

**INFORMATION SOCIETIES TECHNOLOGY (IST) PROGRAMME**



**HOMEY**

**“Home Monitoring through an Intelligent Dialog System”**

**DELIVERABLE: D14 (*external*)**

**WORKPACKAGE: WP8**

**Validation**

**D14 – Validation Protocol**

**Author:** Martin Beveridge (Cancer Research UK)  
Toni Giorgino (Consorzio Bioingegneria e Informatica Medica)

**Submission Date: 31/12/2003**

**Partners: Engineering Ingegneria Informatica, (I), Reitek (I), Consorzio Bioingegneria e Informatica Medica (I), Istituto Trentino di Cultura (I), Cancer Research United Kingdom (UK), Language & Computing (B)**

## **SUMMARY**

This document is part of the result of the research project HOMEY funded by the IST Programme within the 5<sup>th</sup> Framework Programme as project number IST-2001-32434.

One of the goals of the HOMEY project is to develop a technology to be used for deploying innovative tele-medicine services. These new services will be based on an Intelligent Dialog System (IDS), designed and developed to effectively manage an incremental dialog between a tele-medicine system and a patient, taking into account user needs, preferences and time course of her/his disease.

The purpose of work package 8 is to evaluate accuracy and efficiency of the Intelligent Dialog System in four different show cases: 1) advice for urgent referral of suspected cancer patients, 2) advice about “family tree” related diseases, 3) hypertensive patients, 4) type 2 diabetes mellitus patients. In all showcases a prototype will be run in order to assess performance and to derive potential useful indicators for improvement and exploitation of the technology.

The purpose of this deliverable is to set out the evaluation protocol to be used to investigate linkage between decision support in healthcare and speech dialogues. This protocol will allow for the evaluation of system performance (acoustic accuracy, semantic accuracy, etc) for all showcases and will test the usability of speech based chronic diseases and cancer applications. Formal technical investigations will be obtained by assessing the use of semantic “knowledge sources” for constraining speech interpretation and for dialogue control.

# CONTENTS

<b>1. Abstract</b> .....	<b>4</b>
1.1. PURPOSE OF THE HOMEY PROJECT .....	4
1.2. PURPOSE OF WORK PACKAGE 8.....	4
1.3. PURPOSE OF THIS DELIVERABLE .....	4
1.4. AUTHORSHIP OF THIS DOCUMENT .....	4
1.5. LIST OF ABBREVIATIONS .....	5
<b>2. Introduction</b> .....	<b>6</b>
<b>3. Background</b> .....	<b>7</b>
3.1. EVALUATION MEASURES .....	7
3.1.1. <i>Speech Recognition Performance</i> .....	7
3.1.2. <i>Dialogue Manager Performance</i> .....	7
3.1.3. <i>Dialogue Manager Competence</i> .....	9
3.2. EVALUATION METHODOLOGY.....	10
3.2.1. <i>Standard Scripts</i> .....	10
3.2.2. <i>Unscripted Interaction</i> .....	10
3.2.3. <i>Live Setting</i> .....	11
<b>4. Validation of the Cancer Demonstrators</b> .....	<b>12</b>
4.1. APPLICATIONS.....	12
4.1.1. <i>Breast Cancer Referrals</i> .....	12
4.1.2. <i>Genetic Risk Assessment</i> .....	12
4.2. EVALUATION.....	12
4.2.1. <i>Variables</i> .....	12
4.2.2. <i>Measures</i> .....	16
4.2.3. <i>Protocol</i> .....	17
<b>5. Validation of the Chronic Disease Demonstrators</b> .....	<b>19</b>
5.1. STUDY DESIGN.....	19
<b>References</b> .....	<b>21</b>

# **1. Abstract**

## **1.1. Purpose of the HOMEY project**

The purpose of the HOMEY project is to carry out research and develop technology to be used for deploying innovative tele-medicine services. The new services will be based on an Intelligent Dialogue System (IDS), designed and developed to effectively manage an incremental dialogue between a tele-medicine system and a patient, taking into account user needs, preferences and the time course of her/his disease. Intelligent dialogue requires the representation of goals, intentions, and beliefs about the effectiveness of the interaction in terms of quality of health care management. The dialogue system will require dynamic adaptation in order to understand the patient's medical problems and the physician's goals, handle misunderstandings, and carry-out argumentation regarding therapy options. In order to support such adaptation, a representation of the medical domain knowledge, the evolution of the disease of a specific patient, and the history of user-system interactions need to be represented.

## **1.2. Purpose of Work Package 8**

The purpose of this work package is to evaluate accuracy and efficiency of the Intelligent Dialog System in four different show cases: 1) advice for urgent referral of suspected cancer patients, 2) advice about "family tree" related diseases, 3) data collection from hypertensive patients, and 4) type 2 diabetes mellitus patients. In all the showcases a prototype will be run in order to assess performance and to derive potential useful indicators for improvement and exploitation of the technology. The first two showcases will allow evaluation of the possible roles of the *PROforma* knowledge model and the domain ontology in the speech interpretation process. The last two applications target the task of collecting data from patients over the phone.

