the utterance could be decoded correctly by using the optional garbage word, the correct hypothesis is too expensive compared to the winning (but incorrect) hypotheses. This example shows the difficulty the ASR has once the word lattice has started along a false partial hypothesis.

### Example 8: Correct interpretation of out-of-grammar input

The utterance was: ”you should move the window to the left wall <breath>”

```
/tmp/rossignol/recognized_speech_english/log/wavelog_047_2004-09-
16T16:40:01.784\
0   \
50  FORK1 &lt;sil&gt; FORK2  \
100  FORK1 &lt;sil&gt; FORK2  \
150  FORK1 here FORK2  \
200  FORK1 here GARB FORK2  \
250  FORK1 here is a window FORK2  \
300  FORK1 here GARB move the window to the left FORK2  \
350  FORK1 here GARB move the window to the left wall FORK2  \
400  FORK1 here GARB move the window to the left wall GARB FORK2  \
450  FORK1 here GARB move the window to the left wall GARB FORK2  \
h0 FORK1 here GARB move the window to the left wall GARB FORK2
h1 FORK1 GARB move the window to the left wall GARB FORK2
h2 FORK1 here GARB move the window to the left wall GARB FORK2
h3 FORK1 here GARB move the window to the left wall GARB FORK2
h4 FORK1 GARB move the window to the left wall GARB FORK2
h5 FORK1 GARB move the window to the left wall GARB FORK2
h6 FORK1 here GARB move the window to this wall GARB FORK2
h7 FORK1 here GARB here move the window to the left wall GARB FORK2
h8 FORK1 here GARB move the window to the left wall GARB FORK2
h9 FORK1 here GARB move the window to the left wall GARB FORK2
```

Although the utterance is not according to the used grammar, the decoding is able to locate the right fragments (as sub-grammars). Despite the high word error rate (33 percent) according to the default ASR performance assessment, the interpretation of the utterance should be unambiguous for the subsequent modules.

### Example 9: Compound input

The utterance was: ’the top wall is three meters the left wall is three meters’

```
/tmp/rossignol/recognized_speech_english/log/wavelog_036_2004-09-
17T12:33:11.013\
```

```
0   \
50  FORK1 GARB FORK2  \
100  FORK1 GARB FORK2  \
150  FORK1 GARB FORK2  \
200  FORK1 GARB three meters FORK2  \
250  FORK1 GARB three meters eleven FORK2  \
300  FORK1 GARB three meters eleven GARB FORK2  \
350  FORK1 GARB three meters eleven GARB FORK2  \
400  FORK1 GARB three meters eleven GARB three meters FORK2  \
450  FORK1 GARB three meters eleven GARB three meters FORK2  \
500  FORK1 the size of the wall is three meters &lt;sil&gt; GARB here
is a wall GARB FORK2  \
550  FORK1 the size of the wall is three meters &lt;sil&gt; GARB here
is a door three meters FORK2  \
600  FORK1 the size of the wall is three meters &lt;sil&gt; GARB here
is a door three meters FORK2  \
650  FORK1 the size of the wall is three meters &lt;sil&gt; GARB here
is a door three meters FORK2  \
h0 FORK1 GARB three meters eleven GARB three meters FORK2
h1 FORK1 GARB three meters eleven GARB three meters FORK2
h2 FORK1 there GARB three meters eleven GARB three meters FORK2
h3 FORK1 GARB three meters eleven GARB three meters FORK2
h4 FORK1 GARB three meters eleven GARB three meters FORK2
h5 FORK1 there GARB three meters eleven GARB three meters FORK2
h6 FORK1 GARB three meters eleven GARB three meters FORK2
h7 FORK1 GARB three meters eleven GARB three meters FORK2
h8 FORK1 GARB three meters eleven GARB three meters here FORK2
h9 FORK1 there GARB three meters eleven GARB three meters FORK2
```

This example, showing the results of a compound input utterance, shows that the interpretation of the ASR outcome is not always evident. Evidently, the interpretation of the N-best list (h0-h9) is simplified by taking into account the results from early recognition. For this particular case, the garbage model has the tendency to be too cheap.

### Example 10: Correct ASR decoding of input outside ontology

The utterance was: 'Move the window to the opposite wall'

```
/tmp/rossignol/recognized_speech_english/log/wavelog_090_2004-09-
16T12:02:08.286\
0   \
50  FORK1 there FORK2  \
100  FORK1 here is a window FORK2  \
150  FORK1 here is a window GARB FORK2  \
200  FORK1 move the window to the opposite wall FORK2  \
250  FORK1 move the window to the opposite wall FORK2  \
300  FORK1 move the window to the opposite wall FORK2  \
h0 FORK1 move the window to the opposite wall FORK2
h1 FORK1 move the window to the opposite wall FORK2
h2 FORK1 move the window to the opposite wall FORK2
h3 FORK1 move the window to the opposite wall FORK2
h4 FORK1 move the window to the opposite wall here FORK2
h5 FORK1 move the window to the opposite wall FORK2
```

```
h6 FORK1 move the window to the opposite wall FORK2
h7 FORK1 move the window to the opposite wall here FORK2
h8 FORK1 move the window to the opposite wall there FORK2
h9 FORK1 move the window to the opposite wall here FORK2
```

This utterance has been decoded correctly. This example is interesting, however, because the example shows in another way the incapability of the system to correctly process it due to a mismatch between input and ontology. In the section discussing the Natural Language Processing (NLP) module, this utterance falls therefore in the category 'Not Representable'. Although the ASR LM can handle the utterance and the specified wall object 'opposite wall' itself is part of the ontology, the adverb 'opposite' itself is not. This prohibits the NLP to attach any semantic labeling, thereby blocking the entire further processing of the input.

### Global analysis

On the basis of a partially annotated set of acoustic logfiles (412 of the 1491 wavefiles have been annotated for this analysis), the following results have been obtained:

- Of the 412 utterances that have been annotated, 120 (29%) are out-of-grammar; the remaining 291 utterances (71%) that are within-grammar include utterances that contain only silence.
- On the set of all 291 within-grammar utterances, the estimated word error rate is 28% (this result to be made more precise after completion of annotation). The number of compound utterances within grammar equals 141 (almost 50%). These compound utterances correspond to all prompts requesting for compound information, that is, prompt number 9, 10, 11, 12, 13, 14, 15, 30 (see below for a list of all system prompts).

```
1 - Please, enter the TOP wall
2 - Please, enter the length of the TOP wall
3 - Please, enter a door in one of the walls
4 - Please, enter a window in one of the walls
5 - Please, enter the LEFT wall
6 - Please, enter the length of the RIGHT wall
7 - Please, specify the height of the window
8 - Please, enter the width of the window
9 - Please, enter the TOP wall and specify its length
10 - Please, enter the RIGHT wall and specify its length
11 - Please, enter the RIGHT wall and complete all wall length
specifications
12 - Please, enter the vertical walls and specify all missing wall
lengths
13 - Please, complete the outline of the bathroom and specify all
wall lengths
14 - Please, specify the remaining two lengths
15 - Please, enter the remaining three walls
16 - Please, enter the complete outline of the bathroom
17 - Please, move the window to the left of the door
18 - Please, move the window below the door
19 - Please, move the window above the door
20 - Please, move the door to the opposite wall
21 - Please, move the window to the right of the door
22 - Please, move the window to the left wall
23 - Please, move the window some distance to the right
24 - Please, move the window above the door
25 - Please, move the door to the right
```

```
26 - Please, move the window to the opposite wall
27 - Please, erase the BOTTOM wall
28 - Please, erase the window
29 - Please, erase the door
30 - Please, erase the window and the wall to which it is attached
```

- A global analysis of the user utterances reveals the following. In general, subjects echo the written system prompt. A prompt such as 'please, enter the length of the right wall' will typically be replied to by either '2 meters', 'the length is 2 meters', 'the length of this wall is 2 meters', or by 'the length of the right wall is two meters'. The influence of the prompt formulation on the user reply (choice of lexical units, syntax) is substantial. Actually, most length specifications were given by the raw size, e.g. one meter and fifteen centimeter. Sometimes context was added, such as in 'the width is eighty five centimeters'. As indicated in the section on NLP, the current ontology does not support the reference of walls using relative indicators such as 'top', 'bottom', 'left', 'right'. Some users use variants, such as like a command 'window above the door'. These constructions are not grammatical according to the ASR LM. 'Put' and 'move' in the 'move' commands are used interchangeably (both these options are grammatical according to the ASR LM).
- In a number of cases, subjects use the speech modality to enter pictorial information. The application is not designed to handle these inputs. Processing of such utterances by the ASR is not the essential issue in such cases; the difficulty is in the interpretation of the utterance within an ontology, and in the relation between the ontology and the screen state.
- As already shown in the T16 experiment, subjects exploit a wide variety of possible utterances to present information to the system. This is especially true for naive subjects. As observed above, naive subjects might not have the remotest idea about what they can say to the system. The impact of the difference in behavior between naive and experienced users on the performance of the dialog system is well documented in the literature. The difference in experience between users correlate with large differences in task completion rates as shown by complex multi-modal systems. Also for professional, commercially applied dialog systems that are tuned on hundreds of thousands of utterances, these performance differences due to difference in expertise level remain.

## 6.2 Pen Repertoire

An important research issue in the COMIC project pursues the role of human factors in the design of computer supported interaction. This section provides more insight in the human factor aspects of pen input. The goal of this exploration is to find out what pen gestures or handwriting is employed by subjects for entering compound information (several walls, several lengths, walls and lengths), deictic gestures (tapping, encircling, erasing), or spatial moving commands. In [7], an elaborate overview of atomic pen gestures was presented, including walls and windows. Therefore, these classes will not be discussed in this section.

All data recorded during the experiments were automatically analyzed using the tool developed for the analysis of the T16 experiments [8]. This tool outputs UNIPEN data, labeled with the recognition hypotheses from the input recognizer. Using visualization tools for online pen data, we visually categorized all data into a number of classes. For example, for spatial moving gestures, 16 categories and 15 types of arrow-shaped gestures were distinguished.

The resulting taxonomies presented in this section will be used to improve the pen input recognition technologies to be delivered at T32 and to be incorporated in the final demonstrator. As discussed before, the new turn-taking protocol and the possibility to enter compound information provide far more challenging pen input. For example, the order of pen gestures may differ for one and the same compound input (e.g., a subject may first write down a wall length before drawing the wall, or vice versa). The question is to assess the performance of the current pen input interpretor and to identify cases for which it was not designed yet.

The organization of this section reflects the issues mentioned in the previous chapters. In the next section, the results of entering compound objects are presented (stimuli 9 to 16 described in Section 5.4.2). Section 6.2.2 describes the results for spatial moving commands (stimuli 17 to 26 described in Section 5.4.3). In Section 6.2.4, a brief overview over deictic gestures is given, with a focus on erasing gestures (stimuli 27 to 30 described in Section 5.4.4).

### 6.2.1 Compound Gestures

As described in Section 5.4.2, 8 stimuli were presented to the subjects, requesting compound information. These stimuli were particularly designed to test the capabilities of the pen input recognizer to distinguish walls from handwriting, walls from other walls, and lengths from other lengths. A semi-automated analysis of the recorded loggings revealed the statistics described in Table 6.1. Given that the stimuli requested for combinations of walls and lengths, the pen-based user responses could contain one or more walls and one or more lengths. The recorded user input is distinguished in categories based on the number of walls and the number of lengths each response contained. For example, the first row reads as follows: "Stimuli (S) numbered 9 and 10 requested subjects to enter one wall plus one length. In total, this compound information was requested 172 times (total). In one case, the user did not use the pen (no pen). In 78 cases, only one single wall (1w) was entered, in 2 cases, a single length (1l) was input. In 91 cases, the pen input contained all required information, i.e., one wall and one length (w+l)." The latter cases, indicating that all input is available in the pen signal, are marked with a box. Note that all cases may be accompanied by speech, yielding multi-modal input. An analysis of multi-modal input is presented in Section 6.4.

| | | | #walls | | | | #lengths | | lengths + walls | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S | requests | no pen | 1w | 2w | 3w | 4w | 1l | 2l | w+l | w+2l | 2w+l | 2w+2l | 3w+3l | 4w+4l | total |
| 9+10 | 1w + 1l | 1 | 78 | | | | 2 | | **91** | | | | | | 172 |
| 11 | 1w + 2l | 1 | 45 | | | | 6 | | 14 | **138** | | | | | 204 |
| 12+13 | 2w + 2l | 1 | 30 | 53 | | 1 | 1 | | 16 | | 33 | **153** | | | 288 |
| 14 | 2l | | | | | | 6 | **51** | | | | 1 | | | 58 |
| 15 | 3w | 2 | 22 | 8 | **63** | | | | 1 | | | | 5 | | 101 |
| 16 | 4w | | 7 | 1 | 3 | **44** | | | | | | | | 3 | 58 |
| total | | 5 | 182 | 62 | 66 | 45 | 15 | 51 | 122 | 138 | 33 | 154 | 5 | 3 | 881 |

**Table 6.1:** User input divided over compound categories. Each row contains a distinction of pen-based user input in response to a certain request.

