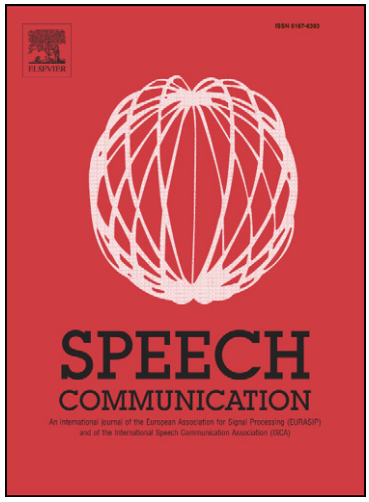
Relations between de-facto criteria in the evaluation of a spoken dialogue system

Zoraida Callejas, Ramón López-Cózar



Relations between de-facto criteria in the evaluation of a spoken dialogue system

Zoraida Callejas, Ramón López-Cózar

PII: S0167-6393(08)00052-6
DOI: 10.1016/j.specom.2008.04.004
Reference: SPECOM 1705

To appear in: *Speech Communication*

Received Date: 31 August 2007
Revised Date: 21 March 2008
Accepted Date: 5 April 2008

Please cite this article as: Callejas, Z., López-Cózar, R., Relations between de-facto criteria in the evaluation of a spoken dialogue system, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.04.004

Relations between de-facto criteria in the evaluation of a spoken dialogue system

Zoraida Callejas * Ramón López-Cózar

Dept. of Languages and Computer Systems

Faculty of Computer Science and Telecommunications

University of Granada

18071 Granada Spain

Abstract

Evaluation of spoken dialogue systems has been traditionally carried out in terms of instrumentally or expert-derived measures (usually called "objective" evaluation) and quality judgments of users who have previously interacted with the system (also called "subjective" evaluation). Different research efforts have been made to extract relationships between these evaluation criteria. In this paper we report empirical results obtained from statistical studies, which were carried out on interactions of real users with our spoken dialogue system. These studies have rarely been exploited in the literature. Our results show that they can indicate important relationships between criteria, which can be used as guidelines for refinement of the systems under evaluation, as well as contributing to the state of the art knowledge about how quantitative aspects of the systems affect the user's perceptions about them.

Key words: Spoken dialogue systems, Evaluation, Field test

* Corresponding author.
Email addresses: zoraida@ugr.es (Zoraida Callejas), rlopezc@ugr.es (Ramón

1 Introduction

Dialogue systems are becoming increasingly attractive for a wide range of applications (McTear, 2004; López-Cózar and Araki, 2005; Wahlster, 2006). In order to minimize costs and optimize results, there is a need for standard methods, architectures and criteria to test, compare and predict the performance and usability of the systems. Several initiatives have arisen since the late 80s to establish these methods. In the USA, the main funding institution for this kind of research is DARPA (Defense Advanced Research Projects Agency), with their project COMMUNICATOR (Walker et al., 2002a), which was aimed at cost-effective development of multimodal dialogue systems. This was achieved by using different plug-and-play components which were evaluated paying special attention to user satisfaction maximization. In Europe, the major institutions concerned with evaluation of dialogue systems have been COCOSDA (Coordinating Committee on Speech Databases and Speech I/O Systems Assessment), which focuses on obtaining corpora that can be shared to study evaluation criteria [1], EAGLES (1996) and DISC (1999). These last two international projects established some best practice guidelines for the development and evaluation of dialogue systems, both at system and component level.

These research efforts have successfully established a common background of criteria for quantitative evaluation. However, there is still no systematic understanding, nor consensus on the criteria that must be taken into account to optimize the usability of dialogue systems. Some projects have tried to

López-Cózar).

[1] http://www.cocosda.org/

address the problem of predicting system usability and user satisfaction from measurable performance criteria. This is the case of the PARADISE framework (Walker et al., 2000a), which has become one of the reference frameworks for system evaluation.

Because of the complexity and effort demanded by the application of the PARADISE framework, many approaches in the literature apply qualitative and quantitative measures separately. For example, Hartikainen et al. (2004) propose a methodology for subjective evaluation that has been used for evaluating the MUMS Multimodal Route Navigation System (Hurtig, 2004). Recently, the VIrtual CO-driver system (Geutner et al., 2002), the MASK multimedia service kiosk (Lamel et al., 2002) and the SAMMIE dialogue system (Becker et al., 2006), have been also evaluated only subjectively. Other authors, for example Robinson et al. (2006), evaluate their systems both with instrumentally-derived measures and quality judgments, but without establishing links between the different evaluation measures employed. In this paper we obtain empirical results on the relationship between both types of criteria from the evaluation of our spoken dialogue system. This is done via correlation studies, which we believe are a reliable method that can be applied to both whole system and component level evaluation. However, when the statistical studies are carried out over a large number of metrics, there is a possibility that some of the findings are due to chance, and thus reliability and significance studies are also reported. This method has been applied successfully for the evaluation of other dialogue systems, e.g. BoRIS (Möller, 2005), yielding some interesting relationships between evaluation criteria.

However, results in the literature are usually based on restricted laboratory interactions, in which some users are asked to interact with the system in

accordance with predefined scenarios. In some cases the users are also given evaluation questionnaires in which they express their personal opinion about different interaction aspects. The main disadvantage of this method is that the scenarios may differ from the tasks that a user would have selected in a non-predefined interaction. In contrast, field evaluation requires real users interacting with the final system in their appropriate environments. Although as stated by Bernsen and Dybkjaer (2000), field tests can fail to be representative of the full functionality of the systems, we believe they offer the most realistic results and cover real user motivations. Field evaluations are not repeatable as the interaction context is highly variable. This is also their main advantage as they gather results from different users (difference in gender, voice, knowledge, experience using the system), who talk on different devices (mobile phones, usual phones or PCs), and in different environments (different noise conditions). As the results obtained from field tests are robust to this heterogeneity, they are more relevant at predicting the real behaviour of the systems. Our contribution to the state-of-the-art system evaluation relies on obtaining new empirical evidence by means of a field study carried out employing our spoken dialogue system.

The rest of the paper is structured as follows. Section 2 presents an overview of the main evaluation trends that can be found in the literature. Section 3 briefly describes our spoken dialogue system. Section 4 describes the computation of the evaluation criteria, distinguishing between interaction parameters and quality judgments. Section 5 presents the statistical studies that we carried out, whereas Section 6 discusses the experimental results obtained. Finally, Section 7 presents the conclusions and points out some future research guidelines.

2 Related work

Evaluation of dialogue systems has been used in the literature for a wide range of purposes, for example, to measure the system's performance, to compare a system with its previous versions to measure the adequacy of changes, to compare different systems and to predict the system behaviour.

Traditionally, authors have differentiated between objective and subjective evaluation criteria. The former take into account measures computed from system performance features such as word error rate (WER). The latter consider measures that judge some property, for example intelligibility of the synthesized speech. This notation has been widely used in previous studies, for example, Larsen (2003), Minker et al. (2004) and Robinson et al. (2006). However, as argued by Möller (2005), humans are always involved in determining the systems' performance. For example, in the so-called objective measures human expert evaluators are often used (i.e. to calculate WER, experts have to compare real user input with the recognizer output). Thus, Möller (2005) proposes to differentiate between quality judgments (subjective), interaction parameters (which can be instrumentally measured or expert derived) and quality predictions (which can be instrumentally derived). In this paper we will focus on the first two categories.