## **1.3. Purpose of this Deliverable**

The purpose of this deliverable is to set out the evaluation protocol to be used to investigate linkage between decision support in healthcare and speech dialogues, in particular for the show cases related to chronic diseases and cancer referral and genetic risk assessment. This protocol will allow for the evaluation of system performance (acoustic accuracy, semantic accuracy, etc) for all showcases and will test the usability of speech based chronic diseases and cancer applications. Formal technical investigations will be obtained by assessing the use of semantic "knowledge sources" (e.g. the *PROforma* knowledge model or domain ontology) for constraining speech interpretation and by comparing and contrasting the use of either *PROforma* or voiceXML as the scripting language for dialogue control.

## **1.4. Authorship of this Document**

Responsibility for authorship is divided as follows. Sections 1, 2, 3 and 4 were written by Martin Beveridge with suggestions and advice from David Milward and John Fox (CRUK). Section 5 was written by Toni Giorgino (CBIM).

## 1.5. List of Abbreviations

---

<i>CRUK</i>	Cancer Research UK	Partner responsible for WP6 & WP8
<i>L&amp;C</i>	Language & Computing n.v.	Partner responsible for WP5
<i>CBIM</i>	Consorzio di Bioingegneria e Informatica Medica	Partner responsible for WP4

---

## 2. Introduction

This deliverable sets out the evaluation measures and methodologies that can be employed to investigate linkage between decision support in healthcare and speech dialogues for the show cases related to:

1. Breast cancer referral
2. Breast cancer genetic risk assessment
3. Hypertension,
4. Type 2 diabetes

For all four showcases, an evaluation of system performance (acoustic accuracy, semantic accuracy, etc), competence (range of dialogue phenomena handled), and usability can be made using standard measures as described in Section 3.

In the first two cases the system to be evaluated is a research prototype, as described in Deliverable D11 (Beveridge and Milward, 2003a), which is intended to investigate the use of existing knowledge representation schemas for medicine as the basis for dialogue management. These dialogue showcases are therefore built upon pre-existing clinical applications, as described in Section 4, and so the proposed evaluation is not one of clinical efficacy. Instead a formal technical investigation will be carried-out to assess the use of semantic “knowledge sources” – the CR-UK *PROforma* knowledge model (Fox et al., 2003) and L&C ontology (Ceusters et al., 2001) – for constraining speech interpretation, and as an alternative to hand-crafting dialogues in a scripting language such as voiceXML.

In the last two cases the system is a mature dialogue-based clinical application, as described in Deliverable D6 (Stefanelli et al., 2002), which has recently undergone internal evaluation and refinement (see Deliverable D7; Stefanelli et al., 2003). For these showcases, the main objective of the evaluation study is therefore to determine the medical effectiveness of the developed systems in clinical trials. The protocol for validation of these chronic disease demonstrators is described in Section 5.

## 3. Background

### 3.1. Evaluation Measures

The following measures can be employed for system evaluation: speech recognition performance, dialogue manager performance and dialogue manager competence.

#### 3.1.1. Speech Recognition Performance

There are four measures that are typically used: word accuracy, string recognition, concept accuracy and semantic recognition.

##### *Word Accuracy*

A commonly used measure of speech recognition performance is the accuracy of the system in recognising individual words. This is typically calculated using the formula below:

$$WA = 100 \left( 1 - \frac{W_s + W_i + W_d}{W} \right) \%$$

This measures accuracy in terms of the number of word substitutions ( $W_s$ ), deletions ( $W_d$ ) and insertions ( $W_i$ ) relative to the total number of words ( $W$ ) in the actual spoken utterances. Here *substitution* means that a different word was recognised from the one spoken, *deletion* means that a word was spoken but not recognised, and *insertion* means that a word was recognised even though it wasn't spoken.

##### *Sentence Recognition*

This measures the percentage of sentence strings that were completely correctly recognised (i.e. where every word in the sentence was correctly recognised).

##### *Concept Accuracy*

Another useful measure is the accuracy of the system in acquiring concepts (i.e. degree of semantic understanding). Based the standard measure of word accuracy given above, Boros et al. (1996) propose the following formula to calculate concept accuracy:

$$CA = 100 \left( 1 - \frac{SU_s + SU_i + SU_d}{SU} \right) \%$$

This measures accuracy in terms of the number of substitutions ( $SU_s$ ), insertions ( $SU_i$ ) and deletions ( $SU_d$ ) of semantic units relative to the total number of semantic units uttered ( $SU$ ).

##### *Semantic Recognition*

This measures the percentage of completely correctly understood sentences (i.e. where every concept in the input utterance was correctly acquired) (cf. Semantic Error Rate; Knight et al., 2001).

#### 3.1.2. Dialogue Manager Performance

This is typically measured by: degree of success in achieving the desired task, cost of successful completion, and the overall usability of the system. Different evaluation schemes,

e.g. SUNDIAL<sup>1</sup> (Simpson and Fraser, 1993), PARADISE<sup>2</sup> (Walker et al., 1997), Behavioural Coding Scheme (Sutton et al., 1995), address different aspects of these measures, as described in detail in Deliverable D7 (Stefanelli et al., 2003).