From the 881 requests, users generated compound pen information in 679 cases. As the focus of our research was targeted at extracting automatic objects from these compounds, we consider this as a success. Furthermore, when considering Table 6.1, it is noted that for each request, the larger part of the user responses contains all required information. In 91/172=52.9% cases, users entered one wall and one length via pen upon being asked to do. Similarly, in 67.4%, users replied one wall and two lengths via pen, 53.1% cases contained 2 walls and 2 lengths, 87.9% contained two lengths, 62.4% 3 walls and 75.9% contained 4 walls.

#### Exceptional user input

In the development of the T28 system, WP3 and WP4 have tried to prepare the underlying recognition technology for all kinds of user input. However, a number of cases were observed that we did not anticipate:

- In several cases, subjects tried to enter a wall via speech only, e.g. by saying "Please enter the top wall". These cases were rejected by the system.

- In other cases, subjects encountered problems in using the pen. Our expectations were that for entering information, subjects would not press the side buttons of the pen. However, in some occasions, subjects erroneously pressed these buttons, although the instructions explicitly warned for this situation.
- Other particular cases that were rejected by the pen module were caused by subjects entering one length on several lines, or subjects that employed a diagonal or vertical writing direction.
- Some subjects drew non-rectangular bathrooms upon being asked to draw four walls, containing 6 or even 7 single walls.
- Some subjects spontaneously entered windows and doors upon being asked to draw four walls.
- Only few subjects entered a corresponding length for each wall that was entered for stimuli 15 (three walls) and 16 (four walls).

In Section 6.2.1, the recognition performance for compound objects and the number of rejects made by the system are described.

### Object ordering in pen input

The stimuli requesting compound information were of different complexity. The more information requested, the more subjects had the possibility to vary the order in which the individual objects were entered. This was also observed above, where in the case of the request to enter four walls, some subjects entered extra information that was not requested. However, as the list of observations below indicates, in most cases subjects first drew wall(s) and subsequently entered the corresponding length(s). This information is important for the distinction between handwriting and drawing modes. Mode detection is a segmentation step that precedes recognition and therefore, knowledge about the order in which subjects enter heterogeneous (consisting of multiple modes) compound information can be used to increase the confidence in recognition.

1w+1l In all 91 cases, subjects first entered the wall and subsequently specified its length.

1w+2l In response to this request, subjects entered one wall and one length in 14 cases. In 7 of these cases, the written lengths concerned the length of the drawn wall, in the other 7 cases, subjects entered the length of the other wall. In one particular case, subjects first wrote down the length of the other wall and subsequently drew the wall, so in 13/14 cases, the wall preceded the length. Next to that, in total 138 responses contained one wall and two lengths. First drawing the wall and subsequently entering two lengths was done in 117 (84.8%) cases. In 18 (13.0%) of the cases, subjects started by entering a length, subsequently drew the wall and finally entered the second length. In two particular cases, a subject started by entering a length; subsequently drew a wall and finally wrote two lengths. In another particular case, one subject began to write a length, then stopped and entered the wall, and finally finished by writing the second length.

2w+2l The following table shows how subjects entered two walls and the corresponding lengths using the pen. The 154 observed cases were distinguished as:

| | | |
|---|---|---|
| $wall_0$; $wall_1$; $length_0$; $length_1$ | 68 (44.1%) | |
| $wall_0$; $length_0$; $wall_1$; $length_1$ | 48 (31.2%) | In the last category, 36 cases were encountered where |
| { $wall_0$, $wall_1$ }; $length_0$; $length_1$ | 36 (23.4%) | |

subjects drew 'L'-shaped walls. It is apparent that in 151 out of 154 cases, a wall always precedes a corresponding length. In one of the two remaining cases, one subject started a length, did not finish it and subsequently drew $wall_0$; $length_0$; $wall_1$; $length_1$. In the other remaining case, another subject drew the first wall twice: $wall_0$; $wall_1$; $length_0$; $wall_0$; $length_1$.

As mentioned above, the conclusion that walls precede lengths in the majority of cases will be incorporated in the next generation of mode detection algorithms, to be delivered at T32.

### Stream occurrence in multiple walls

The trajectory of subsequent pen coordinates between two pen lifts is called a stream. The majority of subjects enter a wall via one stream, either a horizontal or vertical line. For the benefit of the required recognition algorithms, knowledge about the way in which subjects produce multiple walls is important. The responses to the questions where 3 walls (63 cases) or 4 walls (44 cases) were requested are analyzed below.

| request | nstreams | noccurrences | shape description |
|---|---|---|---|
| 3 walls | 1 | 30 | 'U'-shaped |
| 3 walls | 2 | 3 | 'L'-shaped followed by a line |
| 3 walls | 3 | 21 | Three atomic walls |
| 3 walls | 4 | 6 | Three atomic walls and two extra lines |
| 3 walls | 5 | 3 | Three atomic walls and two extra lines |
| 4 walls | 1 | 31 | rectangular shaped |
| 4 walls | 2 | 3 | 'U'-shaped followed by a line |
| 4 walls | 3 | 0 | |
| 4 walls | 4 | 8 | Four atomic walls |
| 4 walls | 5 | 2 | Four atomic walls and one extra line |

What is noted is that in order to draw three walls, in 33% of the cases subjects entered three atomic walls. On the other hand, in only 18% of the cases, subjects entered four atomic walls to draw the complete outline of the bathroom. The reason why subjects draw relatively more rectangular shapes then 'U'-shaped shapes using one stream is difficult to assess. It may be that the existence of one example wall in the stimuli requesting for three walls influenced subjects in their way of generating the remaining three walls. Nevertheless, these findings show that 'L'-shapes occur relatively less frequently than 'U' or rectangular shaped compound walls.

### Performance: turn recognition rates

An in depth analysis and annotation of all entered compound objects resulted in Table 6.2 below. In this table, all encountered responses are distinguished as before in categories containing one or more walls and/or one or more lengths. For each category, the total number of cases (n), the number of cases that were rejected by the system (nrej) and the number of cases correctly recognized by the system (nok) are listed. Recognition performance (rec) is computed by the dividing the number of correct cases by the total number of cases (nok/n):

| category | n | nrej | nok | 100*(nok/n) | 100*nok/(n-nrej) |
|---|---|---|---|---|---|
| 1 wall | 182 | 10 | 148 | 81.3% | 86.0% |
| 2 walls | 62 | 8 | 51 | 82.3% | 94.4% |
| 3 walls | 66 | 3 | 52 | 78.8% | 82.5% |
| 4 walls | 45 | 0 | 40 | 88.9% | 88.9% |
| 1 length | 15 | 3 | 8 | 53.3% | 66.7% |
| 2 lengths | 51 | 0 | 28 | 54.9% | 54.9% |
| 1 wall + 1 length | 122 | 23 | 54 | 44.3% | 54.5% |
| 1 wall + 2 lengths | 138 | 15 | 55 | 39.9% | 44.7% |
| 2 walls + 1 length | 33 | 5 | 8 | 24.2% | 28.6% |
| 2 walls + 2 lengths | 153 | 13 | 47 | 30.7% | 33.6% |
| 3 walls + 3 lengths | 5 | 0 | 2 | 40.0% | 40.0% |
| 4 walls + 4 lengths | 3 | 0 | 0 | 0.0% | 0.0% |
| total | 875 | 80 | 493 | 56.3% | 67.8% |

**Table 6.2:** Recognition performance of PII. Each of the rejected cases actually did not contain a valid input of the user. Therefore, the recognition performance listed in the right-most column is considered as more appropriate.

Turn recognition performance (i.e., in order to judge a recognition result as correct, none of the information entered in one turn may be incorrect) is relatively low, as the average 56.3% indicates. However, since all the rejected cases did not contain any valid input of the user, the listed 67.8% performance number should be used. And, since most performance numbers refer to compound objects and turn recognition rates are presented here, we do not consider these numbers as particularly bad.

Furthermore, it should be noted that a high recognition performance was not the goal of the current study. Rather, we wanted to explore for what cases our algorithms apparently failed. Nevertheless, it is clear that the recognition algorithms have to be adapted and tuned. The acquired data will be used to improve the current technologies, to be delivered at T32.

### 6.2.2 Moving Gestures

In this section, the results on the explorations in "pen input in the context of spatially moving objects" are described. No recognition results are provided yet, although a subjective impression of the experimenters as well as the qualitative analyzes performed by our team indicate that in most cases, the recognizer had no difficulties in recognizing the pen input. Basically, each pen input could indicate:

- a particular location (i.e., *source* location of the object to be moved, or target *destination* location where the object should be placed.).
- a particular move comprising *source* and *destination*

What we call *moving gestures* are triggered by stimuli 17 to 26 described in Section 5.4.3. The subjects did not receive any special instructions about how to move the objects (i.e., windows and doors). Subjects only knew that to perform this request they could employ both pen and speech. Therefore, subjects were invited to provide multi-modal information entered in a natural fashion, in the way they preferred. It was expected that the subjects would largely differ in their behavior. From the pen input point of view, this has been noticed: the "moving gestures" were categorized into 15 classes. As a particularly frequent class of gestures contained arrow-like shapes, the next subsection is dedicated on the taxonomy of 15 arrow classes that were the result of our analyzes.

### 6.2.3 Observed arrow repertoire

A large part of the ways subject employ the pen to indicate spatial move operations contains arrows. Such an arrow may point at a certain location (source or destination), or its tail may start at the source location, with its head ending at the destination. Note that in the document [7], which describes our pilot studies with unconstrained pen input in bathroom design applications, a repertoire of the observed arrows has been presented. However, the current repertoire was acquired in another context. In [7], subjects only used arrows to "link" objects to each other, whereas this is not the case in the current experiments. For example, the arrows produced during the pilot-experiments described in [7] were used to indicate to which "wall" a given "measure" was associated, or to which wall a certain window belonged. The majority of such shapes were rounded lines or arrows with two heads.

The arrows acquired in the current T28 Human Factors experiments can be classified into fifteen categories. The categories were obtained through visual inspections of the logged data. Criteria were the shape of the tail and the shape of the head. Tails could be (i) straight, (ii) rounded, or (iii) angular. Next to arrows without a head, two further distinctions could be made between heads: connected or separated from the tail. Furthermore, two distinctions in the shape of the head were observed: drawn in a 'v'-like shape (simple heads) and drawn as a closed triangle (complete heads). Figure 6.1) displays the resulting models for arrows.

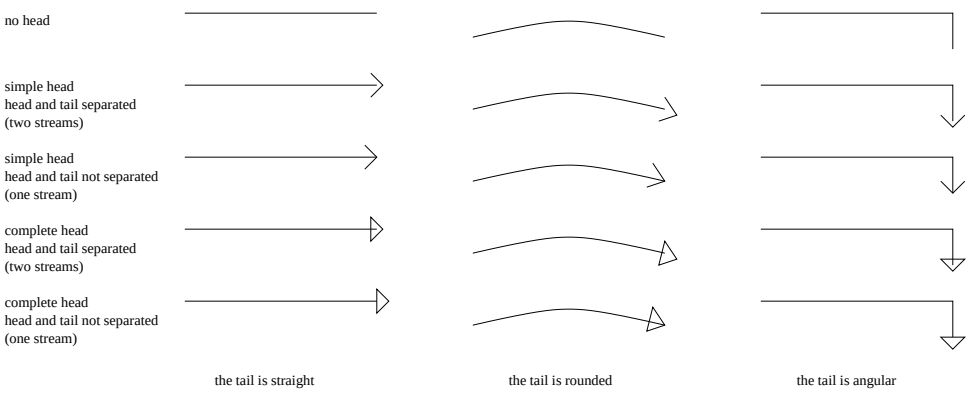



**Figure 6.1:** The fifteen classes of arrows

### A taxonomy of moving gestures

As mentioned above, a first qualitative analyzes of the moving gestures was performed manually. Two experts on pen computing examined all loggings on turns where the loggings contained pen input. From these loggings, in total 546 turns contained moving gestures, which are classified into 16 classes as described below. For each of the classes, an example figure is added. In case such figures contain an arrow, it should be noted that any of the arrow categories described above may apply.

**- 01.** "Moving arrows":

The subject draws an arrow, with the tail starting at or near the object to be moved. The head of the arrow points at its target position. Two examples of such arrows are provided below.