There have been several attempts to create a full list of criteria to be used for evaluation by employing interaction parameters, quality predictions and quality judgments. For example, Dybkjaer and Bernsen (2000) propose a list of 15 criteria to guarantee system usability: adequate use of modalities, accurate input recognition, flexibility of the accepted vocabulary, system voice quality,

adequate response generation, adequate domain coverage, and user satisfaction, among others. The Expert Advisory Group on Language Engineering Standards (EAGLES, 1996), proposed quantitative (e.g. system response time) and qualitative (e.g. user satisfaction) measures, that were applied and interpreted following an innovative framework. This framework provided guidelines on how to carry out the evaluation and how to make results available in such a way that they could be easily interpretable and comparable. In the DISC project (DISC, 1999) there were other best practice guidelines that completed the EAGLES proposal using life cycle development methodologies. Other authors have focused on how to obtain and study speech corpora to compute evaluation measures. These are frequently large corpora extracted from system usage, or from human-to-human dialogues. In the latter case, human behaviour can be used as a baseline to compare with the system behaviour (Paek, 2001). For example, the EVALDA project (Devillers et al., 2004) focuses on evaluation 'campaigns' that consider various aspects of natural language interaction. One of them is the MEDIA campaign, which evaluates the interaction between users and dialogue systems. Their evaluation methodology employs test sets obtained from real corpora along with the commonly used evaluation criteria. Degerstedt and Jönsson (2006) proposed the LINTEST tool to carry out evaluation of dialogue systems using the JUNIT corpus. A very detailed review of the most relevant efforts on generalization of evaluation criteria and practices can be found in (Dybkjaer et al., 2004) and (López-Cózar and Araki, 2005), whereas Möller et al. (2007) present a review of the de-facto criteria extracted from all these studies and an example of their usage to evaluate a particular dialogue system.

As commented above, the PARADISE framework (Walker et al., 1998) is

the most widely embraced evaluation method proposed so far to specify the relative contribution of various factors to the overall system performance. This method models performance as a weighted function of: task success (exact scenario completion), dialogue efficiency (task duration, system turns, user turns, total turns), dialogue quality (word accuracy, response latency) and user satisfaction (sum of TTS performance, ease of task, user expertise, expected behaviour, future use). Additionally, it has been used to develop models for user satisfaction prediction, based again on the weighted linear combination of different measures (Walker et al., 2000b). The goal of this evaluation method is to maximize user satisfaction by maximizing task success and minimizing interaction costs. These costs are quantified using different efficiency and quality measures. The weights of each measure are computed via a multivariable linear regression that considers user satisfaction as the dependent variable and task success, efficiency and quality measures as independent variables. Recently, the PARADISE framework has been enhanced to enable evaluation of multimodal dialogue systems. For example, it was used in the SmartKom Project, creating the so-called PROMISE framework (Beringer et al., 2002).

The application of PARADISE to evaluate a dialogue system requires dialogue corpora extracted from controlled experiments in which users have to evaluate satisfaction on a scale after they have interacted with the system. This approach has been successfully used for evaluating and comparing eight COMMUNICATOR systems (Walker et al., 2002a,b), firstly in controlled laboratory experiments, and secondly in a less restricted context where the systems were accessible on the phone. Strictly, this second evaluation was not an open field study because the authors had control over the users, who were specifically recruited and assigned to the different systems. Nevertheless, the

tasks they had to complete were not predefined in all cases. A similar approach was employed in the ARISE project (den Os et al., 1999), where evaluation was based on the responses of users who either called a dialogue system from home or interacted with it in the laboratory. In either case, the tasks to be carried out by users were predefined (Sanderman et al., 1998).

There is no universal agreement on the distinction between "field" and "laboratory" studies. We call "field tests" the evaluations carried out taking into account real system-user interactions in which the user employs the system freely, without following predefined scenarios created by evaluators. However, in the literature some authors employ the term "field test" or "field trial" to describe studies in which the interactions are carried out by users who employ the telephone network instead of a laboratory environment, even when they are following predefined scenarios. This is the case, for example, for the evaluation of the ARISE spoken dialogue system (Baggia et al., 2000), which measures the impact of a train timetable system on the working routines of human operators, and on the callers who are traditionally served by the operators. Although the authors report experiments as "field studies", in the first experiment they contacted different callers who were asked to use the system by following different scenarios, and to fill in a questionnaire where they expressed their opinions about the system performance. In a second study, the system was enhanced taking into account the results of the first study, and it was used in a railway station in Milan. All the telephone calls were recorded and analysed. The authors obtained very interesting results about the benefits of introducing language technologies in train stations. For example, they found out that when using the dialogue system, with the same number of human operators the number of calls served per month could triplicate. How-

ever, the authors did not present any comparison between the results of their first trial (a laboratory test) and their second trial (a field test) which could have pointed out some differences in performance between laboratory and field interactions.

We can also find in the literature a distinction between "internal" and "external" tests, regarding whether they were carried out by users from the development team of the dialogue system (internal evaluation) or by users who did not have any previous knowledge about the system (external evaluation). However this is not equivalent to the "field" vs. "laboratory" studies distinction, as external tests may involve using predefined scenarios. For example, Rajman et al. (2004) propose a Rapid Dialogue Prototyping Methodology to produce, for any given application, a quickly deployable dialogue-driven interface which can be later enhanced through an iterative Wizard-of-Oz process. To refine the dialogue models developed using this methodology, the authors propose to use an internal and an external test. The internal test is used to further adapt the prototype and its successive modifications. The external test is employed for the final evaluation of the resulting dialogue interface. In both cases the evaluation was carried out in the form of a satisfaction questionnaire which was submitted to the users after they had interacted with the prototype, on the basis of a set of predefined scenarios involving specific contexts for a restaurant search.

To study the implications of using field tests, some authors have focused on non-restricted evaluation studies. This is the case of the Let's Go system (Raux et al., 2003), which was evaluated using interactions of real users who phoned the system to get information about bus schedules. The evaluation was carried out by reporting results of interaction parameters (Raux et al., 2006). Unfortu-

nately, although these parameters are relatively easy to compute, they do not provide sufficient information on quality. Qualitative judgments, on the other hand, are difficult to extract and compare when they are related to subjective opinions. Only in a few cases, performance parameters which can be measured quantitatively are also able to express quality. Our work focuses on using both quantitative and qualitative de-facto standard measures (Möller et al., 2007) in a field study, to evaluate our spoken dialogue system, which is described in Section 3. Our main objective is to empirically obtain relationships between these measures by employing statistical significance studies. Similar methods have been widely used in the area of system acceptance, more specifically for predicting the adoption of new technologies, e.g. in risk studies by companies investing in the technologies. One of the most widely used models is the Technology Acceptance Model, which relates several user judgment criteria to the final adoption of the technologies by users (Legris et al., 2003). However, no quantitative parameters are considered in this model. In the area of dialogue systems, only a few authors have exploited correlation studies to measure such relationships, for example Litman and Pan (2002), Möller (2005) and Schiel (2006), who applied them to controlled laboratory studies.

3 The UAH spoken dialogue system

Universidad Al Habla (UAH - University on the Line) is a spoken dialogue system that we developed in 2005 to provide spoken access to academic information about our Department (Callejas and López-Cózar, 2005). As shown in Figure 1, the system is comprised of the five typical modules of current spoken dialogue systems, concerned with automatic speech recognition (ASR),

dialogue management (DM), database access (DB Access), data storage (DB) and oral response generation (RG). In addition, we implemented a module called GAG (Generación Automática de Gramáticas - Automatic Grammar Generation) to automatically create ASR grammars.