### ***Task Success***

Aspects of task success include:

1. The number of users who managed to complete a dialogue
2. Correctness/appropriateness of data provided by the system (e.g. SUNDIAL 'Transaction Success' metric)
3. Correctness of data acquired from user, i.e. some measure of the match between filled fields & values and the actual dialogue (e.g. PARADISE Kappa coefficient)

The SUNDIAL Transaction Success metric (Simpson and Fraser, 1993) is described in Deliverable D7. The PARADISE Kappa coefficient (Walker et al., 1997) is defined as follows:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where  $P(A)$  reflects the proportion of times that the correct fields were filled with correct values and  $P(E)$  reflects the proportion of times that this would be expected to happen by chance. Hence, when the agreement between the acquired fields & values and the actual dialogue is no better than would be expected by chance then  $\kappa = 0$  and when there is complete agreement then  $\kappa = 1$ . Because this metric simply compares the matrix of acquired fields & values with the dialogue, it is independent of the actual tasks being performed by a dialogue system (and so can be used to compare different systems).

### ***Dialogue Costs***

Aspects of dialogue 'cost' include:

1. System response time, i.e. the average amount of time for the system to respond to a user's utterance.
2. Amount of time required to complete a dialogue.
3. The number of turns required to complete a dialogue (e.g. SUNDIAL 'Number of Turns' metric)
4. The proportion of turns that were spent correcting errors such as misrecognitions, misunderstandings etc (e.g. SUNDIAL 'Correction Rate' metric)

The SUNDIAL 'Number of Turns' and 'Correction Rate' metrics (Simpson and Fraser, 1993) are described in Deliverable D7.

### ***Usability***

Aspects of usability include:

---

<sup>1</sup> Speech UNderstanding in DIALogue, ESPRIT contract P 2218

<sup>2</sup> PARAdigm for DIAlologue System Evaluation

1. The number of times a user made use of 'help'
2. The quality of system responses, e.g. whether the response is appropriate, inappropriate or incomprehensible (e.g. SUNDIAL 'Contextual Appropriateness' metric)
3. The quality of user responses, e.g. the degree to which user answers were responsive, usable and/or concise (e.g. Behavioural Coding Scheme)
4. User report, e.g. how the user rates their experience (friendly, competent, would you use it again etc.)

The SUNDIAL 'Contextual Appropriateness' metric (Simpson and Fraser, 1993) and the Behavioural Coding Scheme (Sutton et al., 1995) are both described in Deliverable D7.

### **3.1.3. Dialogue Manager Competence**

In addition to dialogue manager *performance*, two previous projects, TRINDI<sup>3</sup> and DISC<sup>4</sup>, have provided criteria for evaluating a dialogue manager's *competence* in handling certain dialogue phenomena. These are the TRINDI tick-list and the DISC dialogue management grid, described below.

#### ***TRINDI Tick-List***

The TRINDI Tick-List (Bohlin et al., 1999) consists of three sets of questions that are intended to elicit explanations describing the extent of a system's competence. Apart from the TRINDI project, this tick-list has also been employed in other projects, such as SIRIDUS<sup>5</sup> (Berman et al., 2000).

The first set of criteria relate to the flexibility of dialogue provided by a dialogue manager:

1. Can the system deal with answers to questions that give more information than was requested?
2. Can the system deal with answers to questions that give different information than was requested?
3. Can the system deal with answers to questions that give less information than was actually requested?
4. Can the system deal with negatively specified information?
5. Can the system deal with 'help' sub-dialogues initiated by the user?
6. Does the system deal with 'non-help' subdialogues initiated by the user?
7. Can the system deal with inconsistent information?
8. Can the system deal with belief revision?

The second set of criteria relate to overall system functionality:

1. Can the system deal with noisy input?
2. Can the system deal with barge-in input?

---

<sup>3</sup> Task Oriented Instructional Dialogue, European Telematics Applications Programme project LEA-8314.

<sup>4</sup> Esprit Long-Term Research Concerted Action No. 24823.

<sup>5</sup> Specification, Interaction and Reconfiguration In Dialogue Understanding Systems, IST-1999-10516.

3. Can the system deal with no answer to a question at all?
4. Can the system check its understanding of the user's utterance?
5. Does the system only ask appropriate follow-up questions?

The third set of criteria relate to the system's ability to make use of knowledge to provide appropriate responses:

1. Is utterance interpretation sensitive to dialogue context?
2. Can the system deal with ambiguous designators?

### ***DISC Dialogue Management Grid***

The DISC Dialogue Management grids (Heid et al., 1998) include a set of questions, similar to the Trindi tick-list above, that are intended to elicit some factual information regarding the potential of a dialogue system:

1. What initiative can the system cope with? (System/User/Mixed)
2. Free or bound order of main tasks?
3. Does the system initiate repair dialogues?
4. Does the system initiate clarification dialogues?
5. Can the user initiate repair dialogues?
6. Can the user initiate clarification dialogues?
7. Can indirect speech acts be handled?
8. Is there any difference between the system's use of speech acts and its ability to do topic spotting?
9. Does the system deal with ellipsis?

## **3.2. Evaluation Methodology**

In order to evaluate the measures described above, the following methodologies can be employed: interaction based on standard scripts, unscripted interaction and a live setting.

### **3.2.1. Standard Scripts**

In this scenario, users are provided with a standard script to follow in interacting with the system. Because all users follow the same dialogue structure, this approach makes it possible to evaluate the effect of variables such as differing voice characteristics, ambient noise, and grammar complexity on the speech recognition performance, separate from any effects of the dialogue strategy. This scenario is typically used only for internal evaluations.