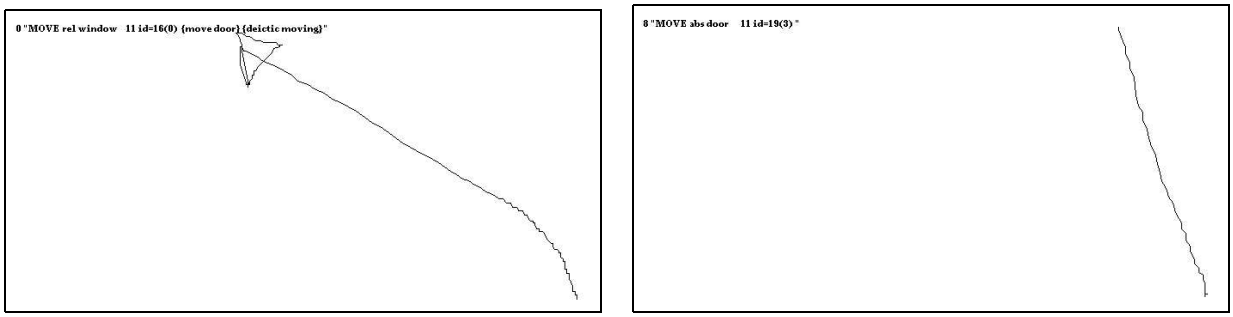



**Figure 6.2:** A "Complete arrow" and an "Arrow without head"

**- 02.** "One tapping":

The subject produces one tapping gesture. The location may be close to the object to move, or close to its target position. In order for the system to understand which of these two options apply, the pen input should be accompanied by speech input. Such a "tapping gesture" can contain a few samples (like a dot), can be a small circle, or can a small cross mark.

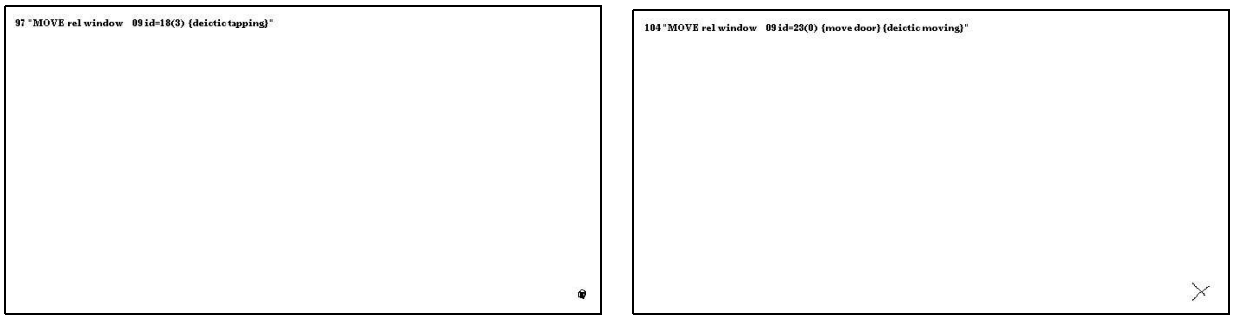



**Figure 6.3:** Two examples of “one tapping”, one with a dot and the other with a cross mark.

**- 03.** “Two tappings”:

The subject produces two tapping gestures, one close to the object to move, and the second one close to its new position, or vice versa. In order to determine which of both locations is source or destination, additional information as may be present in the speech signal is required.

**- 04.** “Redraw”:

This fourth category of user input was not expected when we implemented the system. In such cases, the subject redraws the object (s)he has to move at its new location. Below, three examples of this behaviour are depicted.

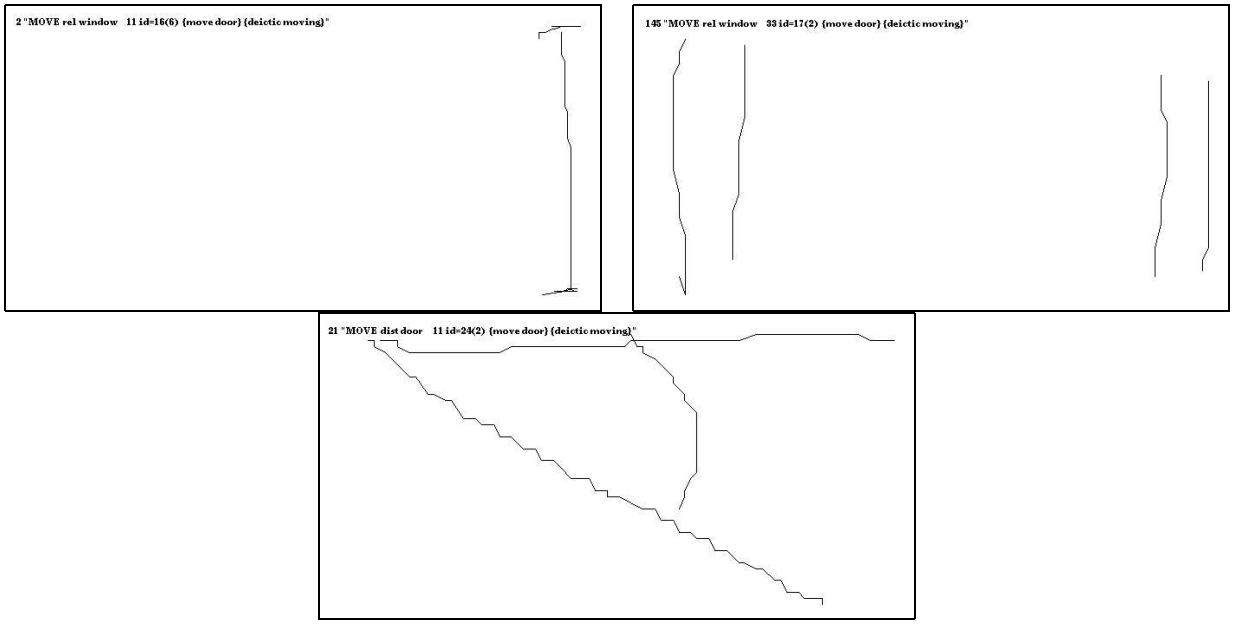



**Figure 6.4:** Three cases in which the subjects redrew a door at the destination.

**- 05.** “Arrow+Redraw”:

Here, the subject not only draws an arrow to indicate source and destination, but also redraws the object s(he) has to move at its new location.

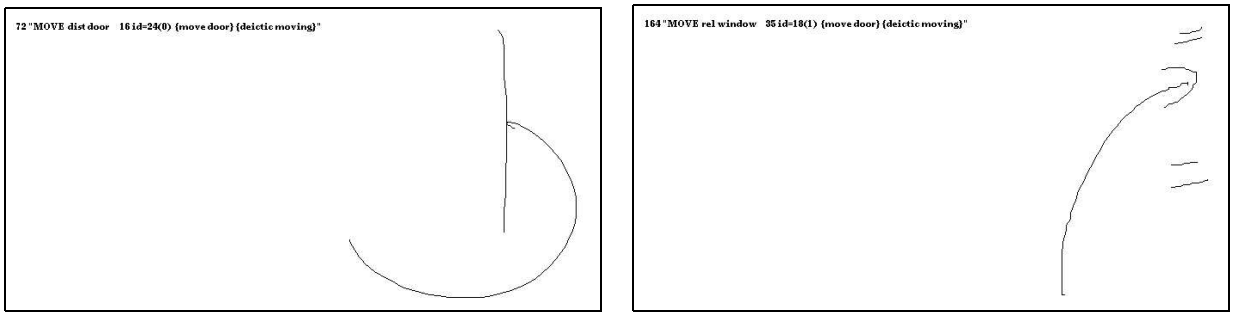



**Figure 6.5:** “Arrow+Redraw”

**- 06.** “Tapping+Redraw”:

The subject produces a tapping gesture close to the object to move, and redraws the object (s)he has to move in its new location.

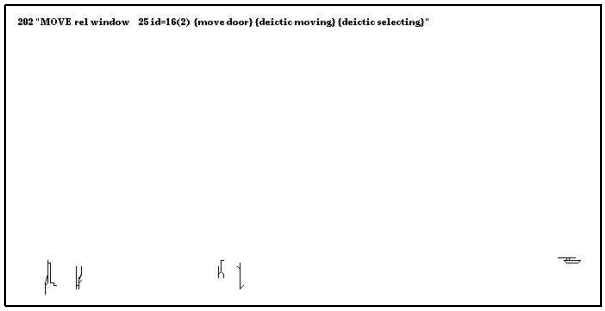



**Figure 6.6:** “Tapping+Redraw”

**- 07.** “Cross marks+Redraw”:

The subject produces cross marks, trying to erase the object to be moved. In the example depicted below, the two edges of the window are crossed out; and the window is redrawn in its new location.

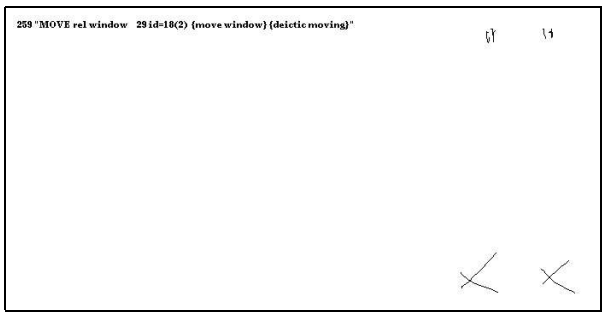



**Figure 6.7:** “Two cross marks+Redraw”

**- 08.** “Scribble+Redraw”:

The subject produces a scribble gesture on the object to move; and redraws the object (s)he has to move in its new location. The scribble was intended to erase or cross the source object out. As with all examples of redrawing, such scribbling gestures were not expected in the design of the pen input recognition module.

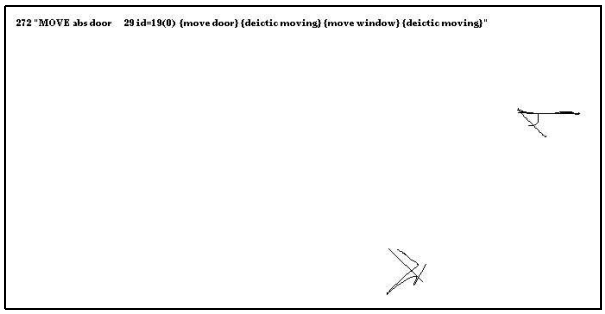



**Figure 6.8:** “Scribbling+Redraw”

**- 9.** “Encircle”:

The subject encircles the object (s)he wants to move. In order to further disambiguate the intention of the user, accompanying speech or additional pen input is required.

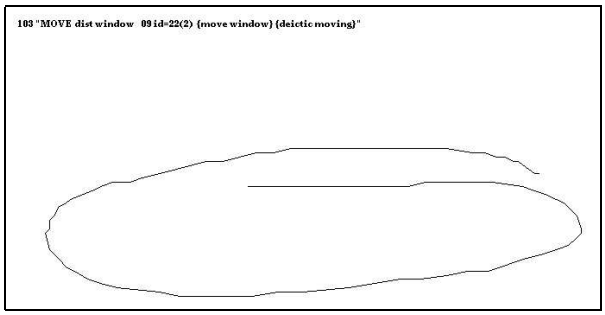



**Figure 6.9:** “Encircle”

**- 10.** “Encircle+Tapping”:

The subject encircles the object (s)he wants to move and produces a tapping gesture close to its new position.

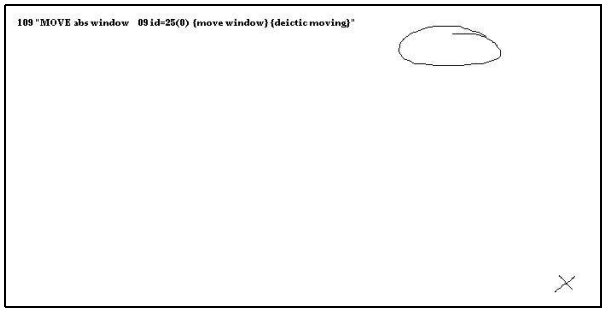



**Figure 6.10:** “Encircle+Tapping”

**- 11.** “Encircle+Arrow”:

The subject encircles the object (s)he wants to move and draws an arrow from the object to move to its new position.

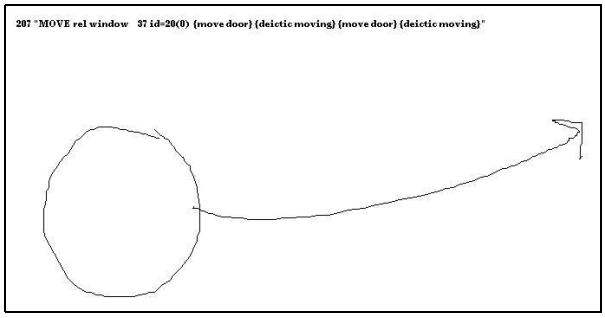



**Figure 6.11:** “Encircle+Arrow”

**- 12.** “Encircle+Redraw”:

The subject encircles the object (s)he wants to move and redraws it in its new position.

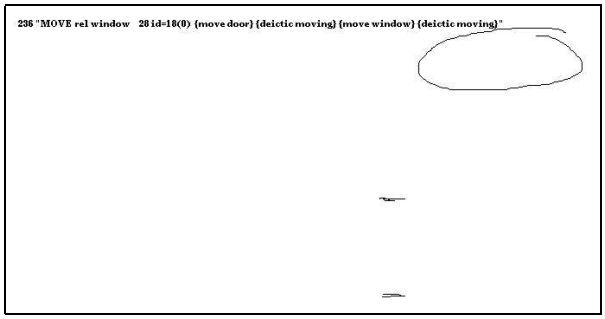



**Figure 6.12:** “Encircle+Redraw”

**- 13.** “Encircle+Arrow+Redraw”:

The subject encircles the object (s)he wants to move, draws an arrow from the object to move to its new position, and redraws the object (s)he has to move in its new location.