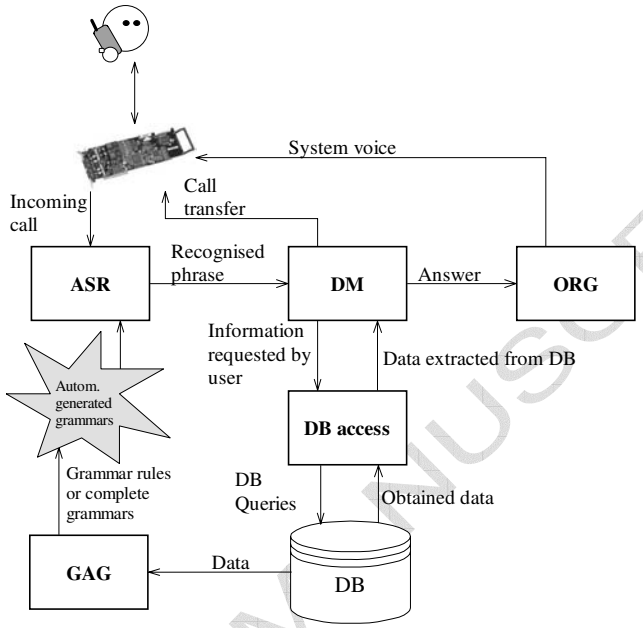


Fig. 1. Modular architecture of the UAH system

A telephony card receives the sentences uttered by the user and provides the associated speech signals to the ASR module. By employing a set of rule-based grammars representative of the permitted sentences, this module provides word sequences (recognition hypotheses) in text mode. The GAG module automatically creates the grammars employing a trigger-based technique that automatically updates the grammars if there are changes in the vocabulary, which is stored in databases. In this way, this technique avoids possible inconsistencies between the vocabulary in the databases and the vocabulary in the grammars (Callejas and López-Cózar, 2007).

The grammar rules include semantic values specified via the Semantic Inter-

pretation for Speech Recognition (SISR) standard [2] . These values are obtained when the user's utterances are recognized and they are mapped on to one or more VoiceXML ECMAScript variables.

According to the data extracted from the user utterances, the dialogue manager decides the next system response. This module employs VoiceXML documents that are dynamically created using PHP. The system employs dialogue management strategies with different flexibility (Turunen and Hakulinen, 2001). More specifically, information about professors and subjects is given by employing a system-directed initiative strategy, whereas information about registration procedures and post-graduate studies is provided by means of mixed-initiative dialogues.

The dialogue manager dynamically adapts the system responses to the context of the conversation and the dialogue state, which improves the naturalness of the interaction. For example, the help messages provided by the system take into account the topic that the user and the system are addressing at a particular moment. The context is used as well to decide the confirmation strategy to use. The UAH system uses explicit confirmations in dialogue states where it is critical to attain a very high recognition rate. For example, before transferring a call to a professor the system confirms explicitly his/her name (e.g. "Did you say professor Smith?") to avoid transferring the call to another professor. In non-critical dialogue states, the system uses explicit confirmations if the confidence scores attached to the recognized words are not high, and otherwise it uses implicit confirmations. The response generation is carried out by instantiating text patterns, and sending the obtained sentences to a TTS system.

[2] http://www.w3.org/TR/semantic-interpretation/

For purposes of illustration, we show below a sample dialogue translated from Spanish into English (S = system, U = user):

S1> Welcome to the UAH system. How can I help you?

U1> I need information about the registration procedure in the University.

S2> What do you want to know about the registration procedure?

U2> The deadline.

S3> The registration deadline is on the 31st of August. Do you need any other information about the registration procedure?

U4> No, thanks.

S5> Thank you for using the UAH system. Have a nice day.

4 Evaluation criteria

The UAH system evaluation was carried out both with interaction parameters and quality judgments. Interaction parameters were employed to measure the system performance (e.g. number or errors made by the speech recognizer), and the dialogue course (e.g. duration of the dialogue or number of turns). These measures allowed us to carry out different studies about performance and reliability of the system as well as discovering interaction points which can be improved. Although interaction parameters are a good indicator of the quality of the evaluated interaction, they do not necessarily provide reliable information about user satisfaction (López-Cózar and Araki, 2005). Thus, it is necessary to carry out a qualitative judgment evaluation to register users' opinions about these aspects of the interaction. In the experiments presented in the paper, the subjective evaluation was carried out by employing user tests.

4.1 Interaction parameters

To compute the values for the interaction parameters, we have used a dialogue corpus collected from telephone calls made to the UAH system from September 2005 to September 2006. This corpus consists of 85 dialogues and 422 user turns, with an average of 5 user turns per dialogue. Each dialogue was automatically annotated with two timestamps, corresponding to the call starting and ending times respectively. Each user utterance was stored in .wav format along with information about the recording starting time, the previous system turn, and the speech recognition result, which included confidence scores attached to the recognized words.

A human annotator took into account whether each utterance was correctly understood by the system, regardless of the speech recognition errors. For example, in response to the system prompt: "What type of information do you want?", the user answered: "I want information about a subject", but the recognition result was: "Information about subjects", where there are three deletions and one substitution. Regardless of these errors, the utterance was correctly understood by the system, as the semantic values returned by the speech recognition grammar were correct. Hence, the annotator tagged the utterance as "correctly understood".

At the dialogue level, the annotator registered the gender of the speaker, whether the dialogue was complete (i.e. whether the user did not hang up before finishing the dialogue) and whether the dialogue was successful. As it was a field study, there were no predefined tasks for the users to accomplish. Thus, a strategy had to be defined to consider dialogue success. More specifically,

we considered that the dialogues were successful when the user obtained the information he requested.

All the annotations were stored in a database from which the values for the interaction parameters were automatically computed. For example, dialogue duration was computed from the timestamps, and the number of confirmation turns was computed using the information about the previous system turn. Table 1 sets out the interaction parameters that we used in the experiments, while Figure 2 shows a sample parameter computation.

Parameter	Description	Human annotation needed
Task success	Binary value which indicates whether the user obtained from the system the information he requested	Yes
Dialogue completion	Binary value which indicates whether the user waited until the end of the dialogue before hunging up	Yes
Dialogue duration	Duration of the dialogue in seconds	No
Number of user turns	Number of user turns in the dialogue	No
Average number of words per turn in the dialogue	Avg. number of words along all the user utterances in the dialogue	Yes
WER	Number of incorrectly recognized words divided by the total number of words uttered by the user in the dialogue	Yes
Average recognition confidence	Average confidence score of the words in all the speech recognition results in the dialogue	No
% correctly understood utterances	Percentage of correctly understood utterances with respect to the total number of utterances in the dialogue	Yes
Number of confirmation turns	Number of times the system asked for explicit confirmation during the dialogue	No

Table 1
Interaction parameters employed

4.2 Quality judgments

The interaction with the UAH system starts with a welcome message in which the system introduces itself, and asks the user to visit a web page where he can complete a questionnaire with his opinion about the system performance. To be able to link the results of this test with the recordings of