### **3.2.2. Unscripted Interaction**

In this scenario, users are provided only with a set of information and goals to achieve. They then interact with the system in order to try to achieve those goals given the information available to them. This approach allows for evaluation of the dialogue manager competence and performance, as well as the speech recognition performance. It can be used in both internal and external evaluations.

### **3.2.3. Live Setting**

In this scenario, the dialogue system is accessed live by members of the target group for the system for some extended period. This scenario allows for a full evaluation of the speech recognition performance and dialogue manager competence/performance, in particular the system's usability. This scenario is necessarily an external evaluation.

## 4. Validation of the Cancer Demonstrators

### 4.1. Applications

Two cancer demonstrators will be evaluated. The first is a breast cancer referrals application, and the second is a system for determining a patient's genetic risk of developing cancer.

#### 4.1.1. Breast Cancer Referrals

In other work, CRUK has developed a system (ERA) for advising doctors on whether patients require urgent referral for suspected cancer (Bury et al., 2001). The system is currently accessed by a standard web browser that generates web pages for collecting patient data and reporting on results (see <http://www.infermed.com/era>). The demonstrator developed as part of the HOMEY project uses the knowledge representation developed for ERA, along with an ontology provided by L&C, to provide a spoken dialogue interface for entering data into this system.

#### 4.1.2. Genetic Risk Assessment

A previous project undertaken by CRUK (RAGs) developed a system to assess genetic risk in breast and ovarian cancer, and to support professional counsellors helping people make personal healthcare decisions in these areas (Emery et al., 2000). RAGs applied decision theories and technologies, and cognitive models and theories of risk perception, which had been developed in a series of earlier projects. As with the breast cancer referral demonstrator, the genetic risk assessment demonstrator developed as part of the HOMEY project attempts to re-use these pre-existing knowledge representations in order to develop a spoken dialogue risk assessment system.

### 4.2. Evaluation

As stated in Section 2, the medical effectiveness of the applications that the dialogue demonstrators are based on has already been determined in previous studies (Bury et al., 2001; Emery et al., 2000) and so the evaluation here concentrates on the spoken dialogue interfaces to these applications.

#### 4.2.1. Variables

In evaluating the cancer demonstrators the following variables must be considered: domain size, degree of flexibility allowed at any point in the dialogue, verification strategy, variation in voices, and level of ambient noise. Each of these could impact on the effectiveness of the system as described below.

##### *Domain Size and Structure*

The CRUK dialogue system extends the set of concepts specified in the dialogue domain plan with related concepts from the domain ontology, as described in Deliverable D9 (Ceusters et al., 2002). The number of concepts represented in the dialogue state therefore depends on the number of concepts referenced in the domain plan and the number of concepts in the domain ontology that are related to each concept in the domain plan. The latter in turn depends on various factors:

1. The types of relations between concepts to be considered (currently only subsumption relations)

2. The topological distance from a target concept that should be considered (currently only immediate source and target concepts, i.e. path distance of 1)
3. The difference between the information content of a target concept and related concepts, i.e. semantic distance (calculated according to the algorithm<sup>6</sup> in (Van Buggenhout and Ceusters, 2003)).

The number of concepts involved in the dialogue state in turn determines the complexity of the language model generated for the speech recogniser: the higher the number of concepts, the more terms there will be in the speech grammar.

In addition, the ontological relations between concepts in the domain ontology give rise to informational relations between dialogue segments (games) in the representation of the dialogue state, and these are used (in addition to intentional relations derived from the domain plan) to structure the dialogue, as described in Deliverable D11 (Beveridge and Milward, 2003). Hence, the degree of inter-connectedness between concepts affects the constraints placed on the dialogue manager in determining which dialogue segment should receive attention next. The degree to which the ontology size and structure affects the performance of the dialogue system largely depends on the degree of flexibility of user response that is allowed. The more flexible the range of possible responses, the more scope there is to make use of ontological relations in the dialogue planning.

### ***Degree of Flexibility of User Response***

The degree of user-initiative allowed in interaction obviously affects the complexity of the language model for the speech recogniser, and the complexity of the overall dialogue. In the CR-UK system the dialogue manager can use one of four levels of flexibility, defined by a 'tolerance' parameter specified by the dialogue client on initiating a conversation. These are outlined below:

1. *Tolerance=1*. This specifies a fully system-initiative interaction using only concepts specified in the domain plan.
2. *Tolerance=2*. This specifies a mixed-initiative interaction, but only using concepts specified in the domain plan. In this case the user can answer more than one question in one utterance but must use the exact terms expected by the domain plan.
3. *Tolerance=3*. This specifies mixed-initiative interaction using a wider range of concepts than those specified in the domain plan. In this case the user can answer more than one question with a single utterance, and the range of replies that the user can give includes references to underspecified concepts (i.e. more general concepts than those referred to by the domain plan) and overspecified concepts (i.e. more specific concepts than those referred to by the domain plan) as described in deliverable D9 (Ceusters et al., 2002).
4. *Tolerance=4*. This specifies a wider form of mixed-initiative in which the user can not only answer more than one question with a single utterance, but can also explicitly take the initiative by asking questions of the system (e.g. clarification questions). The range of replies also includes references to overspecified and underspecified concepts as before.

---

<sup>6</sup> In this algorithm the information content of a concept depends on the number of links from it to other concepts and the information content of those concepts. Hence information content is related to the degree of inter-connectivity between concepts.