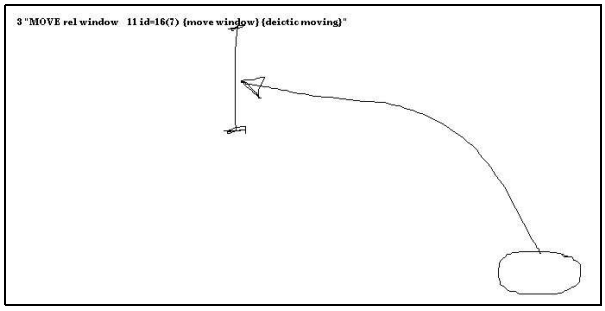



**Figure 6.13:** “Encircle+Arrow+Redraw”

**- 14.** “Multiple arrows”:

The subject encircles the object (s)he wants to move, and afterwards draws a few arrows in order to indicate how the object should be moved towards its new location.

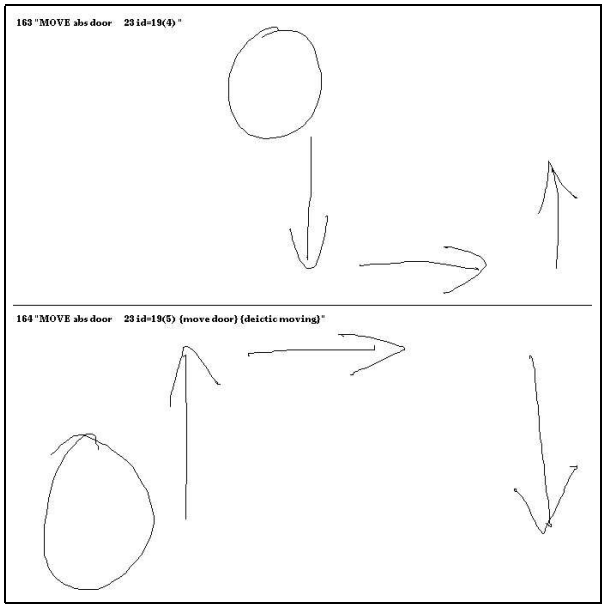



**Figure 6.14:** “Multiple arrows”

**- 15.** “Erasing”:

The subject tries to erase the object to move. This was is not supported by the current system.

### Summary of moving gestures

In total, 546 “moving gestures” have been collected during the experiment. Table 6.3 below, presents the frequency of occurrence of the 15 categories described above:

| class | frequency | class | frequency | class | frequency |
|---|---|---|---|---|---|
| 01 arrows | 286 (52.4%) | 06 tapping+redraw | 4 (0.75%) | 11 encircle+arrow | 30 (5.5%) |
| 02 one tapping | 36 (6.6%) | 07 cross marks + redraw | 8 (1.5%) | 12 encircle+redraw | 14 (2.5%) |
| 03 two tappings | 9 (1.6%) | 08 scribble + redraw | 2 (0.37%) | 13 encircle+arrow+redraw | 12 (2.2%) |
| 04 redraw | 96 (17.6%) | 09 encircle | 13 (2.3%) | 14 multiple arrows | 6 (1.1%) |
| 05 arrow+redraw | 8 (1.5%) | 10 encircle+tapping | 4 (0.75%) | 15 erasing | 18 (3.3%) |

**Table 6.3:** Frequency of occurrence of each kind of moving gesture.

Given the taxonomies presented in this section, it can be concluded that similar to the observations made in previous human factors reports, the variability in pen input is high. However, please note that more than 52 % of the moving gestures are arrows, which indicates that our efforts must be concentrated on the recognition of such arrows. To this end, the acquired data in this experiment as well as some databases we have available in our Department will be used for training and testing the required recognition technologies.

### 6.2.4 Deictic erasing Gestures

Compared to the previous human factors experiments, we expected a more prominent part of pen input to contain deictic gestures. Indeed, as the previous section on moving gestures shows, many encircling gestures (see, e.g., Table 6.3), arrows, and tapping gestures were collected. Note that these deictic gestures can be considered as atomic items, where multiple deictic gestures build up a moving gesture comprising compound gesture information.

Figure 6.15 depicts an overview over the typical classes of deictic erasing gestures observed in the collected loggings.

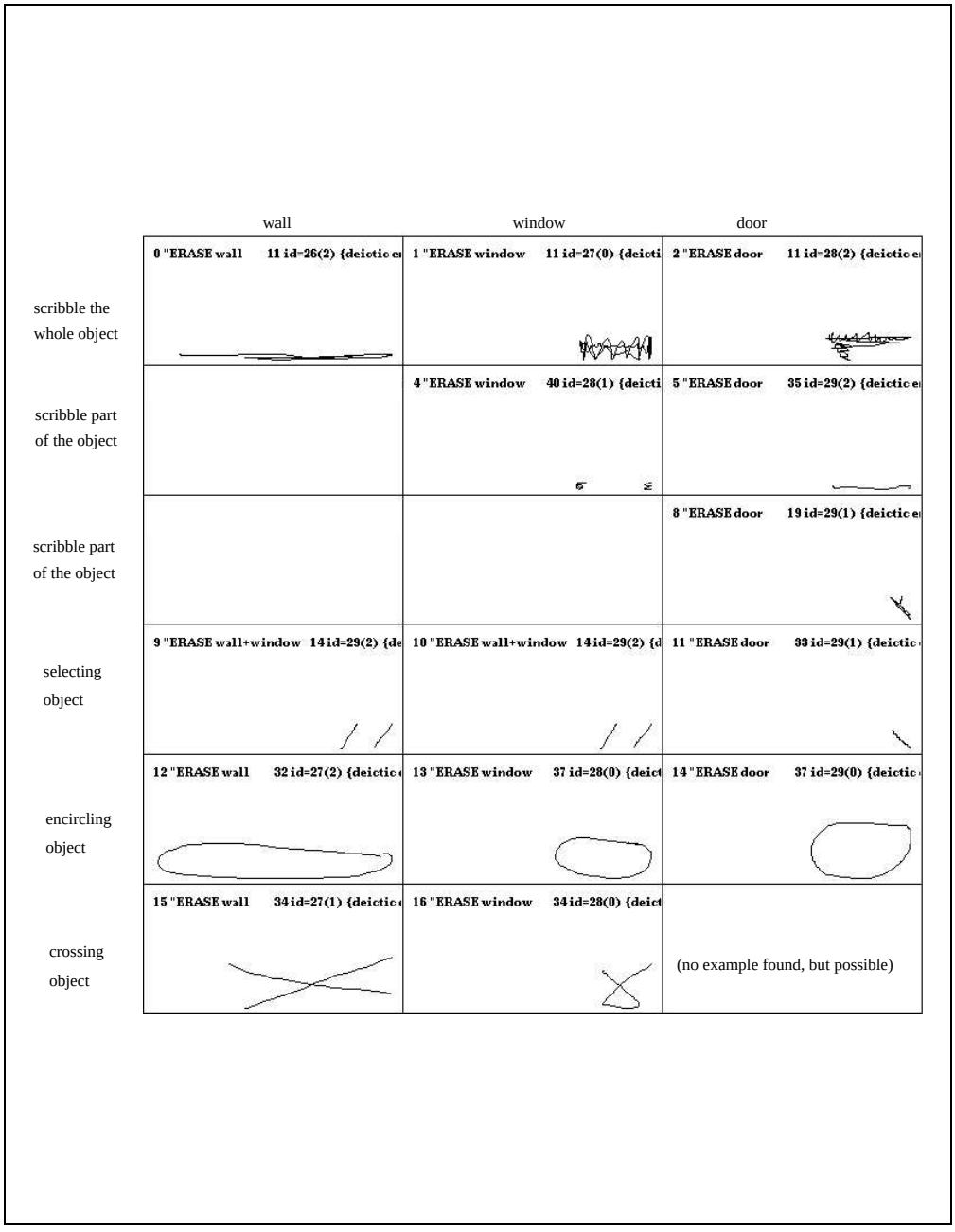



**Figure 6.15:** How subjects erase objects using the erasing feature of the pen.

A qualitative analysis of the recorded pen input was performed to categorize deictic gestures into the following three categories (i) tapping, (ii) encircling, (iii) erasing. In none of the cases, arrows were used

- Tapping. A tapping gesture is composed of a small number of samples and covers a small surface. Examples are small dots, small circles, and cross marks.
- Encircling. Encircling gestures are produced in two cases. When the user wants to move an object, then (s)he encircles the object to move. And when the user wants to erase an object, then (s)he encircles the object to erase.
- Erasing. It is also possible to erase an object using the back of the pen (or the button on its side). In that case, the subjects encircle the object they have to erase, or produce a scribble gesture, crossing over the object to become erased. The objects that were asked to be erased are walls, doors and windows. The scribble gesture produced for each of these objects is different and depends on the way objects are beautified:
  - The beautification of a wall consists of a vertical or an horizontal line. Such scribbles follow the orientation of the wall in serrate, perpendicular strokes.
  - The beautification of a window consists of four parallel short segments, two for each edges. Sometimes, subjects explicitly erase the two edges and nothing else. In some other occasions, they erase the window and the portion of the wall comprised between its two edges. This can be observed in Figure 6.15.
  - The beautification of a door consists, for an horizontal one, of an horizontal segment representing its position; and of a diagonal segment representing its its opening. Sometimes, the subjects explicitly erase these two segments. Sometimes, they erase only the diagonal one. This can be observed in Figure 6.15.

## 6.3 Natural Language Processing

The NLP module analyzes the spoken utterances resulting in an abstract representation of the utterances. In this section, we examine which kinds of utterances occurred in the human factor experiments and how well they were processed.

### 6.3.1 Phenomena

The utterances collected in the T28 human factor experiments will be used to enhance the knowledge bases of the NLP module for the T30 and T36 demonstrators. The utterances are categorized in utterances creating objects (incl. walls, doors and windows), utterances specifying sizes (incl. lengths, widths and heights) and utterances which move or erase objects. The following subsections present examples of user utterances and show the implications on the further development of the NLP module and the system-wide used ontology.

#### Creation of walls, windows and doors

We expected that the subjects only draw the objects. But 6 subjects tried to enter objects using speech, e.g., by uttering *door on the wall to the right*, *put the window on the left wall* or *enter the left wall*. This even happened in one case when an additional length was requested, e.g, *missing wall two point four three meters long*. Currently, such utterances are not processed and it is unlikely that such functionality will be added to the system. But with little effort it would be possible to hint the user that the objects have to be drawn.

One subject counted the objects loudly (*one two three*) while drawing multiple walls. The problem is that they are currently interpreted as sizes. Extending the NLP knowledge bases in a way that such utterances are ignored would avoid that.

#### Specification of sizes

The subjects either uttered the raw sizes, e.g.,*one point five meters* or they added some contextual information, like in *the width of the window is seventy centimeters* or *the top wall is two meters and forty nine centimeters*. While in most cases the provided information can be inferred from the DAM's expectations and is not necessary to generate the correct interpretation, it is nevertheless important that the utterance is processed completely. Otherwise, the unprocessed parts of the utterance would lead to a low score and another hypothesis of the speech recognizer might be selected by NLP.

If the user utters several sizes and the corresponding objects already exist, the additional information becomes important. This is the case in stimuli 14 which asks the user to specify lengths for two already existing walls. Typically, the user referred to the walls using relative locations, e.g., *the top wall is three meters the left wall is three meters* or *three meters on the left and three meters above*. Currently the ontology does not support the referring of walls using relative locations and an extension of the ontology is essential to support such utterances.

#### Move operations

Most subjects just used the phrase presented on the screen, e.g., *move the window below the door*. But also some variants occurred, like

- Omitting the verb, e.g., *window below the door* or even *window right door*
- Exchanging the verb, e.g., *put* instead of *move*
- Addressing the avatar/face directly, e.g., *could you please move the window to the opposite wall* or *you should move the door to the opposite wall*

#### Erase operations

Like in the case of move operations, the phrases presented on the screen were used most of the time to perform the specified task.

If only speech as modality is used and several objects of the same type are shown on the screen, the users again used relative locations, e.g., *erase the bottom wall*. The support of these utterances requires an extension of the system-wide ontology.

### 6.3.2 Performance

To measure the performance of the NLP module all user utterances were transcribed manually. Utterances containing noise or off-talk were filtered out for the evaluation. Afterwards, the transcriptions were processed by the NLP module and the result was stored.

In a next step, each utterance was categorized manually in one of four classes depending on the transcription and the result of the NLP analysis:

**Not representable** Principally, the utterance is not representable in terms of the ontology and therefore, could not be analyzed in an adequate manner. To process the utterance correctly in the future, the ontology as well as the knowledge bases of ASR and NLP have to be extended.

**Out of grammar** The utterance is representable in terms of the ontology, but is not covered by the grammars of ASR and NLP. To process the utterance correctly in the future, it is sufficient to extend the knowledge bases of ASR and NLP.

**Analyzed correctly** The utterance was correctly analyzed by the NLP module.

**Analyzed incorrectly** The utterance was not correctly analyzed by the NLP module although due to a bug in the knowledge bases of NLP. An update of the knowledge bases is required.