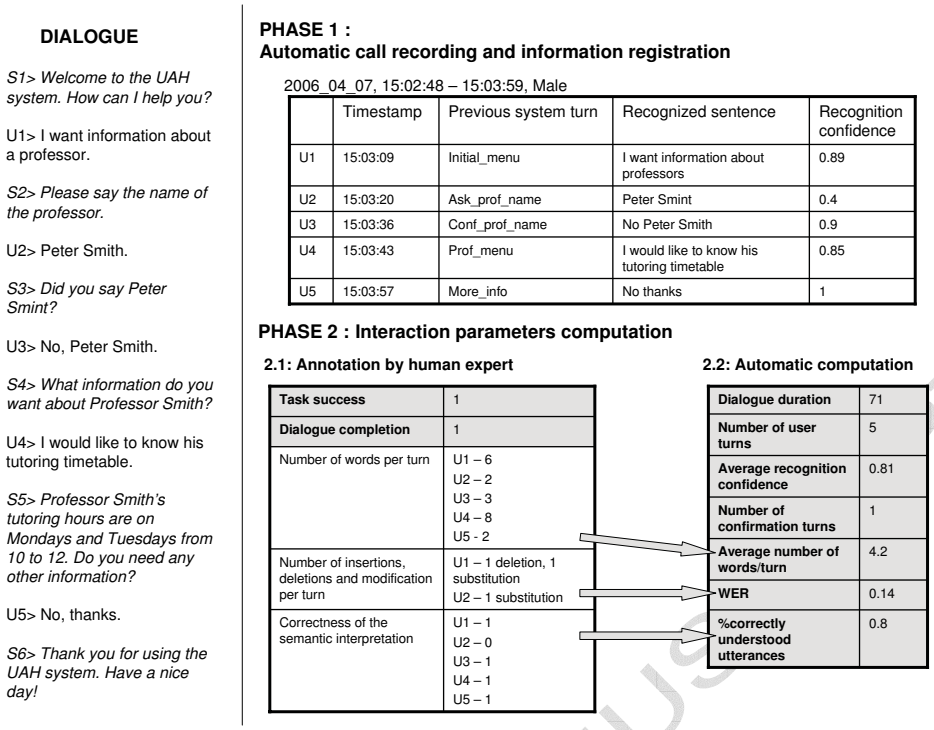


Fig. 2. Example of the computation of the interaction parameters for an UAH dialogue

the user-system interaction, the user is provided with a dialogue identification number. This number is requested in the questionnaire along with the date he made the telephone call to the system and an approximate time for the start of the interaction.

The original Spanish version of the questionnaire can be found in the Appendix (Section A.1), the English translation is as follows:

Q1. State on a scale from 1 to 5 your knowledge about new technologies for information access. (1 = "Low", 5 = "High")

Q2. State on a scale from 1 to 5 your previous experience using telephone-based dialogue systems. (1="Low", 5="High")

Q3. How many times have you used the UAH system before?

- I have not used it before.
- ..... times.

Q4. How well did the system understand you?

- Extremely bad.
- Bad.
- Fair.
- Good.
- Excellent.

Q5. How well did you understand the messages generated by the system?

- Extremely bad.
- Bad.
- Fair.
- Good.
- Excellent.

Q6. In your opinion the interaction was:

- Very slow.
- Slow.
- Adequate.
- Fast.
- Very fast.

Q7. Correcting the errors made by the system was:

- Extremely difficult.
- Difficult.
- Easy.
- Extremely easy.
- The system made no errors.

Q8. Was it easy for you to get the information that you requested?

- No, it was impossible.
- Yes, but with great difficulty.
- Yes, but with certain difficulties.
- Yes, it was easy.
- Yes, it was extremely easy.

Q9. Are you satisfied with the system performance?

- Not satisfied at all.
- Not very satisfied.
- Indifferent.
- Satisfied.
- Very satisfied.

Q10. Were you sure about what to say to the system at every moment?

- No, never.
- No, almost never.
- Sometimes.
- Yes, almost always.
- Yes, always.

Q11. Do you believe the system behaved similarly as a human would do?

- No, never.
- No, almost never.
- Sometimes.
- Yes, almost always.
- Yes, always.

The answers to each question were encoded and appropriately saved in the interactions database. All the answers excepting those corresponding to Q3 were assigned a numeric value between one and five (in the same order as they appear in the questionnaire). The values by default were: Q1=1, Q2=1, Q3=1, Q4=3, Q5=3, Q6=3, Q7=5, Q8=3, Q9=3, Q10=3, Q11=3. From the

results of the test, the measures listed in Table 2 were extracted:

Parameter	Question from which it is extracted
Knowledge about new technologies for information access	Q1
Knowledge about dialogue systems	Q2
Experience using the UAH system	Q3
Perceived extent to which UAH understands the user	Q4
Perceived extent to which the user understands UAH	Q5
Perceived interaction speed	Q6
Perceived presence of errors made by UAH	Q7
Perceived ease of UAH error correction	Q7
Perceived easy of obtaining the requested information	Q8
User satisfaction	Q9
Extent to which the user knew what to say at each moment of the interaction	Q10
Perceived human-like behaviour of the UAH system	Q11

Table 2
Perceived quality and user profile parameters employed

The first three measures listed in Table 2 are not quality judgments, but information about users. With the help of these questions, we intended to obtain an approximate idea of the users' background. However, as the UAH users were mainly students and professors of our Faculty, knowledge about new technologies for information access was high in almost all cases, as it is shown in Figure 3. Only 36% of our test participants were women.

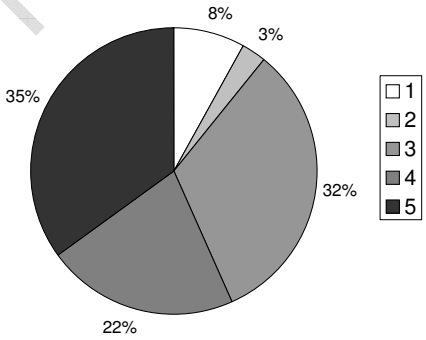


Fig. 3. Users' knowledge about new technologies for information access (1 = Low, 5 = High)

As our experiments were based on calls made by users who phoned the system on their own initiative, we think that the results obtained are very

realistic, given that the interaction was based on a real need of the users. Besides, dialogues were more heterogeneous as they take place in different contexts. The disadvantage of this approach was that, although the users were encouraged to answer the questionnaires, some of them did not do it, and thus there were no quality judgments for all the recorded dialogues. Specifically, only 37 of the 85 dialogues have subjective measures along with the objective ones. Figure 4 shows the demographic data of the two types of users: those who answered the subjective test, and those who did not.

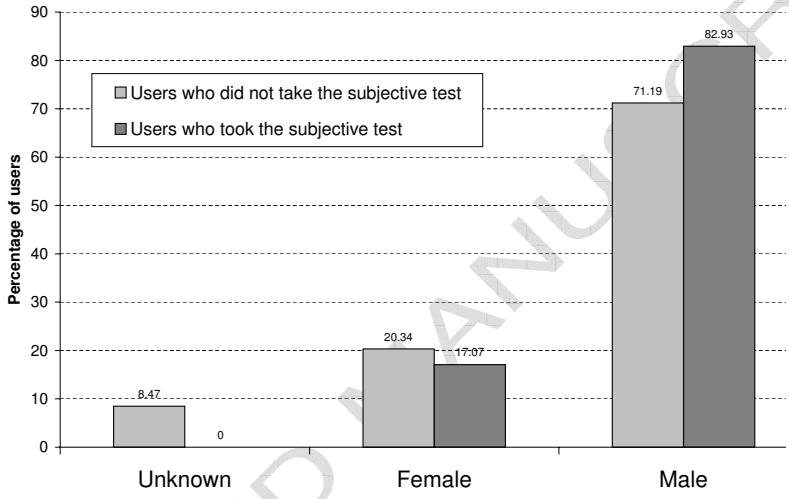


Fig. 4. Demographic data for the different user types

As can be observed, from the dialogues that corresponded to users who did not fill in the questionnaire, 8.47% were annotated with an unknown gender of the speaker. This is because these users hung up after the first prompt of the system and said nothing in response. The first system prompt clearly stated that the user was about to talk to an automatic system, and that the call was going to be recorded for research purposes. Hence, we think that two plausible reasons why some users hung up before their first turn are that they did not feel confident in talking to a computer, and that they were not happy with having their interactions recorded.