Clearly, the first case will lead to the smallest and simplest language models for the speech recogniser, and also the simplest (although longest) dialogues. An example of this type of dialogue is given below.

```
[1] S: What is the patient's age?  
[2] U: 35  
[3] S: ok, what is the patient's sex?  
[4] U: female  
[5] S: ok, does the patient have an acquired nipple deformity?  
[6] U: no  
[7] S: ok, does the patient have a skin ulcer?  
[8] U: yes  
[9] S: ok, ...
```

The second case will lead to a more complex language model and allows for the dialogue itself to vary as the system responds to extra information provided by the user by re-ordering the priorities of remaining questions. An example of this type of dialogue is given below. Here, when the user responds to the system's question regarding patient age in utterance [2], they also specify some additional information: the patient's sex and the fact that they have some nipple discharge. The system responds in [3] by following-up this information with a related question before returning to its default dialogue plan in utterance [4]. In utterance [6], the user again takes the initiative and supplies extra information.

```
[1] S: What is the patient's age?  
[2] U: she is 35, female with some nipple discharge  
[3] S: ok, and is it a bloodstained nipple discharge?  
[4] U: no  
[5] S: ok, does the patient have an acquired nipple deformity?  
[6] U: no, but she has a skin ulcer  
[9] S: ok, ...
```

The third case will lead to greatly increased complexity of the language model as the range of possible terms that can be used increases. The dialogue structure will also be more complex as the system responds not only to extra information but also to overspecified and underspecified concepts. In the former case, a single concept can answer more than one question due to the ontological relations between concepts (e.g. if there is no nipple discharge then there must also be no bloodstained nipple discharge). In the latter case, the system will have to initiate clarification sequences in order to try to answer questions.

An example of this type of dialogue is given below. Here the user's reply [2] to the system's initial question includes the term 'nipple disorder', which is a more general (underspecified) term than the system needs, so it follows-up with a clarification question [3]. In [4] the user informs the system that the patient has 'bloodstained nipple discharge' which implicitly tells the system that the patient has 'nipple discharge' so it need not raise the question "does the patient have any nipple discharge" as in the previous example. This is therefore an example of an overspecified reply. In [6] the user uses the term 'cyst' which is too general (underspecified) for the system to use and so it tries to clarify the user's utterance [7].

```
[1] S: What is the patient's age?  
[2] U: she is 35, female, and she has a nipple disorder  
[3] S: ok, what type of nipple disorder does the patient have?  
[4] U: she has some bloodstained nipple discharge  
[5] S: ok, does the patient have any nipple eczema?  
[6] U: yes, and she has a cyst  
[7] S: ok, do you mean a breast cyst?  
[8] U: yes  
[9] S: ok, ...
```

The fourth case will obviously lead to the most complex language model and the most complex dialogue structure as the system responds not only to extra information and over- or under-specified replies, but also to meta-level queries regarding the dialogue (e.g. what are valid replies? Why are you asking this? etc). An example of this type of dialogue is given below. Here the user is able in [4] to issue a clarification request to the system, and the system provides an appropriate response [5].

```
[1] S: What is the patient's age?  
[2] U: she is 35, female, and she has a nipple disorder  
[3] S: ok, what type of nipple disorder does the patient have?  
[4] U: what do you mean?  
[5] S: for example: nipple anomaly or nipple discharge  
[6] U: she has some nipple discharge  
[7] S: ok, ...
```

Whilst, it would be expected that a high degree of flexibility in user responses would provide the most usable system, that flexibility also requires a complex language model and so increases the chance of misunderstandings. Another important variable is therefore the verification strategy employed in order to handle mis-recognitions etc.

### ***Verification Strategy***

Typically, verification of the system's recognition of utterances falls into one of the following categories:

1. *Explicit*: E.g., "ok, I understood that the patient is female. Is that correct?"
2. *Implicit*: E.g. "ok, the patient is female. What is her age?"
3. *None*

In the first case, the system explicitly states the information it understood and asks the user to confirm whether it was correct. This therefore "signals understanding but also a lack of confidence in the interpretation" and so tries "to elicit a response regarding the correctness of the interpretation from the hearer" (Larsson, 2003; p76). In the second case, the system simply restates the information it has understood and continues to the next question without requesting confirmation. This "signals confident understanding and ... does not try to elicit a response" (Larsson, 2003; p76). In the final case information is not confirmed at all, hence this is the most 'efficient' strategy, in terms of fastest completion of the dialogue.

Typically, the verification strategy is varied between these categories. For example, Hulstijn and van Hessen (1998) suggest that "for a crucial action, such as the closure of a transaction, information must be verified explicitly" whereas "for actions such as responses to inquiries,

where recognition errors are less costly, the efficiency gain of suppressing verification prompts, outweighs the risk of misunderstanding”.

In the case of the CRUK system, where the dialogue manager follows-up extra information provided by the user with questions regarding semantically related concepts, a misunderstanding can seriously affect the structure of the dialogue and the chances of successful completion. For example, the following fragment shows how mis-recognition of concepts in the user’s utterance leads it to ask questions that are not relevant from the point of view of the user (indicated by ‘??’ in the example). A no-verification strategy is assumed here.

```
[1] S: Does the patient have a skin ulcer?  
[2] U: she has a cutaneous ulcer  
      [system recognises: she has some pain, cyst, ulcer]  
[3] ?? S: Do you mean intractable pain?  
[4] U: no  
[5] ?? S: Do you mean a breast cyst?  
[6] U: no
```

If an explicit verification strategy were used then the irrelevant follow-up questions [3 & 5] could be avoided by confirming each concept that the system heard before continuing the dialogue.