The combined performance of ASR and NLP is measured by comparing the output of using the output of ASR and the output using the transcriptions. This part of the evaluation considers only utterances which were classified as *analyzed correctly*. Table 6.4 and table 6.5 shows the performance for each stimulus.

The results show that an extension of the ontology is advisable. Also, the experiments give a hint how the knowledge bases have to be extended and where some fixes are required.

## 6.4 FUSION

In the following we describe what implications the results of the T28 human factors experiment for COMIC's FUSION component have. To cope with the overall structure of the document this consideration is partitioned into three subsections: (i) turn-taking protocol, (ii) compound objects and (iii) spatial relations.

In general, we used all these findings to update and enhance the current FUSION component that will be used for the T30 and T36 COMIC demonstration system. However, those modifications focused mainly on the rule base as the actual architecture of FUSION is stable and revealed no more unforeseen problems. The current rule base of FUSION for phase 1 (phase 2 and phase 3 were not part of this experiment) comprises 39 active rules[1]. During the experiment, the production rule system underlying FUSION went in total 18230 times through the so-called *fetch execute circle* which means that 18230 times a rule fired.

### 6.4.1 Turn-Taking

An important result of the T24 evaluation was that the T24 turn-taking protocol turned out to be too restrictive. Many subjects for example did not know when they were supposed to speak and were often cut off by the closing channels. These obstacles can be traced back to the fixed time-out window (about 8 seconds) within which the user was able to communicate with the system. The subjects also criticized in the T24 evaluation the speed or pace of the interaction. Event though most of the time is spent on recognition and generation of multi-modal utterances, the turn-taking protocol itself sometimes caused additional delay. If a user for example entered the requested information fairly quickly through a single channel, FUSION had to wait for the other modality to finish its time-out before it could pass on the recognition result to DAM.

The new turn-taking protocol permits a less restricted interaction style as the fixed recognition window is replaced by a dynamic extending recognition window managed by FUSION. This enables users to start their contribution whenever they want— after the system has passed over the floor. Moreover, users can virtually input an unlimited number of objects or commands within one turn without being interrupted by the system[2].

The most crucial aspect of this new turn-taking protocol is the identification of the end of the user turn—the so-called *end-of-turn* detection. This end-of-turn detection needs to be reliable and fast in order to achieve a natural and efficient exchange of turns. Our approach (as described in Chapter 3) monitors all activity on the input channels and waits until no more pending recognition results are in the pipeline. To ensure that no ongoing input event will be disregarded, the end-of-turn detection waits one second after the last input was received before any FUSION output is generated.

[1] Active rules refers to those rules that were actually used to process the interaction patterns performed by the subjects.

[2] For a detailed description of the new turn-taking protocol see chapter 3

| # | not rep. | oog | incorrect | correct | ASR (n=1) | ASR (n=2) | ASR (n=3) |
|---|---|---|---|---|---|---|---|
| *1 - Please, enter the TOP wall* | | | | | | | |
| 1 | 1 (100%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| *2 - Please, enter the length of the TOP wall* | | | | | | | |
| 26 | 3 (11.5%) | 1 (3.8%) | 2 (7.7%) | 20 (76.9%) | 9 (45.0%) | 9 (45.0%) | 9 (45.0%) |
| *3 - Please, enter a door in one of the walls* | | | | | | | |
| 9 | 7 (77.8%) | 0 (0.0%) | 0 (0.0%) | 2 (22.2%) | 2 (100%) | 2 (100%) | 2 (100%) |
| *4 - Please, enter a window in one of the walls* | | | | | | | |
| 5 | 5 (100%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| *5 - Please, enter the LEFT wall* | | | | | | | |
| 5 | 5 (100%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| *6 - Please, enter the length of the RIGHT wall* | | | | | | | |
| 27 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 27 (100%) | 14 (51.9%) | 14 (51.9%) | 14 (51.9%) |
| *7 - Please, specify the height of the window* | | | | | | | |
| 32 | 0 (0.0%) | 4 (12.5%) | 1 (3.1%) | 27 (84.4%) | 14 (51.9%) | 15 (55.6%) | 15 (55.6%) |
| *8 - Please, enter the width of the window* | | | | | | | |
| 40 | 0 (0.0%) | 6 (15.0%) | 0 (0.0%) | 34 (85.0%) | 14 (41.2%) | 14 (41.2%) | 14 (41.2%) |
| *9 - Please, enter the TOP wall and specify its length* | | | | | | | |
| 19 | 0 (0.0%) | 3 (15.8%) | 0 (0.0%) | 16 (84.2%) | 12 (75.0%) | 12 (75.0%) | 12 (75.0%) |
| *10 - Please, enter the RIGHT wall and specify its length* | | | | | | | |
| 39 | 6 (15.4%) | 0 (0.0%) | 2 (5.1%) | 31 (79.5%) | 16 (51.6%) | 16 (51.6%) | 17 (54.8%) |
| *11 - Please, enter the RIGHT wall and complete all wall length specifications* | | | | | | | |
| 39 | 13 (33.3%) | 0 (0.0%) | 0 (0.0%) | 26 (66.7%) | 17 (65.4%) | 17 (65.4%) | 17 (65.4%) |
| *12 - Please, enter the vertical walls and specify all missing wall lengths* | | | | | | | |
| 37 | 10 (27.0%) | 0 (0.0%) | 1 (2.7%) | 26 (70.3%) | 14 (53.8%) | 14 (53.8%) | 14 (53.8%) |
| *13 - Please, complete the outline of the bathroom and specify all wall lengths* | | | | | | | |
| 41 | 9 (22.0%) | 0 (0.0%) | 6 (14.6%) | 26 (63.4%) | 12 (46.2%) | 12 (46.2%) | 13 (50.0%) |
| *14 - Please, specify the remaining two lengths* | | | | | | | |
| 24 | 6 (25.0%) | 1 (4.2%) | 3 (12.5%) | 14 (58.3%) | 13 (92.9%) | 13 (92.9%) | 13 (92.9%) |
| *15 - Please, enter the remaining three walls* | | | | | | | |
| 3 | 0 (0.0%) | 1 (33.3%) | 0 (0.0%) | 2 (66.7%) | 1 (50.0%) | 1 (50.0%) | 1 (50.0%) |

**Table 6.4:** Performance of the NLP component for stimuli 1-15. The first column contains the amount of utterances. The second column contains the amount of utterances that cannot be represented in terms of the ontology. The third column contains the amount of utterances that can be represented but is not covered by the grammars used in ASR and NLP. The forth and fifth columns contain the amount of utterances which are processed incorrectly resp. correctly by NLP using the annotations instead of the ASR output. The last three columns contain the amount of correctly processed utterances using the ASR with varied lengths of the n-best list. For the last three columns only the utterances are considered which are correctly analyzed by NLP using the transcriptions.

Within the 2839 turns entered by the subjects, we noticed only 64 breakdowns of the turn-taking protocol which means that only 2.25% of the turns were to some extend disturbed. A closer look at those disruptions reveals that there are actually two classes of problems: (i) *late delivery* of a FUSION result (more than two seconds delay) and (ii) *premature delivery* of a FUSION result (not all unimodal input events were delivered to FUSION when FUSION itself produced its output). Whereas the first class simply decelerates the generation of system feedback, the second class is much more crucial to the progression of the dialog as user contributions are most likely cut off in such situations.

As table 6.6 reveals, the late delivery is hardly a problem for our approach; only in two turns the interaction was decelerated by FUSION. The average response time of FUSION is about 1157 milliseconds including the hard-wired one-second wait for additional input events. This means that FUSION needs an average of 157

| # | not rep. | oog | incorrect | correct | ASR (n=1) | ASR (n=2) | ASR (n=3) |
|---|---|---|---|---|---|---|---|
| *16 - Please, enter the complete outline of the bathroom* | | | | | | | |
| 0 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| *17 - Please, move the window to the left of the door* | | | | | | | |
| 25 | 2 (8.0%) | 4 (16.0%) | 0 (0.0%) | 19 (76.0%) | 6 (31.6%) | 6 (31.6%) | 6 (31.6%) |
| *18 - Please, move the window below the door* | | | | | | | |
| 25 | 1 (4.0%) | 5 (20.0%) | 0 (0.0%) | 19 (76.0%) | 7 (36.8%) | 8 (42.1%) | 8 (42.1%) |
| *19 - Please, move the window above the door* | | | | | | | |
| 17 | 3 (17.6%) | 2 (11.8%) | 0 (0.0%) | 12 (70.6%) | 5 (41.7%) | 5 (41.7%) | 5 (41.7%) |
| *20 - Please, move the door to the opposite wall* | | | | | | | |
| 26 | 25 (96.2%) | 0 (0.0%) | 0 (0.0%) | 1 (3.8%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| *21 - Please, move the window to the right of the door* | | | | | | | |
| 21 | 1 (4.8%) | 3 (14.3%) | 2 (9.5%) | 15 (71.4%) | 11 (73.3%) | 10 (66.7%) | 10 (66.7%) |
| *22 - Please, move the window to the left wall* | | | | | | | |
| 35 | 33 (94.3%) | 0 (0.0%) | 0 (0.0%) | 2 (5.7%) | 1 (50.0%) | 1 (50.0%) | 1 (50.0%) |
| *23 - Please, move the window some distance to the right* | | | | | | | |
| 23 | 0 (0.0%) | 7 (30.4%) | 0 (0.0%) | 16 (69.6%) | 8 (50.0%) | 8 (50.0%) | 8 (50.0%) |
| *24 - Please, move the window above the door* | | | | | | | |
| 20 | 1 (5.0%) | 3 (15.0%) | 0 (0.0%) | 16 (80.0%) | 7 (43.8%) | 7 (43.8%) | 7 (43.8%) |
| *25 - Please, move the door to the right* | | | | | | | |
| 29 | 0 (0.0%) | 10 (34.5%) | 0 (0.0%) | 19 (65.5%) | 2 (10.5%) | 2 (10.5%) | 2 (10.5%) |
| *26 - Please, move the window to the opposite wall* | | | | | | | |
| 28 | 27 (96.4%) | 0 (0.0%) | 0 (0.0%) | 1 (3.6%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| *27 - Please, erase the BOTTOM wall* | | | | | | | |
| 39 | 28 (71.8%) | 0 (0.0%) | 0 (0.0%) | 11 (28.2%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| *28 - Please, erase the window* | | | | | | | |
| 17 | 0 (0.0%) | 3 (17.6%) | 0 (0.0%) | 14 (82.4%) | 5 (35.7%) | 5 (35.7%) | 5 (35.7%) |
| *29 - Please, erase the door* | | | | | | | |
| 18 | 0 (0.0%) | 0 (0.0%) | 13 (72.2%) | 5 (27.8%) | 2 (40.0%) | 1 (20.0%) | 1 (20.0%) |
| *30 - Please, erase the window and the wall to which it is attached* | | | | | | | |
| 31 | 21 (67.7%) | 0 (0.0%) | 0 (0.0%) | 10 (32.3%) | 3 (30.0%) | 3 (30.0%) | 3 (30.0%) |
| *all stimuli* | | | | | | | |
| 701 | 207 (29.5%) | 53 (7.6%) | 30 (4.3%) | 411 (58.6%) | 195 (47.4%) | 195 (47.4%) | 197 (47.9%) |

**Table 6.5:** Performance of the NLP component for stimuli 16-30. The first column contains the amount of utterances. The second column contains the amount of utterances that cannot be represented in terms of the ontology. The third column contains the amount of utterances that can be represented but is not covered by the grammars used in ASR and NLP. The forth and fifth columns contain the amount of utterances which are processed incorrectly resp. correctly by NLP using the annotations instead of the ASR output. The last three columns contain the amount of correctly processed utterances using the ASR with varied lengths of the n-best list. For the last three columns only the utterances are considered which are correctly analyzed by NLP using the transcriptions.

milliseconds to produce a result if the end of a turn was identified. Premature delivery, however, occurred in 62 turns (2.18%). An in-depth analysis of the loggings recorded during the experiments revealed that there are basically two reasons for this premature delivery:

- suppressed *input received* messages
  In certain circumstances it happened that FUSION received one or more input events which were not announced by a corresponding input received message. In this case FUSION was faced with additional input after it had published an integration result and responded by publishing a second result for the same turn.

| turn-taking error | total quantity | percentage of turns |
|---|---|---|
| late output | 2 | 0.07% |
| premature output | 62 | 2.18% |
| total errors | 64 | 2.25% |

**Table 6.6:** End-of-turn detection problems viewed by the two error-classes

- *complex input*
  67% of the turn-taking problems occurred only for a restricted set of five stimuli. Those stimuli are all related to the so-called compound objects where the subjects were asked to enter multi-modal commands employing multiple modes. Most of the problems were caused by one or two input events that were not considered in the eventual FUSION result.

Both problem sources are smoothed out for the upcoming T30 system by updating the rule base so that FUSION can not produce any additional output within a single turn in case it already published some output. Additionally, the rules responsible for processing complex input were updated according to the observed user behavior.