The descriptive statistics of all the parameters regarding the type of users involved are shown in Table 3, where the minimum, maximum and range values of all the measures used in our study are indicated. Section 6.1 presents a detailed study of the differences in performance and perceived quality between the interactions of these two user groups.

Parameter	User type	Range	Min.	Max.	Avg.	Typ. Dev.	Variance
Knowledge about new technologies for information access	Subj. test	4	1	5	3.77	1.14	1.30
Knowledge about dialogue systems	Subj. test	4	1	5	3.23	1.28	1.65
Experience using the UAH system	Sub. test	9	1	10	2.80	3.20	10.22
Perceived extent to which UAH understands the user	Subj. test	4	1	5	3.69	1.25	1.57
Perceived extent to which the user understands UAH	Subj. test	2	3	5	4.37	0.69	0.48
Perceived interaction speed	Subj. test	3	1	4	2.71	0.62	0.39
Perceived presence of errors made by UAH	Subj. test	1	0	1	0.54	0.50	0.25
Perceived ease of UAH error correction	Subj. test	3	1	4	2.47	0.90	0.82
Perceived easy of obtaining the requested information	Subj. test	4	1	5	3.37	1.437	2.06
User satisfaction	Subj. test	4	1	5	3.63	1.09	1.18
Extent to which the user knew what it was expected from him at each point of the dialogue	Subj. test	3	2	5	4.29	0.893	0.798
Perceived human-like behaviour of the UAH system	Subj. test	4	1	5	3.57	1.04	1.08
Task success	Subj. test	1	0	1	0.77	0.43	0.18
	No subj. test	1	0	1	0.46	0.50	0.25
Dialogue completion	Subj. test	1	0	1	0.74	0.44	0.20
	No subj. test	1	0	1	0.36	0.48	0.23
Dialogue duration	Subj. test	153	21	174	96.66	37.06	1373.70
	No subj. test	297	0	297	90.14	64.65	4179.88
Number of user turns	Subj. test	9	1	10	5.34	2.26	5.11
	No subj. test	16	1	17	4.7	3.94	15.52
Avg. words per turn	Subj.test	3	1	4	1.81	0.69	0.48
	No subj. test	4.33	0	4.33	1.73	0.78	0.61
WER	Subj. test	0.67	0.00	0.67	0.19	0.18	0.03
	No subj. test	0.83	0	0.83	0.25	0.28	0.05
Avg. recognition confidence	Subj. test	0.16	0.82	0.98	0.93	0.04	0.002
	No subj. test	0.23	0.77	1	0.93	0.05	0.003
% correctly understood utterances	Subj. test	0.50	0.50	1	0.95	0.12	0.15
	No subj. test	0.74	0.33	1	0.89	0.19	0.04
Number of confirmation turns	Subj. test	2	0	2	0.80	0.63	0.40
	No subj. test	3	0	3	0.62	0.88	0.77

Table 3
Descriptive statistics of the criteria used

5 Statistical studies employed for evaluation

In order to find relevant relationships between the criteria used, we correlated all the variables, obtaining the absolute value of the *Pearson correlation*

coefficient. However, the value of the correlation coefficient by itself was not enough to obtain reliable results, as it was also necessary to know the probability of obtaining the results by chance. This was done by computing the significance (or *p-value*) of each correlation coefficient. If the significance level was very small (less than 0.05) then the correlation was significant and the two criteria were considered linearly related.

As most of the variables were inter-correlated, we studied the effect that each criterion had on the significance of the relationships between the rest. It is possible that two criteria are correlated just because they are both affected by a third one. Thus, when eliminating the effect of this criterion, they would not be significantly correlated. To study the relationships in isolation, eliminating the effect of the rest of the criteria, we computed the *partial correlation coefficients* along with their significance levels.

The *Pearson correlation coefficient* is suitable for *scale* variables, whose values represent ordered categories with meaningful metrics, such as dialogue duration in seconds, so that distance comparisons between the values are appropriate. However, we do not only use scale variables but also *ordinal* and *dichotomous* variables (a classification can be found in Table 4). The values of the ordinal variables represent categories with an intrinsic rating, such as the perceived quality parameters described in Section 4.2. Dichotomous variables, such as "task success" or "dialogue completion", can only have two values (0 or 1 in our case). Thus, in order to obtain reliable results, we built contingency tables for the ordinal criteria. These tables allow us to study these variables and discover associations between them. To measure the strength of their relationships, we employed the *Kendall's Tau-b* and the *Spearman's rho* coefficients. The interpretation of these coefficients is equivalent to that of the

Pearson coefficient. However, as they are based on the ordinal properties of the data, their values and significances may not be the same.

Additionally, we carried out analyses of variance (ANOVA). Essentially, ANOVA models try to describe a dependent variable as the result of the weighted sum of several factors. Specifically, we used one-way ANOVA, in which there is only one independent variable, and computed the *F coefficient*. When *F's* critical level is below 0.05, it is possible to discard the average equality and conclude that not all the poblational averages that are being compared are equal. We also obtained *Eta square* which is an estimation of the degree to which each factor affects the dependent variable. To obtain more information on which to base our interpretations, especially for the case of dichotomous variables, we also obtained *Phi* and *Cramer's V* coefficients, which allow us to contrast the independence hypothesis in contingency tables.

All the experiments were carried out using the SPSS 14 [3] predictive analysis software. For the experiments in which we aimed to obtain important relationships between all the evaluation criteria including both interaction parameters and quality judgments, we used the 37 dialogues in which the users answered the subjective test. For the experiments in which we studied the possible reasons for the users to take the test or not, we used both types of dialogues (85 in total).

[3] Statistical Product and Service Solutions - http://www.spss.com/

Parameter	Type
Knowledge about new technologies for information access	Ordinal
Knowledge about dialogue systems	Ordinal
Experience using the UAH system	Ordinal
Perceived extent to which UAH understands the user	Ordinal
Perceived extent to which the user understands UAH	Ordinal
Perceived interaction speed	Ordinal
Perceived presence of errors made by UAH	Dichotomous
Perceived ease of UAH error correction	Ordinal
Perceived easy of obtaining the requested information	Ordinal
User satisfaction	Ordinal
Extent to which the user knew what it was expected from him at each point of the dialogue	Ordinal
Perceived human-like behaviour of the UAH system	Ordinal
Task success	Dichotomous
Dialogue completion	Dichotomous
Dialogue duration	Scale
Number of user turns	Scale
Avg. words per turn	Scale
WER	Scale
Avg. recognition confidence	Scale
% correctly understood utterances	Scale
Number of confirmation turns	Scale

Table 4
Type of variables used for the statistical studies

6 Evaluation results

Appendix A.2 presents a summary of the numeric results obtained from the statistical studies. For each pair of criteria, Table A.1 sets out the *Pearson correlation* coefficient and its significance level. Significance levels below 0.05 are marked in light grey, those below 0.01 are marked in dark grey, and non-significant relations are left white. Table A.2 shows a summary of the results obtained with the partial correlations. For reasons of space we have not reported all the 21 partial correlations tables with their numeric values. Instead, we report all the significant correlations found between all the tables, along with the number of control criteria for which they were significant (i.e. the number of partial correlation tables in which the relationship was significant).