### ***Variation in Voices***

Speaker variation may also have an impact on system performance. In particular, the following parameters may be significant

1. Accent
2. Speed
3. Prosody

### ***Level of Ambient Noise***

The level of ambient noise in a particular environment should also be considered. We can distinguish 2 types of noise:

1. *Structured*: e.g. background voices that could be accidentally recognised by the system
2. *Unstructured*: e.g. background noise from air-conditioning, etc.

In general, structured noise poses more of a problem for speech recognition than unstructured noise. In addition the quality of the acoustic signal may differ depending on the type of microphone used, in particular the use of telephone vs. microphone, and the acoustic properties of the environment, e.g. degree of reverberation (Haderlein et al., 2003).

### **4.2.2. Measures**

Section 3 described some standard measures of the competence, performance, and usability of a spoken dialogue system. In addition to these evaluations, it is also our intention to assess the use of the CR-UK PROforma process specification language (Fox et al., 2003) and L&C ontology (Ceusters et al., 2001) as an alternative approach to developing dialogue systems to the usual method of handcrafting using a scripting language such as voiceXML. Key to this assessment is an evaluation of factors such as the following for both approaches:

1. *Reconfigurability*: the degree to which the dialogue system can be re-configured for use with other domains.
2. *Scalability*: the extent to which the dialogue system can scale-up to much larger domains

### **4.2.3. Protocol**

Due to the technical nature of the domain for the Breast Cancer Referrals demonstrator, it will only be evaluated internally using scripted interactions. However, this should provide information on the effect of variables such as differing voice characteristics, ambient noise, and grammar complexity on speech recognition performance, and allow an assessment of the dialogue manager competence. The range of evaluation measures that may be appropriate for this demonstrator are given in Table 1.

The Genetic Risk Assessment demonstrator has a much more general domain (family histories) than that of the Breast Cancer Referrals system, and so it will be evaluated both internally and externally using both scripted and unscripted interactions. In addition to speech recognition performance, the internal evaluation should provide information on the effect of variables such as domain size & structure, degree of flexibility of responses, and verification strategy on dialogue manager performance. This can then be used to select an appropriate system configuration for external evaluation, in order to provide further data on performance, and, in particular, usability. It is intended that the external evaluation be carried-out by recruiting members of the general public to use the system.

Development of the Genetic Risk Assessment demonstrator will also allow us to determine the degree to which the dialogue system, which was originally built to implement the Breast Cancer Referrals system, can be re-configured for use with other domains. The range of evaluation measures that may be appropriate for this demonstrator are given in Table 1.

Furthermore, it is intended to use the dialogue system developed in this project as part of a much larger project, CREDO (Cancer Research UK, 2002), which is intended to address the full care pathway for breast cancer patients from diagnosis through treatment and then follow-up. This will provide information on how well the dialogue system can scale-up to larger domains.

**Table 1: Evaluation measures appropriate to each demonstrator.**

Evaluation Measure	Breast Cancer Referrals	Genetic Risk Assessment
<b>Speech Recognition Performance</b>		
Word Accuracy	✓	✓
Sentence Recognition	✓	✓
Concept Accuracy	✓	✓
Semantic Recognition	✓	✓
<b>Dialogue Manager Competence</b>		
TRINDI Tick-List	✓	✓
DISC Dialogue Management Grid	✓	✓
<b>Dialogue Manager Performance</b>		
<i>Task Success</i>		
Number of users to complete		✓
Correctness of data provided by system		✓
Correctness of data acquired from user		✓
<i>Dialogue Costs</i>		
System response time		✓
Dialogue completion time		✓
Number of turns		✓
Correction rate		✓
<i>Usability</i>		
Use of help		✓
Quality of system responses		✓
Quality of user responses		✓
User report		✓
<b>System Attributes</b>		
Reconfigurability		✓
Scalability		

## 5. Validation of the Chronic Disease Demonstrators

The evaluation of the chronic disease prototypes, as described in the previous deliverables, has been divided in two phases: an internal evaluation, performed collecting and analyzing simulated dialogues from volunteers, followed by a controlled clinical trial. The former evaluation process targeted usability and stability of the system. The testing phase, which has been described in previous Homey deliverables (see Deliverable D7; Stefanelli et al., 2003), is now complete. Qualitative and quantitative measures described therein showed that recognition and usability requirements for public use have been met. The prototype validation was therefore pushed to a more advanced stage, i.e. the controlled clinical trial, whose design is described in the following section.

The hypertension management system is currently being used in three major Italian hospitals. Physicians in specialist centers for the care of hypertension are enrolling their patients in an Electronic Health Records (EHR) system linked to the dialog system and using it to store the outcomes of encounters and laboratory tests.

A motivation for the adoption of the system is that its use will prove economically viable and beneficial to the collaborating parties. Clinical evidence suggests that careful monitoring and self-monitoring of patients can have beneficial impact not only on the allocation of resources in hospitals (Young et al., 2001), but also on the actual health condition of subjects (Rogers et al., 2001), and their motivation to change behaviours in favour of healthier habits (Pinto et al., 2002; Ramelson et al., 1999).