### 6.4.2 Compound Objects

The stimuli related to the compound objects can be broadly partitioned into two classes: (i) stimuli where more than one type of object was requested (e. g., a wall and its length were requested) and (ii) stimuli where only a single type of object is requested (e. g., the four walls that form a room were requested). Basically, there are two hypotheses that influenced the development of the rule base for FUSION and that we wanted to verify in this experiment:

1. If users are requested to input only a single type of object they will most likely use only a single mode.
2. The longer a turn lasts the more likely is the input pattern a complex one (e. g., involving odd numbers, repositioning of the pen).

| stimulus | description | multi-modal turns |
|---|---|---|
| multiple object types | | |
| 09 | “one wall and its length” | 52.27% |
| 10 | “one wall and its length” (more complex variation of stimulus 9) | 70.69% |
| 11 | “one wall and two lengths“ | 66.98% |
| 12 | “two walls and two lengths” | 67.85% |
| 13 | “two walls and two lengths” (variation of stimulus 12) | 61.71% |
| single object type | | |
| 14 | “two lengths” | 20.24% |
| 15 | “three walls” | 14.79% |
| 16 | “four walls” | 10.20% |

**Table 6.7:** Percentage of multi-modal input partitioned for the two conditions *multiple object types* and *single object type*

Table 6.7 reveals clearly that true multi-modal contributions comprising more than one input mode are more likely to occur when the user tries to input two or more different types of objects. The average percentage of multi-modal *multi-mode* input is for the single typed objects about 15.07% and for multiple type objects 63.90%.

Table 6.8 shows the average turn duration for the different compound stimuli. Interestingly, there is already an effect between stimulus 9 and 10 whose only difference is a slightly more complex length of the opposite wall for stimulus 10 (“3m” compared to “2.43m”). For more complex stimuli this difference is even more

prominent (see for example stimulus 11). If we consider the single object types, this effect is also visible as stimulus 14—asking for two lengths—causes longer turns than the other two stimuli.

| multiple object types | | single object type | |
|---|---|---|---|
| stimulus | average turn duration | stimulus | average turn duration |
| 9 | 5.79 sec | | |
| 10 | 7.18 sec | 14 | 7.55 sec |
| 11 | 10.67 sec | 15 | 5.30 sec |
| 12 | 9.49 sec | 16 | 5.78 sec |
| 13 | 8.78 sec | | |

**Table 6.8:** Average turn duration partitioned for the two conditions *multiple object types* and *single object type*

### 6.4.3 Spatial Relations

Interestingly, it looks as if it was not difficult for the subjects to figure out how they can move objects with a pen and speech based interface. In total, the 10 different moving stimuli were presented 749 times which means that every subject accepted the system result after 1.87 attempts. Noteworthy is also that not a single subject tried to use handwriting when faced with these stimuli.

The data obtained from the spatial relation stimuli shows several interesting phenomena (see table 6.9). In general, the amount of multi-modal contributions is rather small (only about 5 % - 15 % contributions were multi-modal ones for moving stimuli). But it is not clear whether this is related to the performance of the ASR or directly reflects the preferences of the subjects. However, a clear observation the experimenters made is that once a subject found out that it is possible to move an object simply by drawing an arrow the subjects hardly ever tried to use spoken commands again.

What we also noticed is a dependency of the target location and use of speech. When there is no clear reference point for the target location (e. g., stimulus 23) people tend to make more use of the spoken commands (in this case 40% compared to the average speech/pen ration of 29.84%). This might be an important insight for the further development of the COMIC demonstration system as the anticipated end-user will be naïve persons that walk into a store and want to get a rough impression of what a new bathroom interior would look like. These people will most likely not bring an exact blueprint of their bathroom with them and thus the system will often be faced with rather vague commands as in stimulus 23.

| stimuli | description | speech ratio | multi-modal input |
|---|---|---|---|
| 17 | “Move the window to the left of the door” | 27.72% | 8.85% |
| 18 | “Move the window below the door” | 32% | 14.75% |
| 19 | “Move the window above the door” | 23.88% | 6.25% |
| 20 | “Move the door to the opposite wall” | 30.02% | 7.48% |
| 21 | “Move the window to the right of the door” | 29.89% | 9.52% |
| 22 | “Move the window to the left wall” | 28% | 10.42% |
| 23 | “Move the window some distance to the right” | 40% | 8.96% |
| 24 | “Move the window above the door” | 24.91% | 13.16% |
| 25 | “Move the door to the right” | 32.39% | 4.94% |
| 26 | “Move the window to the opposite door” | 29.62% | 13.13% |

**Table 6.9:** Speech / pen ration and percentage of multi-modal input for the *moving* stimuli

## 6.5 Questionnaire

To investigate the subjective experience, subjects were asked to complete a questionnaire (see appendix B). It is composed of a mix of 5 more personal questions, of 30 5-points Likert-scale questions, and of 5 open questions. In this section we briefly cover the most important results of the closed questions.

The results are summarized below. For the Likert-scale questions, when the mean is higher than 3.5, it is written in bold; and when the mean is lower than 3, it is written in italic. None of the 40 subjects who participated in the experiment was native English speaker. Three of them were left-handed; one was ambidextrous. They were between 20 and 45 years old (mean: 28.525). Most subjects reported to have quite some computer experience (question 2.1, mean: 3.875), but they reported to have very little experience with speech recognition systems (question: 2.3, mean: 1.9) and even less with pen recognition systems (question: 2.2, mean: 1.625), and in general with computer programming (question 2.4, mean: 2.65).

| | 1.1 | 1.2 | 1.3 | 1.5 | 2.1 | 2.2 | 2.3 | 2.4 | 3.1 | 3.2 | 3.3 | 3.4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | 0.500 | 28.525 | 0.950 | 1.000 | **3.875** | *1.625* | *1.900* | *2.650* | **3.500** | **3.725** | 3.125 | **3.850** |

| | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 3.10 | 3.11 | 3.12 | 3.13 | 3.14 | 3.15 | 3.16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | *2.725* | 3.100 | *2.200* | *2.850* | *2.700* | 3.211 | 3.100 | 3.050 | 3.000 | *2.175* | *2.525* | **3.875** |

| | 3.17 | 3.18 | 3.19 | 3.20 | 3.21 | 3.22 | 3.23 | 3.24 | 3.25 | 3.26 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | *2.000* | **3.625** | **3.575** | 3.025 | **3.925** | *2.700* | **3.525** | **4.050** | 3.425 | 3.375 | | |

**Task understanding.** Subjects did not have difficulties to understand the task (question 3.4, mean: 3.85) and to understand the meaning of the beautifications (question 3.21, mean: 3.925). They always knew what the system expected from them (question 3.2, mean: 3.725; question 3.23, mean: 3.525). The system was not confusing (question 3.7, mean: 2.2).

**Usability of system.** The subjects did not find that the system was too fast (question 3.8, mean: 2.85; question 3.14, mean: 2.175) for them.

**Usability of input modalities.** Subjects found the pen easy to use (question 3.16, mean: 3.875; and question 3.18, mean: 3.625). In general, they did not find the system complicated to use (question 3.17, mean: 2.0; question 3.22, mean: 2.7; question 3.15, mean: 2.525).

**System performance.** Concerning the system performance, subjects think it needs to be improved (question 3.24, mean: 4.05). They didn't find it very efficient (question 3.5, mean: 2.725), nor reliable (question 3.22, mean: 2.7). However, they like to use the system (question 3.19, mean: 3.575), and they did not find it frustrating (question 3.9, mean: 2.7). And they have been able to use it successfully (question 3.1, mean: 3.5)

**Usability of the multi-modal system.** Finally, they like a lot the idea of being able to draw and to speak at the same time to the system (question 3.25, mean: 4.05)!!!

Further, three kinds of analyzes have been performed. First, a correlational analysis between the questions. Second, the results obtained for the T16 Human-Factors experiment (document [7]) and the T28 Human-Factors experiment are compared. Third, a between conditions analysis between subjects that performed the experiments in the sequential and random conditions.

### 6.5.1 Correlation between questions

We performed a Factor Analysis on the data to investigate whether there was an underlying structure in the Likert scales. Nine factors with an eigenvalue $> 1$ were obtained. However, except for a general factor, with relatively high loadings of many scales, and which explained 25% of the total variance, all remaining factors had substantial loadings of a single scale only. Thus, it appears that the scales do not sample clearly distinct dimensions of user satisfaction, at least not for the present data set. We repeated the Factor Analysis for the two conditions, with essentially the same results: 8 or 9 factors, of which only the first has high loadings from many scales.

### 6.5.2 Comparison with the T16 questionnaire

The T16 questionnaire and the T28 questionnaire are very similar. Therefore, it can be of some interest to compare the results provided in the document [7] and the results obtained now, even if it must be realized that many differences in the scores may be due to differences in the tasks. Specifically, the tasks in the present experiment showed a much higher variation than the tasks in the T16 experiment.

| | T16 expe. mean | T28 expe. mean | T16 expe. std | T28 expe. std |
|---|---|---|---|---|
| 2.1 "Experience with computer." | 3.8 | 3.875 | 0.676 | 0.911 |
| 2.2 "Experience with tablet." | 1.533 | 1.625 | 0.916 | 1.005 |
| 2.3 "Experience with ASR." | 2.133 | 1.9 | 1.356 | 1.194 |
| 2.4 "Experience with programming." | 1.667 | 2.65 | 1.175 | 1.562 |
| 3.2 "The task was obvious." | 3.5 | 3.725 | 1.225 | 0.933 |
| 3.3 "The system was easy to use." | 3.0 | 3.125 | 1.254 | 0.939 |
| 3.4 "The instructions were clear." | 4.133 | 3.85 | 0.916 | 0.70 |
| 3.5 [T16: 3.6] "The system was efficient." | 2.467 | 2.725 | 0.916 | 0.987 |
| 3.6 [T16: 3.7] "It took long to enter all data." | 3.866 | 3.1 | 1.126 | 1.105 |
| 3.7 [T16: 3.8] "The system was confusing." | 2.4 | 2.2 | 1.352 | 0.911 |
| 3.9 "The system was frustrating." | 3.267 | 2.7 | 1.28 | 1.265 |
| 3.10 "The prompts were easy to understand." | 4.533 | 3.211 | 0.64 | 1.756 |
| 3.11 "To combine pen and speech was easy." | 3.533 | 3.1 | 1.060 | 1.081 |
| 3.12 "System under control." | 2.4 | 3.05 | 0.910 | 0.932 |
| 3.13 "The combination pen/speech was natural." | 3.267 | 3.0 | 1.033 | 0.934 |
| 3.14 "The system was too fast." | 2.533 | 2.175 | 1.457 | 0.958 |
| 3.15 "It took a lot of effort to use the system." | 2.667 | 2.525 | 1.397 | 0.987 |
| 3.16 "The handling of the pen was easy." | 3.933 | 3.875 | 0.799 | 0.822 |
| 3.17 "The system was complicated." | 2.2 | 2.0 | 1.014 | 0.641 |
| 3.18 "The rubber was easy to use." | 4.067 | 3.625 | 1.033 | 1.213 |
| 3.19 "It was fun to use the system." | 3.13 | 3.575 | 1.126 | 1.059 |
| 3.20 "It took long before reaction was possible." | 3.133 | 3.025 | 1.060 | 1.165 |
| 3.21 "What was recognized by the system was clear." | 4.267 | 3.925 | 0.884 | 1.095 |
| 3.22 "The system was reliable." | 2.133 | 2.7 | 0.743 | 0.853 |
| 3.23 "If input not recognized well, what to do was clear." | 3.933 | 3.525 | 1.28 | 1.261 |
| 3.24 "The system needs to be improved." | 4.333 | 4.050 | 0.617 | 0.846 |
| 3.25 "Go idea to be able to draw and speak together." | 4.267 | 3.425 | 0.884 | 1.059 |
| 3.26 "Later stimuli easier to enter than the first ones." | 3.714 | 3.375 | 1.267 | 1.055 |

**Table 6.10:** Comparison of the means and of the standard deviations

The subjects who performed the T28 experiment found the system more efficient (question 3.5, man: 2.725) than the subjects who performed the T16 experiment (question 3.6, mean: 2.467).

The new turn-taking protocol and the fact that entering compound objects was possible, in combination with the more varied tasks affect the impression of the speed with which the system responds (question 3.6: compare 3.866 to 3.1). However, because of the differences between the tasks we cannot be certain that this difference is mainly due to the changes made to the system, relative to T16.

The subjects who used the T16 system found it quite frustrating (question 3.9, mean: 3.267), while the subjects in the T28 experiment were much less negative in this respect (mean: 2.7). They felt also to have much more the system in control (from 2.4 for the T16 system to 3.05 for the T28 one); and that the system was more reliable (from 2.133 for the T16 system to 2.7 for the T28 one).

The T28 system (question 3.19, mean: 3.575) was more fun to use than the T16 one (mean: 3.13), although especially here the more varied tasks may be more important than the changes made to the system per se.