As can be observed, there were no significant relations regardless of the control criteria used (i.e. none of them appeared in the 21 tables). In fact, the best case was achieved when the relationship between two criteria was shown to be significant when we eliminated the effect of 17 of the 21 variables. This showed that all the variables were deeply related. Finally, in Table A.3 we report a summary of the results for the *Tau-b* and *Rho* coefficients, only emphasizing the relations for which significance differs from those obtained in the Pearson correlation studies. In the following sections we will discuss and interpret the main findings derived from these results.

6.1 Impact of the interaction performance on the user decision to answer the subjective test

As was described in Section 4.2, not all the users answered the subjective test from which we computed the perceived quality criteria. In order to study if there were some interaction parameters that influenced the users' decision to answer the test, we introduced a dichotomous variable indicating whether the user answered the test or not, and carried out Pearson correlation and ANOVA studies to find its relationship with the interaction parameters. Table 5 shows the results obtained.

The only relations that were shown to be significant for the "user taking the subjective test" were with the "dialogue completion" and the "task success" metrics. These are two criteria that were also very significantly correlated with each other, with an *ANOVA F* of 180.159, and a 0.000 significance. *Eta square* was 0.685, and as both are dichotomous variables we also calculated *Phi* and *Cramer's V*, obtaining for both coefficients a value of 0.827 and a

Relationship	ANOVA F (Sig)	Eta square	Pearson (Sig)
Task success	7.156(0.009)	0.079	0.282(0.009)
Dialogue completion	7.775(0.007)	0.086	0.293(0.007)
Dialogue duration	0.245 (0.622)	0.003	0.054 (0.622)
Number of user turns	0.729 (0.396)	0.009	0.093 (0.396)
Avg. recognition confidence	0.122 (0.728)	0.001	-0.159 (0.150)
WER	2.107 (0.150)	0.025	0.010 (0.927)
Avg. words per turn	0.008 (0.927)	0.000	0.038 (0.728)
% correctly understood utt.	3.759 (0.056)	0.043	0.208 (0.56)
Number of confirmation turns	0.592 (0.447)	0.18	0.133 (0.447)

Table 5
Significance of the relationship between "The user taking the subjective test" and the interaction parameters

0.000 approximate significance.

One conclusion to be derived from these results is that the users carried out the subjective test mainly when they succeeded in getting the information they wanted. The fact that the successful dialogues were related to dialogue completion might be because unsuccessful dialogues were usually prematurely finished by the user.

To check whether the interaction parameters that affect task success are the same for all the user groups, we carried out additional ANOVA studies, which yielded the results shown in Table 6.

As can be observed in the table, the only differences related to task success appeared for its relationships with the number of user turns, the percentage of correctly understood words per turn, and the number of words per turn. The three relationships were significant for the users who did not answer the test, but not for those who answered it, although the first two cases can be considered as almost significant at the 0.05 level. This change might be due to the degree of cooperation of the different types of user. For example, the users who did not answer the test and had unsuccessful dialogues, hung

Relationship	User group	F	Sig
Dialogue completion - Task success	Users who did not take the subjective test	93.312	0.000
	Users that took the subjective test	19.951	0.000
	All users	180.159	0.000
Dialogue duration - Task success	Users who did not take the subjective test	17.814	0.000
	Users that took the subjective test	9.638	0.004
	All users	21.532	0.000
Number of user turns - Task success	Users who did not take the subjective test	13.025	0.001
	Users that took the subjective test	3.977	0.054
	All users	16.231	0.000
Avg. recognition confidence - Task success	Users who did not take the subjective test	0.105	0.748
	Users that took the subjective test	0.026	0.874
	All users	0.789	0.377
WER - Task success	Users who did not take the subjective test	0.171	0.681
	Users that took the subjective test	0.009	0.925
	All users	0.292	0.590
Avg. words per turn - Task success	Users who did not take the subjective test	12.787	0.001
	Users who took the subjective test	0.964	0.333
	All users	15.452	0.000
% correctly understood utt. - Task success	Users who did not take the subjective test	5.891	0.019
	Users who took the subjective test	3.992	0.054
	All users	12.539	0.001
Number of confirmation turns	Users who did not take the subjective test	0.528	0.471
	Users who took the subjective test	0.789	0.381
	All users	0.963	0.334

Table 6
ANOVA table for task success and the rest of the interaction parameters regarding the different user groups

up immediately: 70.37% of the times before the fourth user turn. However, the users who answered the subjective test were more patient and tried to overcome the interaction problems even when in the end they could not obtain the information that they were asking for.

The main difference detected between both user groups was in the relationship between the number of words per turn and task success. For the users who did not answer the test, F had a value of 12.787 and it was significant below the 0.01 level, whereas for those who answered the test, F was 0.964 and it was not significant. This was probably because the distribution of the number of words per turn for the unsuccessful and successful dialogues was more balanced in the case of the users who answered the subjective test. For them, successful and unsuccessful dialogues had a similar number of words per turn. However, the users who did not answer the test employed no more than an average of one word per turn in their unsuccessful dialogues, and more than two turns in the successful ones. Thus, an average of words per turn less or equal to one was an indicator of dialogue failure in the case of users who did not answer the subjective test.

6.2 Criteria with highest impact on user satisfaction and task success

Relationship	*Pearson* (sig)	*Tau-b* (sig)	*Rho* (sig)	*ANOVA F* (sig)
Perceived easy of obtaining the requested information - User satisfaction	0.844 (0.000)	0.750 (0.000)	0.814 (0.000)	31.071 (0.000)
Task success - User satisfaction	0.827 (0.000)	0.732 (0.000)	0.787 (0.000)	33.140 (0.000)

Table 7
Statistical significance of the most important relationships with "user satisfaction"

Table 7 shows the two highest correlation values with user satisfaction, which were obtained in all the statistical studies for the criteria "ease of obtaining information" and "task success". Thus, as we expected, a user was highly satisfied when he found it easy to get the information he wanted. However, it is remarkable that the way of gathering information had the same order of significance with user satisfaction as with the final obtaining of the informa-

tion. In (Möller, 2005), user satisfaction was also correlated with the fact that the user finally obtained the information he was looking for. However, Möller's indicator of ease of communication (which he classified as a comfort factor) did not provide a significant contribution to the overall user satisfaction. This might suggest that ease of interaction is more important for users who have a real need to obtain the information from the system compared with those for whom the interaction is carried out by following predefined scenarios.

In addition, the item of the subjective questionnaire from which the measure "perceived ease of use" is computed, implicitly takes into account the perceived success of the dialogue. Specifically, the answers to question Q8 in the questionnaire (Section 4.2) ranged from "No, it was impossible to get the information" to "Yes, it was very easy to get the information". Thus, we had two different task success measures: an interaction parameter that indicated whether the user was able to get the information that he was looking for, and another that indicated perceived task success. This second measure was extracted from the "ease of obtaining information" parameter by assigning 0 (unsuccessful) to the answer "No, it was impossible" and 1 (success) to the rest.

Contingency tables showed that both task success measures had the same value for all the dialogues. Hence, in our experiments we only considered task success as an interaction parameter. Previous studies such as (Rajman et al., 2004) found that as the users in laboratory tests are not given the possibility to contrast the information provided by the dialogue system, they trust the system responses. For example, they do not check whether the information is correct or useful. Thus, they consider the fact of obtaining a piece of information from the system equivalent to obtaining a correct result. The authors

studied this behaviour by employing laboratory test users who could not discern whether the information about restaurants, menus and prices provided by a dialogue system was correct. In our experiments, the UAH users were provided with real academic information. As they had a real need for this information, they could contrast it and know whether it was accurate or not. Thus, among the unsuccessful dialogues (both from the interaction parameter and the quality perception points of view) there were cases where the system provided information to the user but it was not what they desired, as is shown by the fact that some complete dialogues were unsuccessful. It is a benefit of test fields to allow this separation between the quality of the interaction and the quality of the results.