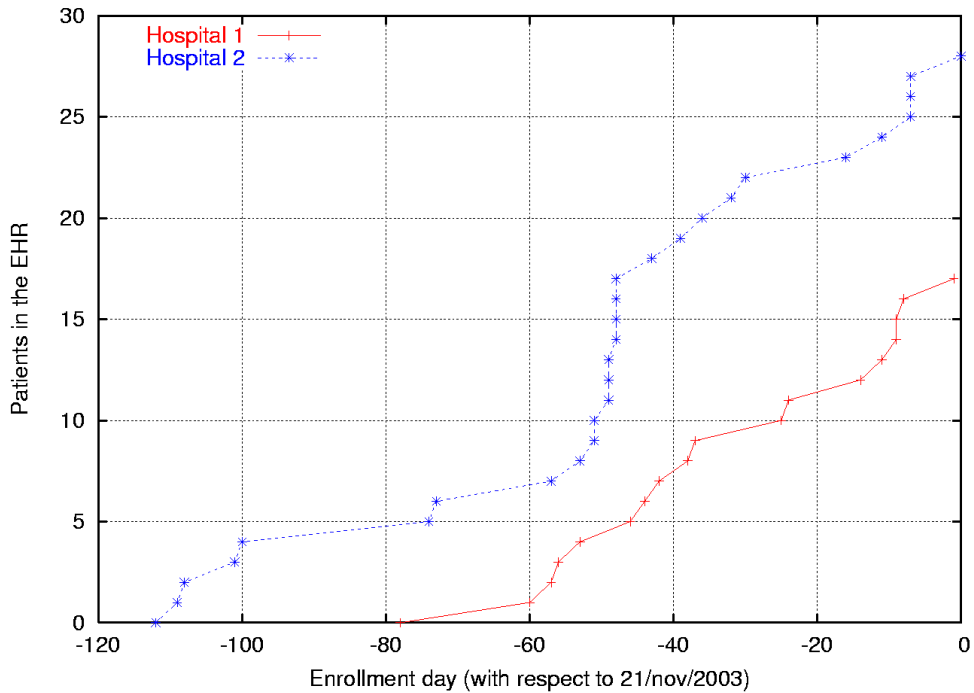
### 5.1. Study design

To verify this hypothesis, in collaboration with the caregivers we designed a randomized controlled clinical trial. Patients, after being informed of the scope of the study, are enrolled in the study and their personal data and clinical history are entered in the EHR. They are then randomly assigned to either a control group or a treatment group (population ratio 1:1, stratified by sex).

Physicians have encounters with patients of both groups with the same frequency; subjects in the treatment group, in addition to the face to face visit, are also asked to periodically dial a toll free number which connects them to the Homey service and interact with it providing the information required. The frequency of calls, which they are asked to make, is once a week for the first two months of use of the system, and once every two weeks later.

The hypothesis under test is whether systolic or diastolic blood pressure in the treatment group decreases by at least 5 mmHg with respect to the baseline of  $140 \pm 15$  over  $90 \pm 9$  (mm Hg  $\pm$  SD) within six months of study. The decrease is sought with respect to conventional treatment, i.e. additional to the decrease already attained due to pharmacological or behavioral therapy. Transient effects are in fact avoided by enrolling patients that are in a steady therapy regimen.

Study power and significance levels are taken as  $p=0.80$  and  $\alpha=0.05$ ; this yields a study size of 304 patients (152+152), distributed in equal proportion in the four collaborating medical institutions (76 patients per hospital, 38 males+38 females). Criteria for enrollment include stage 1 hypertension according to JNC7 (Chobanian et al., 2003), absence of diabetes, organic damage, and cardiovascular damage. Patients will use their own



measurements instruments, which will be calibrated in the ambulatory before the study. The study last follow-up measure, to be compared with the initial one, will be taken during an encounter.

The rate with which patients are enrolled into the database depends on physician's schedule, personal judgment, and other factors. The number of patients enrolled, as a function of time, has been checked for two hospitals, and is shown in the plot in this page. The plot displays the number of patients (grouped by hospital) entered in the EHR which have been visited at least once.

Important, albeit qualitative, user satisfaction estimators will be collected by administering surveys at the middle and the end of the study period. The clinical trial started on August 2003 and it is scheduled to last six months.