### 6.5.3 Learning effects

In order to examine whether subjects can learn from the interaction, one half of the subjects received the stimuli blocks of similar rasks, while the other subjects got the stimuli in a random order that was different for all subjects. Since the stimuli were sampled from four categories, it was expected that subjects in the group who received te stimuli blockwise would be able to learn how to handle the stimuli belonging to the same category. Furthermore, since in particular the first category of stimuli contained relatively simple requests for atomic information, the hypothisis is put forward that the “blocked” group of subjects perform better than the “random” group, because the former has a better opprotunity to learn.

A detailed analysis was performed of the average number of turns the subject groups needed to enter the requested information. Figure 6.16 shows the average number of turns needed to successfully enter the individual stimuli. On average, the total number of trials in the ”blocked” condition was lower than in the ”random” condition.

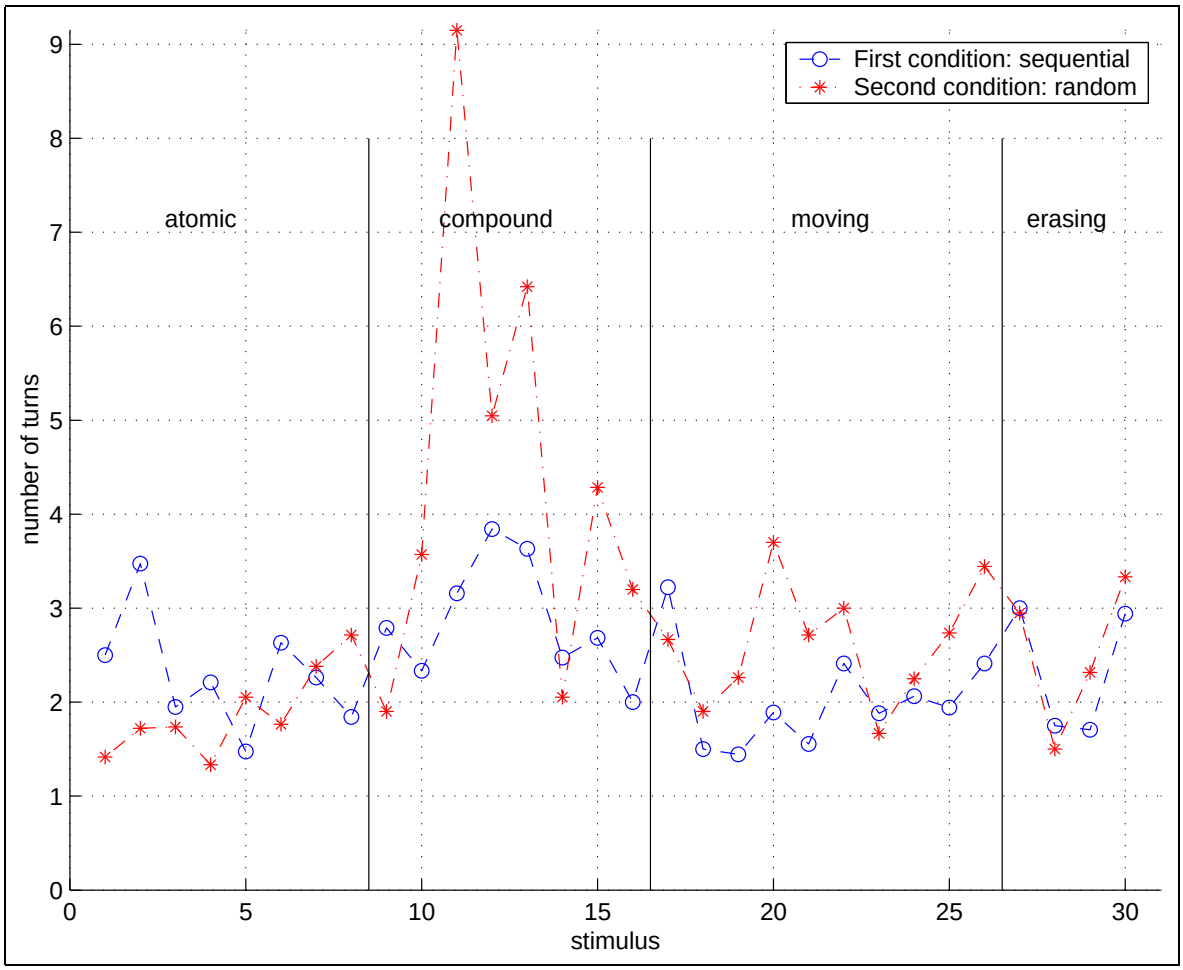



**Figure 6.16:** Average number of turns in the sequential and random conditions.

The ratio between the number of stimuli and the number of responses (user turns) indicates the difficulty of each condition:

| blocked condition | | | random condition | | |
|---|---|---|---|---|---|
| nstimuli | nturns | ratio | nstimuli | nturns | ratio |
| 530 | 1279 | 2.41 | 621 | 1750 | 2.81 |

However, the advantage of the ”blocked” condition is not present for all four categories of stimuli. Yet, the group that got the stimuli blocked shows a slight tendency for the average number of turns to increase more sharply at the border between the blocks than is the case for the ”random” group. This suggests that a so called “short term learning effect” exists, due to the fact that all the “atomic” stimuli, all the “compound” stimuli, etc. are grouped together.

## Atomic information

First, the number of trials needed by the subjects in the ”random” condition is lower than the number of turns requested in the first condition. This is probably due to the fact that the subjects in this condition already had some training when they attempted to complete tasks 1 – 3, while the subjects in the ”blocked” condition had had no chance to learn.

Second, in the ”blocked” condition, the number of turns tends to decrease with the number of the stimulus. This indicates a short term learning effect.

## Compound information

The most difficult stimuli are the 12th and the 13th ones (4 atomic objects are asked: 2 walls and 2 lengths). The “difficulty” of these particular stimuli is also reflected in the recognition performance of PII (see Table 6.2), as the system recognizes correctly only 30.7% of the cases. Stimuli 9, 10 and 14 are similar: 2 atomic objects are asked. For the stimulus 11, 3 atomic objects are asked. Stimuli 15 and 16 are apart, and less difficult than the previous ones, as only walls are asked, and no length. For both subject groups this gradation in the difficulty of the stimuli is obvious in Fig. 6.16. For the time being, we have no explanation for the very large differences between the two groups for the most difficult stimuli.

## Spatial relations

The number of turns needed in the “random” condition is higher. This suggests that for this block of stimuli there was a short term learning effect, which is obviously absent in the “random” condition.

## Erasing

There is no noticeable difference between the two conditions. This is due to the fact that, as the questionnaire shows it, the subjects did not find difficult, in both condition, the use of the rubber.

### Subjective measures

We conducted *t*-tests to compare the Likert scale scores between the two groups of subjects. No significant differences were found, except for question 3.19 (It was fun to use the system.) for which the average score of the “blocked” condition (3.7) was significantly higher $p = 0.015)$ than for the “random” condition (3.4). Apparently, the subjects who got the stimuli in random order found the task somewhat less pleasant.

# Chapter 7

# Summary and conclusions

This report describes the results from the T28 experiments performed in WP3 and WP4 in COMIC. These experiments have been a success in many aspects. First of all, the experiments have tested the recommendations for technology development derived from the Human Factors experiments at T16 and the experiments with the T24 system. Following these guidelines, two major changes in the input recognition, NLP and FUSION modules were made. First, a new turn-taking protocol was implemented that made it possible to provide variable, user-driven time slots during which user input was allowed. Second, the system prompts and the task of the subjects were adapted in order to elicit much more complex information, christened as *compound* objects. Moreover, we also provided the possibility to erase multiple objects in one turn. Significant improvements in the technology modules were made to be able to handle compound objects.

In addition, the reviewers requested that we spend attention to conversational handling of spatial relations. This resulted in a third major adaptation of the tasks and the supporting technology: users were provided with the option to modify situations on the screen by supporting absolute and relative specifications of move operations.

To be able to perform the T28 experiments we implemented a new platform, that makes it easy to design stimuli and tasks, as long as these stay within the limitations of one system request with a single – be it potentially very complex – response of the subject.

As a spin off of the T28 experiments, COMIC has acquired a substantial data set comprising multi-modal information. Compared to the previous human factors experiments, the amount of truly multi-modal utterances had increased significantly. This collection will form the basis of future research within COMIC and beyond.

## 7.1 Interactive experiments with a working system

During the design of the T28 research platform, it became clear that in order to perform interactive experiments, users should be aware of how the system handled their input. In our case, a notion of system feedback could only be implemented if the responses generated by the FUSION module would be interpreted and properly rendered through a graphical user interface. At the same time, we were restricted by the fact that many of the issues performed in the current research were not yet implemented in the available COMIC system. In particular the notions of compound objects and spatial relations had to be developed from scratch. Therefore, the decision was made to build a dedicated user interface that emulated the Visoft front-end and that was able to parse and interpret messages generated by FUSION. Since no resources and time for implementing a full-blown dialog manager were available, we designed a system that: (i) presented the user with a particular situation (“stimulus”) on the screen; (ii) collected user input acquired during one user turn; (iii) determined automatically when the user had finished entering information; (iv) processed the collected complex multi-

modal and compound information; and (v) rendered the fused hypotheses in order to provide the user with the appropriate system feedback.

In total 30 stimuli were designed, organized in four categories: (i) requests for atomic information, (ii) requests for compound information, (iii) requests for moving operations through spatial relations and (iv) requests for erasing a single or compound objects from the screen.

The T28 research platform has been very successful. The implementation of the user interface and corresponding mechanism to easily modify stimuli through a control file has made it possible to fine tune the requests for information.

## 7.2 Input recognizers - ASR

Based on a global analysis of the T28 logged data for ASR, the following conclusion can be drawn from the experiment.

- The recognition of compound input is made possible by a careful tuning of the ASR LM in combination with the adjustment of settings in the NLP. In chapter 6, a list of representative examples has been presented that shows the various issues that the ASR has to solve. Partly because of the inevitable errors made by the ASR as a result of the additional options in the language model, the risk that subsequent modules (such as NLP) make erroneous interpretations has increased. The more complex the LM is (such as in the case of compound decoding with long turns), the more complex the interpretation of the ASR output will become.
- The robustness of the ASR is still to be improved. The recognition performance of the ASR in testing conditions is strongly dependent on the speaker. Also the penalty setting of the garbage model is still not very robust. A better tuning of the garbage model (and also the construction of more elaborate garbage models) would be possible with more annotated data.
- The improvement of the ASR crucially depends on the availability of annotated data. To that end, all loggings of the T28 experiment have undergone (or will undergo) an annotation phase.
- The complexity of the turn taking for T28/T30 is about the limit that the current architecture of HTK can handle without substantial additional programming effort.

## 7.3 Input recognizers - PII

The research questions that were set for pen input recognition in the chapters explaining compound information, the new turn taking mechanism, and multimodal specifications of spatial relations (Chapters 2,3,4), can be summarized in two main categories:

- What is the pen repertoire, how do users enter all this information? In Section 6.2, several taxonomies of pen gestures are provided. All acquired data has now been annotated and classified accordingly. To our knowledge, these elaborate explorations in pen input for design applications have never been performed in such a detail yet. Therefore, as a spin off from this study, the collected databases and obtained knowledge will provide a valuable source of information for future research on pen computing. The data is stored in the UNIPEN standard, for which several institutions (including the NICI) are currently developing software to transform UNIPEN to InkML, the standard developed in the W3C multimodal interaction group[1].
- How does the module perform, what cases are typical or problematic? Given that most of the observed pen gestures were quite new for the pen input interpreter, the focus of our analyzes on recognition

[1] see `http://www.w3.org/TR/2004/WD-InkML-20040928/`

performance was on the mode distinction and recognition of compound objects. The average recognition result of 67.8% was judged as relatively good, given that this number represents “turn recognition rates”, where a single digit or a wall that was not recognized is counted an error. Furthermore, as reported by the experimenters that supervised the experiments, the majority of subjects had no problems in using ”moving gestures”, of which a large part contained arrow-like shapes. Similarly, the majority of gestures employed to erase one or more objects, were handled well by the module as well. More detailed analysis will be presented in deliverable 3.2, which is due at T32.

With respect to pen input, the conclusion can be drawn that the T28 experiments were quite successful. A large amount of relevant data was acquired and stored for future research. The performance of the recognition algorithms was judged as satisfactory, although it is also concluded that improvements are required. Finally, an elaborate taxonomy of pen input in the context of natural interactions was presented.

## 7.4 NLP: natural language processing

The experiment has corroborated the hypothesis that the design of the NLP module enables us to extend this module so as to cover additional semantic detail with relatively little human involvement. Also, the NLP module has proved to be quite robust. At the same time it has once again become clear that it is of crucial importance that the ontology covers all relevant objects and operations.

For the remainder of the COMIC project no fundamental changes to the NLP module appear to be necessary. However, substantial adaptations within the context of the present architecture may be needed if the ontology in the COMIC system grows more complex. Additional changes may be necessary if the output of the ASR module changes, for example due to the use of an N-gram language model instead of a grammar.

## 7.5 FUSION: a new turn-taking protocol

The new turn-taking protocol was designed with two goals: (i) to provide more natural, user-driven interactions and (ii) to reduce the amount of latencies in the system. As a result of the experiments, the new protocol has been extensively tested. In total 2839 turns were encountered, of which almost 98% cases were handled correctly. The remaining 2% of the cases pointed at two classes of scenarios that were not covered by the protocol. At present, these problems have been solved.