Within interaction parameters, there is a remarkably high correlation between dialogue completion and task success. As shown in Figure 5, although users could hang up when they received the desired information, without waiting for the system to ask if they needed any other information, if the dialogue was successful, they usually waited until the end. Although the percentage of complete and successful dialogues was higher for more collaborative users (i.e. those who answered the questionnaire), both the users that took the subjective test and those who did not take it were patient enough to wait until the end of the dialogue when it was successful.

This differs from findings of other authors. For example, Turunen et al. (2006) reported that there were highly significant differences on how the interaction was finished in field and laboratory tests carried out with the Stopman system. In the laboratory tests, 65% of the users employed an explicit request to end the call (e.g. "thank you and goodbye"). On the contrary, in the field tests less than 10% of users waited to the end of the call before hanging up.

The number of dialogues in which the users waited until the end of the interaction (i.e. the number of complete dialogues) in our field study is more than 50% higher than in that of Turunen et al. (2006).

Rajman et al. (2004) discuss that a positive attitude of users towards a system does not only depend on its behaviour, but also on the "technophile" or "technophobe" attitude of the users, although they did not control these parameters in their experimentation. In our experiments, 57% of the users rated their knowledge about new technologies for accessing information above 3 in a 1-5 scale, where 1 represented "low" and 5 "high". Thus, the collaborative nature of our users could be a result of their possible technophile disposition.

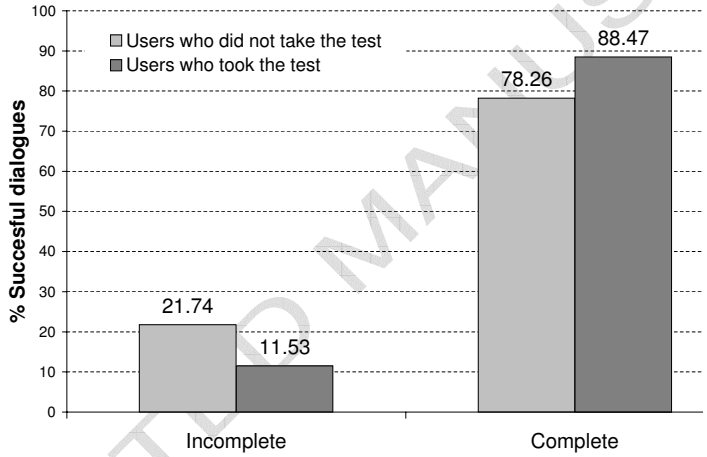


Fig. 5. Percentage of successful dialogues which are also complete regarding the different user groups

Another criterion which is highly correlated with task success and user satisfaction is the perceived ease of error correction. However, the perceived presence of errors is not significantly correlated with any of these criteria. This is probably because although in 48.19% of the successful dialogues the users detected errors, in most cases they managed to circumvent them and obtain the information they were looking for. Specifically, as shown in Figure 6, the 69.23% of the users found it "easy" or "very easy" to correct errors in the

successful dialogues. However, in the non-successful ones, 83.33% of the users found it "difficult" or "very difficult" to correct the errors.

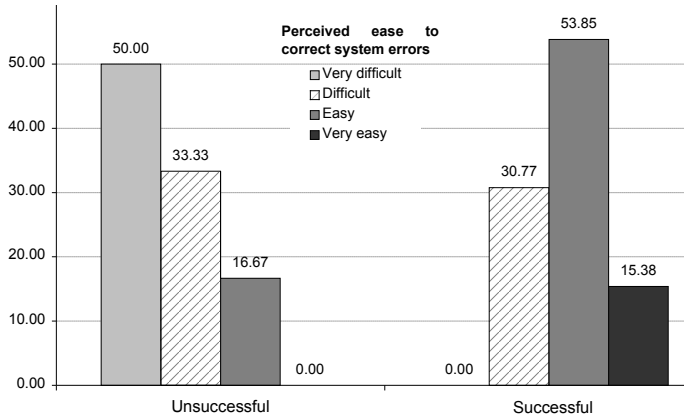


Fig. 6. Task success vs. Perceived ease of error correction

In (Möller, 2005), the users' opinion about whether misunderstandings could be easily clarified, which was classified as a contributing factor to dialogue smoothness, was not a good predictor for user satisfaction. Additionally, the author found that user satisfaction could not be fully predicted by task success, and argued that this result could be because of the unrealistic situation of the laboratory experimentation employed. We have corroborated this finding in our field study (Appendix A.2), as the subjective user tests could not be replaced by the interaction parameters employed without losing information.

6.3 Criteria with highest number of significant relations

The criterion that showed the largest number of significant correlations was the "perceived extent to which UAH understands the user". On the one hand, it was highly correlated with other quality judgments, like the degree to which the user understands the system, the perceived ease of error correction, the perceived ease of obtaining information, user satisfaction, perceived presence of errors (negative correlation in this case), and the perceived human-like

behaviour of the system. Besides, as can be observed in Table A.1, in most of these relations the significance was highest. On the other hand, this perceived quality criterion was highly correlated with interaction parameters such as completion of the dialogue, task success, dialogue duration or percentage of correctly understood utterances per dialogue.

The most significant relationships between this quality perception and other parameters were with task success and user satisfaction. A linear adjustment showed a coefficient of multiple determination of 0.55 (Figure 7), which indicates that 55% of the variability of the perceived UAH understanding could be explained by task success. Perceived system understanding, which is listed by Möller (2005) as an indicator of speech input quality, was also very significantly correlated with user satisfaction in Möller's study.

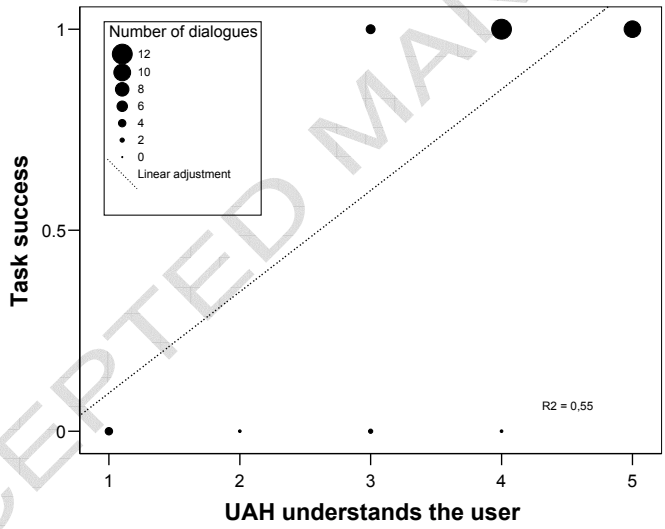


Fig. 7. Relationship between the degree to which UAH was perceived to understand the user and the task success

It is also interesting that the extent to which the user felt that the UAH system understood him was not correlated with the interaction parameters that measure the performance of the speech recognizer, such as WER or confidence scores. However, the percentage of correctly understood utterances was corre-

lated with a significance below 0.01, which indicates that from the user's point of view, speech recognition errors were not important as long as the semantic interpretations were correct and thus these errors were invisible to him. This is reflected in that the perceived presence of errors was related to the percentage of correctly understood utterances and the number of confirmation prompts, but not to WER. However, perceived ease of error correction was not significantly correlated with any of these measures. Both the perceived presence of errors and the perceived ease of correcting them were very highly correlated with the perception that the UAH system understood the user. The perceived presence of errors also negatively affected the user's confidence about what to say next during the interaction.