## References

- Berman, A., Cooper, R., Ericsson, S., Hieronymus, J., Jonson, R., Larsson, S., Milward, D., and Torre, D. (2000). *Implemented SIRIDUS System Architecture (Baseline)*. Deliverable D6.2, SIRIDUS Project, IST-1999-10516.
- Beveridge, M. A., and Milward, D. (2003). The High-Level Task Specification Language. Deliverable D11, HOMEY Project, IST-2001-32434.
- Bohlin, P., Bos, J., Larsson, S., Lewin, I., Matheson, C. & Milward D. (1999). *Survey of Existing Interactive Systems*. Deliverable D1.3, TRINDI Project, LE4-8314.
- Boros, M., Eckert, W., Gallwitz, F., Görz, G., Hanrieder, G., and Niemann, H. (1996). Towards Understanding Spontaneous Speech: Word Accuracy vs. Concept Accuracy. In *Proc. ICSLP'96*, Philadelphia, PA, pp. 1009-1012.
- Bury, J., Humber, M., and Fox, J. (2001). Integrating Decision Support with Electronic Referrals. In R. Rogers, R. Haux and V. Patel (Eds) *Medinfo*. IOS Press, Amsterdam.
- Cancer Research UK (2002). CREDO: A Clinical Trial of PROforma Technology in Improving Consistency, Quality and Safety in the Care of Cancer Patients. Draft Project Overview, (<http://www.acl.icnet.uk/lab/docs/credoJul02.pdf>).
- Ceusters, W., Beveridge, M. A., Milward, D., and Falavigna, D. (2002). *Specification for Semantic Dictionary Integration*, Deliverable D9, HOMEY Project, IST-2001-32434.
- Ceusters, W., Martens, P., Dhaen, C., and Terzic, B. (2001). LinkFactory: an Advanced Formal Ontology Management System. *Proc. Interactive Tools for Knowledge Capture Workshop, KCAP-2001*, Victoria B.C., Canada.
- Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L. J., Jones, D. W., Materson, B. J., Oparil, S., Wright, J. T. J., Roccella, E. J. (2003). The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 Report, *JAMA* 289 (19) pp. 2560-2572, (Guideline).
- Emery J., Walton R., Murphy M., Austoker J., Yudkin P., Chapman C., Coulson A., Glasspool D., Fox J. (2000) Computer Support for Recording and Interpreting Family Histories of Breast and Ovarian Cancer in Primary Care: Comparative Study with Simulated Cases. *British Medical Journal* 321 pp. 28-32.
- Fox, J., Beveridge, M. A., and Glasspool, D. (2003). Understanding Intelligent Agents: Analysis and Synthesis. *AI Communications*, 16 (3) pp. 139-152.
- Haderlein, T., Stemmer, G., and Nöth, E. (2003). Speech Recognition with  $\mu$ -Law Companded Features on Reverberated Signals. In *Proc. Text, Speech and Dialogue (TSD-03)*, České Budějovice, Czech Republic, 8-12 September.
- Heid, U., Bernsen, N., and Dybkjaer, L. (1998). *Current Practice in the Development and Evaluation of Spoken Language Dialogue Systems*. Deliverable D1.8, DISC Project, Esprit Long-Term Research Concerted Action No. 24823.
- Hulstijn, J., and Van Hessen, A. (1998). Utterance Generation for Transaction Dialogues. In *Proc. ICSLP'98*, Sydney.
- Knight, S., Gorrell, G., Rayner, M., Milward, D., Koeling, R. and Lewin, I. (2001). Comparing Grammar-Based and Robust Approaches to Speech Understanding: a Case Study. In *Proc. Eurospeech 2001*.
- Larsson, S. (2003). Interactive Communication Management in an Issue-Based Dialogue System. In *Proc. DiaBruck '03, 7<sup>th</sup> Workshop on the Semantics and Pragmatics of Dialogue*, 4-6 September, Saarbrücken.
- Pinto, B., Friedman, R., Marcus, B., Kelley, H., Tennstedt, S., and Gillman, M. (2002). Effects of a Computer-Based, Telephone-Counseling System on Physical Activity, *Am J Prev Med* 23 (2) pp. 113-120.

- Ramelson, H., Friedman, R., Ockene, J. (1999). An Automated Telephone-Based Smoking Cessation Education and Counseling System, *Patient Educ Couns* 36 (2) pp. 131-144.
- Rogers, M. A., Small, D., Buchan, D. A., Butch, C. A., Stewart, C. M., Krenzer, B. E., and Husovsky, H. L. (2001). Home Monitoring Service Improves Mean Arterial Pressure in Patients with Essential Hypertension. A Randomized, Controlled Trial, *Ann Intern Med* 134 (11) pp. 1024-1032 (Clinical Trial).
- Simpson, A., and Fraser (1993). Black Box and Glass Box Evaluation of the SUNDIAL System. In Proc. 3<sup>rd</sup> European Conference on Speech Communication and Technology (Eurospeech'93), Berlin, Germany.
- Stefanelli, M., Rognoni, C., Quaglino, S., Giorgino, T., and Azzini, I. (2002). Algorithms and Tools for Dialogue Adaptation. Deliverable D6, HOMEY Project, IST-2001-32434.
- Stefanelli, M., Rognoni, C., Quaglino, S., Giorgino, T., Piazza, M., and Azzini, I. (2003). Data Collection and Refinement of the Algorithms. Deliverable D7, HOMEY Project, IST-2001-32434.
- Sutton, S., Hansen, B., Lander, T., Novick, D. G., and Cole, R. (1995). Evaluating the Effectiveness of Dialogue for an Automated Spoken Questionnaire. Technical Report CS/E95-12, Dept. of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology.
- Van Buggenhout, C., and Ceusters, W. (2003). A Novel View on Information Content of Concepts in Extremely Large Ontologies, In Proc. MIE-03, IOS Press, 409-414.
- Walker, M. A., Litman, D. J., Candace, A. K., and Abella, A. (1998). Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies, *Computer Speech and Language* 12 (3).
- Young, M., Sparrow, D., Gottlieb, D., Selim, A., and Friedman, R. (2001). A Telephone-Linked Computer System for COPD Care, *Chest* 119 (5) pp. 1565-1575.