With respect to the first goal we can conclude that the new turn taking protocol was able to support the manner in which subjects entered quite complex compound objects, that required the combination of pen and speech, as well as the combination of multiple atomic objects in the pen and/or speech channels.

With respect to the second goal, the average turn durations listed in Table 6.8 can be used to show that a more efficient interaction results from the possibility to enter multiple information items in one turn. Turn durations between 5.30 and 7.55 seconds were observed for single objects. For compound objects, durations between 5.8 and 10.7 seconds were observed. Average durations are 6.2 seconds for single objects and 8.3 seconds for compound objects, comprising at least two objects. So, it can be deduced that the average time to enter one object in the case of compound objects is less than 4.2 seconds.

## 7.6 Multi-modal information in user-driven interactions

One result from the experiments is particularly striking: the amount of multi-modal information that was observed in the current user-driven context was much larger than in the situation where the system requests for single information items. This was shown in detail in Section 6.4.2, where in the case of requests for

compound information on average 63.9% utterances contained both speech and pen, whereas for atomic objects, a mere 15.1% multimodal utterances were observed.

On the other hand, for specifying move operations through spatial relations, only between 5% up to 15% cases were observed where subjects employed both pen and speech (see Section 6.4.3). The conclusion that subjects figured out that using arrow-like gestures resulted in efficient interactions can be sustained by the observation made in Section 6.2.2 that more than 52% of the pen input comprised arrows.

## 7.7 Subjective user experiences

Although the results of the Likert scales for the present experiment are difficult to interpret, the overall picture that emerges from the scores is that the subjects were able to perform the tasks without too much difficulty and frustration. This makes us confident that we are on the right track, both with the design of the multimodal interaction and with the development of the component technologies.

## 7.8 Directions for future research

The overall conclusion that can be drawn from the experiments described in this report is that interaction design and component technology development are on the right track, but that substantial additional improvements of all component technologies are still necessary. ASR and PII are far from near 100% recognition accuracy, despite the fact that fully natural interaction will definitely require near-perfect performance. In addition, the input patterns will definitely become more complex and less predictable if users no longer have the still relatively strict guidance that they had in the T28 experiments: although users had substantial freedom in choosing the way in which to respond to the system prompts, we still had the situation that subjects entered a single (compound) object in response to a very specific request. If users face the need to enter a number of such complex information objects, in a situation where they cannot always be certain what information must be entered, nor if the pieces of information must be entered in a certain order, user behavior cannot but become more complex. If that happens, also NLP and Fusion, which in the T28 experiments showed a very high performance, will be faced with more difficult tasks, and consequently will require additional development to avoid higher proportions of errors. Future experiments with the T30 and T36 systems will show what the most important remaining problems are.

# Bibliography

[1] all COMIC partners. D1.4, specifications for the t30 demonstrator. Technical report, 2004. Available via `http://www.hcrc.ed.ac.uk/comic`.

[2] L. Boves, A. Neumann, L. Vuurpijl, L. ten Bosch, S. Rossignol, R. Engel, and N. Pfleger. Multimodal interaction in architectural design applications. In *8th ERCIM Workshop on "User Interfaces for All"*, Palais Eschenbach, Vienna, Austria, 28-29 June 2004.

[3] E. den Os and L. Boves. Towards ambient intelligence: Multimodal computers that understand our intentions. In *Proc. eChallenges*, Bologna, October 2003.

[4] S. Rossignol, A. Neumann, and *et al* L. ten Bosch. Description of the COMIC T16 Human Factor experiments. Internal COMIC Document 3.3.1, June 2003.

[5] S. Rossignol, L. ten Bosch, and L. Vuurpijl *et al*. Human-factors issues in multi-modal interaction in complex design tasks. In *HCI International*, pages 79–80, Greece, 2003.

[6] S. Rossignol, D. Willems, A. Neumann, and L. Vuurpijl. Mode detection and incremental recognition. In *Proc. of the Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR9)*, Tokyo, Japan, October 2004. In press.

[7] Stéphane Rossignol, Louis ten Bosch, and Louis Vuurpijl. Internal COMIC Document 3.1 – Pilot studies for phase-I of the Comic demonstrator. Technical report, 2002. Available via `http://www.hcrc.ed.ac.uk/comic`.

[8] L. Vuurpijl, L. ten Bosch, S. Rossignol, A. Neumann, N. Pfleger, and R. Engel. Evaluation of multimodal dialog systems. In *LREC Workshop Multimodal Corpora and Evaluation*, Lisbon, 2004.

[9] Louis Vuurpijl, Louis ten Bosch, Stéphane Rossignol, Andre Neumann, Ralf Engel, and Norbert Pfleger. Reports on human factors experiments with simultaneous coordinated speech and pen input and fusion. Technical report, 2003. Available via `http://www.hcrc.ed.ac.uk/comic`.

# Appendix A

# Instructions

Instructions file:///home/rossigno/T28-Installation/instructions/T28Instructio...

Instructions

Please read these instructions carefully. If something is unclear, do not hesitate to ask the experimenter for h
Follow the instructions as closely as possible during the experiment.

This experiment takes place within the context of the so-called COMIC-project. In this project, the NICI wo
together with six other European partners to investigate new possibilities of interacting with the computer
this case the interaction of speech, handwriting and pen gestures. This project will integrate those interac
possibilities into an already existing graphical program for the designing of bathrooms.

**About the experiment**

The experiment will take about 45 minutes. During the experiment you will wear a microphone via which
can talk to the system. Furthermore, you can use a pen to draw or write on a graphical tablet. You can use
pen also like a rubber for erasing things on the screen. To do this, press the bottom part of the button on the
of the pen, and use the pen to encircle or to point at the items you want to remove.

**Task**

How do I use the system?
The experiment is composed of 30 different sub-tasks – so-called stimuli. For each stimulus, the system
ask you to enter some piece(s) of information.

The requests from the system are written on the bottom side of the drawing surface.

When you have read the request from the system, press the button " Go!" .

The background of the drawing surface turns into light blue and you can interact with the system.

You can use the pen as well as speech (via the microphone) to enter the requested information.

When you have finished entering information, the background turns into red after a while. This me
that you cannot enter anything anymore.

The system will give you visual feedback of what it understood after each input.
You can then " Accept" or " Reject" what the system offers you on the screen.

à If you press " Accept" the system will start the next task.
à If you press " Reject" the system will ask you to re-enter your information.

What kind of information is requested?
For each stimulus, you will see parts of a bathroom blueprint. This may contain one or more walls, a window
a door. Each of these bathroom parts may be accompanied by a corresponding size, like the length of a wall
the width or height of a window. Below the stimulus, you will read a text that requests you to add, modify
delete (erase) particular information.

The system will request information like:
1. The form of the bathroom (walls and their measures). You can only enter rectangular bathrooms.
2. The position of a door (including the opening directions).
3. The position of windows, their width, the height of their lower edge and their height.
4. You may be asked to move a window or a door to another location.
5. You may be asked to erase one or more bathroom parts (window, door, wall, or sizes).

**Before you start:**

Make sure that you have read and understood these instructions;

Do not hesitate to ask the experimenter for help;

Before the experiment, you will be asked to enter some personal information, like age and gender.

Please, also read the statement belonging to the questionnaire and sign the agreement;

After the experiment, you will be asked to fill in a questionnaire. In this questionnaire, you will answer some questions about your experiences with using the system.

The COMIC team

1 of 1 10/01/2004 10:23 AM

**Figure A.1:**

# Appendix B

# Questionnaire

Fragebogen Human Factors Experiment file:///home/rossigno/T28-Installation/instructions/T28Questionn...

Questionnaire P-number:

**Agreement**
You are about to perform an experiment within the framework of the European COMIC project. The information that you enter via speech and/or pen will be recorded. These recordings will be analysed with the goal to evaluate the performance and usability of the bathroom design system. By signing the agreement below, you state that you have no objections to the fact that the COMIC team will analyse and use these data.

" I know that audio- and pen-recordings are made from this experiment. I give my approval to the COMIC-project to use these recordings for analyzing purposes. The recordings will not be put at the disposal of third persons outside of the project. The recordings can, however, be used for scientific presentations and demonstrations. "

Name: _____

Nijmegen, date: _____

Signature: _____

**Part 1. Personal information**

1.1 Gender: male / female
1.2 Age: _____
1.3 Handedness: left / right / both
1.4 Education: (please pick one) High school
University (enter study) _____
other _____

1.5 Are you a native English speaker? Yes / No

Part 2. Computer experience

Mark the number that indicates best your experience with:

2.1 computers in general
very little 1 2 3 4 5 very much

2.2 graphical tablets with pen (e.g. Palm-PDA, TabletPC)
very little 1 2 3 4 5 very much

2.3 automatic speech recognition systems (e.g. train information)
very little 1 2 3 4 5 very much

2.4 computer programming
very little 1 2 3 4 5 very much

## Part 3

In the following, you see some statements about the system you just used.
Please evaluate your experience on a scale from 1 (totally disagree) to 5 (totally agree). We invite you to take your time to answer these questions and if you would like to add a comment or remark to please do so!

| | Totally Disagree | Disagree | Neutral | Agree | T |
|---|---|---|---|---|---|
| 3.1) I was able to use the system successfully | 1 | 2 | 3 | 4 | |
| Remarks: | | | | | |
| 3.2) I knew what I was supposed to do at every step | 1 | 2 | 3 | 4 | |
| Remarks: | | | | | |
| 3.3) It was easy to use the system | 1 | 2 | 3 | 4 | |
| Remarks: | | | | | |
| 3.4) The instructions made it clear how to use pen and speech | 1 | 2 | 3 | 4 | |
| Remarks: | | | | | |
| 3.5) I found the system efficient | 1 | 2 | 3 | 4 | |
| Remarks: | | | | | |
| 3.6) I took long to enter all information | 1 | 2 | 3 | 4 | |
| Remarks: | | | | | |
| 3.7) I found the system confusing | 1 | 2 | 3 | 4 | |
| Remarks: | | | | | |
| 3.8) It was difficult for me to enter information quickly | 1 | 2 | 3 | 4 | |
| Remarks: | | | | | |
| 3.9) I found the use of the system frustrating | 1 | 2 | 3 | 4 | |
| Remarks: | | | | | |
| 3.10) I always understood the questions and hints from the system | 1 | 2 | 3 | 4 | |
| Remarks: | | | | | |



**Figure B.2:**

| | Totally Disagree | Disagree | Neutral | Agree |
|---|---|---|---|---|
| 3.11) To combine pen and speech was easy | 1 | 2 | 3 | 4 |
| Remarks: | | | | |
| 3.12) I felt in control when using the system | 1 | 2 | 3 | 4 |
| Remarks: | | | | |
| 3.13) The combination of pen and speech felt natural | 1 | 2 | 3 | 4 |
| Remarks: | | | | |
| 3.14) The system was too fast for me | 1 | 2 | 3 | 4 |
| Remarks: | | | | |
| 3.15) It took me a lot of effort to use the system | 1 | 2 | 3 | 4 |
| Remarks: | | | | |
| 3.16) The handling of the pen was easy | 1 | 2 | 3 | 4 |
| Remarks: | | | | |
| 3.17) I found the system complicated | 1 | 2 | 3 | 4 |
| Remarks: | | | | |
| 3.18) I found the use of the rubber easy | 1 | 2 | 3 | 4 |
| Remarks: | | | | |
| 3.19) It was fun to use the system | 1 | 2 | 3 | 4 |
| Remarks: | | | | |
| 3.20) It usually took a long time before I could make another entry | 1 | 2 | 3 | 4 |
| Remarks: | | | | |



**Figure B.3:**

| | Totally Disagree | Disagree | Neutral | Agree | T |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| 3.21) The way the system showed what it had recognized was obvious<br>Remarks: | | | | | |
| 3.22) I experienced the system as reliable<br>Remarks: | 1 | 2 | 3 | 4 | |
| 3.23) If an input was not recognized well I knew what I had to do<br>Remarks: | 1 | 2 | 3 | 4 | |
| 3.24) In my opinion, the system still has to be improved considerably<br>Remarks: | 1 | 2 | 3 | 4 | |
| 3.25) It was a good thing to be able to draw and speak at the same time<br>Remarks: | 1 | 2 | 3 | 4 | |
| 3.26) The later stimuli were easier to enter than the first ones<br>Remarks: | 1 | 2 | 3 | 4 | |



**Figure B.4:**

**Part 4:**
To complete this questionnaire, we kindly request that you answer the following questions:

| 4.1 | What was the hardest part about interacting with the system? |
|---|---|
| | |

| 4.2 | What was the easiest part? |
|---|---|
| | |

| 4.3 | Did the system behave unexpectedly? If so, how? |
|---|---|
| | |

| 4.4 | Do you think you should have had some (more) practice before the experiment started? |
|---|---|
| | |

| 4.5 | Please tell us why you would like to use the system again (or why not). |
|---|---|
| | |



**Figure B.5:**