6.4 Impact of user's knowledge and experience

It is noteworthy that the user's knowledge about dialogue systems and new technologies for accessing information were the criteria with the lowest correlation factors with all the others. However, they were significantly correlated with each other. Thus, in our case the knowledge of the user about new technologies for information access was not determinant on the results of the interaction, not in objective terms (e.g. duration, success), nor in perceived terms (e.g. perceived speed, user satisfaction). This may be because the great majority of users had a rather high level of technical knowledge. It is possible that in experiments with other dialogue systems, where users may have more varied backgrounds, these appear to be important criteria.

The previous experience of the user employing the system ("UAH usage") was also not correlated with any of the other variables in all the statistical

studies. However, the sign of the correlation parameters indicated that experienced users perceived fewer errors, needed fewer turns to get the information, provoked fewer recognition errors and required fewer confirmation turns.

The fact that previous UAH usage was not significantly correlated with other factors, such as task success or interaction speed, differs from results found in the literature. For example, Turunen et al. (2006) stated that previous experience in using a system is a very important factor that can help to predict the success and smoothness of the dialogue. Similarly, Park et al. (2007) found that the performance of laboratory test users who had previously employed a system in very strictly predefined interactions was better than for those who had not employed it before. Other authors have studied the effect of user experience on quality judgments. For example Sturm et al. (2005) indicate that a previous prolonged use of the system helps to obtain substantial improvements in quality judgments, such as "ease of use" and "user satisfaction".

We believe that the impact of the user's experience is closely related to the type of evaluation carried out. In laboratory tests users are generally trained on how to employ the system, or at least are informed about how to interact with it. In field studies users commonly employ the system without any previous training, and this is why they are less prone to employ characteristics such as help requests (Turunen et al., 2006), of which they are sometimes not aware. However, these characteristics can be very useful to make interaction easier and to recover from error situations. On the other hand, in some particular areas of study, for example spoken dialogue systems for health applications, it has been argued that, contrary to what the previously commented studies suggest, an increasingly richer previous experience using the system does not

always imply better performance and perceived quality results. For example (Bickmore and Giorgino, 2006) report that individuals who intermittently use health dialogue systems on the telephone, compared to those who use them frequently and those who hardly use them at all, obtain the highest satisfaction levels and the best outcomes in terms of the perceived benefits. However, as discussed by Farzanfar et al. (2004), this can be due to the stress that some users experience if they feel monitored.

6.5 Impact of dialogue management initiative

To study the impact of the initiative used for dialogue management, we carried out the same computations discussed above, but distinguishing between dialogues with system-directed initiative and dialogues with mixed-initiative. The differences between both approaches are reported in Table 8, where significant correlations are marked with 'Y' (yes) and non-significant with 'N' (no).

We found that task success was approximately the same for both dialogue management approaches. This differs from the results that can be found in the literature [4], where a more flexible initiative led to considerably higher task success rates. In our experiments success was higher for mixed initiative, but the difference between both was practically negligible (77.77% of the mixed-initiative dialogues and 76.92% of the system-directed ones were successful).

However, we found that task success was related to different factors in each approach. For example, in mixed-initiative dialogues the user's confidence

[4] A comprehensive summary can be found in (Möller, 2005)

Criterion 1	Criterion 2	Mixed initiative	System-directed initiative
Perc. extent to which the user understands UAH	Perc. extent to which UAH understands the user	N	Y
Perc. interaction speed	Perc. extent to which UAH understands the user	Y	N
Perc. presence of errors made by UAH	Knowledge about dialogue systems	Y	N
Perc. presence of errors made by UAH	Perc. extent to which UAH understands the user	N	Y
User confidence about what to do next	Perc. presence of errors made by UAH	N	Y
User confidence about what to do next	Perc. ease of obtaining the requested information	N	Y
User confidence about what to do next	User satisfaction	N	Y
Perc. human-like behaviour of the UAH system	Perc. presence of errors made by UAH	N	Y
Perc. human-like behaviour of the UAH system	Perc. ease of obtaining the requested information	N	Y
Perc. human-like behaviour of the UAH system	User satisfaction	N	Y
Perc. human-like behaviour of the UAH system	User confidence about what to do next	N	Y
WER	Perc. presence of errors made by UAH	N	Y
Task success	User confidence about what to do next	N	Y
Task success	Perc. human-like behaviour of the UAH system	N	Y
Task success	Dialogue completion	Y	N
Dialogue duration	Dialogue completion	N	Y
Dialogue duration	Perc. ease of obtaining the requested information	Y	N
Dialogue duration	User satisfaction	Y	N
Dialogue duration	Task success	Y	N
Number of user turns	Task success	Y	N
Number of user turns	User satisfaction	Y	N
Number of user turns	Perc. ease of obtaining the requested information	Y	N
Dialogue completion	Perc. ease of obtaining the requested information	N	Y
Avg. recognition confidence	User satisfaction	Y	N
WER	User confidence about what to do next	N	Y
WER	Dialogue completion	N	Y
WER	Number of user turns	Y	N
% correctly understood utt.	Perc. ease of obtaining the requested information	N	Y
% correctly understood utt.	User satisfaction	N	Y
% correctly understood utt.	User confidence about what to do next	N	Y
% correctly understood utt.	Perc. human-like behaviour of the UAH system	N	Y
% correctly understood utt.	Dialogue completion	N	Y
% correctly understood utt.	Task success	N	Y
% correctly understood utt.	Avg. recognition confidence	N	Y
Number of confirmation turns	Perc. presence of errors made by UAH	N	Y
Number of confirmation turns	Dialogue duration	N	Y
Number of confirmation turns	Number of user turns	N	Y
Number of confirmation turns	Avg. recognition confidence	N	Y

Table 8
Criteria that were significantly correlated with one initiative type but not with the other

about what to do next in the dialogue was not correlated with task success, user satisfaction or perceived ease of obtaining information. On the contrary, task success had a significant correlation with user confidence in system-directed dialogues. Probably this is because the user was less constrained in the mixed-initiative interactions, and hence he did not know exactly what he could say (Figure 8). This effect did not result in bad interaction results, as task success was not reduced in the case of mixed-initiative interactions.

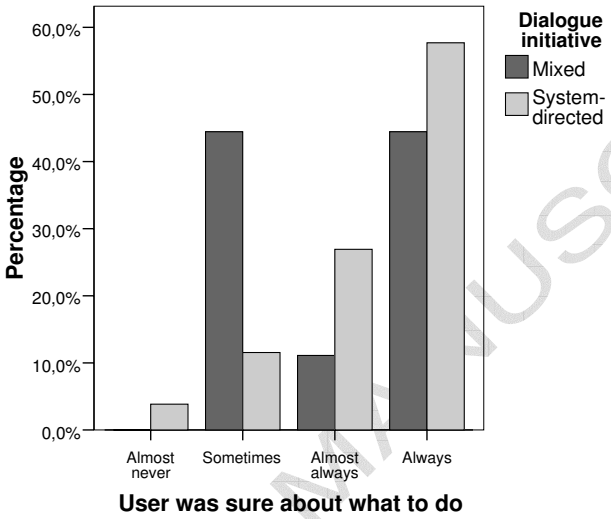


Fig. 8. Dialogue initiative influence on user confidence

Correlations of the perceived ease of obtaining information were also very different in the two cases. In the system-directed case it was related to the completion of the dialogue, the number of correctly understood utterances and the opinion that the user had about the human-like behaviour of the system. On the contrary, for mixed-initiative dialogues the perceived ease was not correlated with these measures, but with duration interaction parameters such as dialogue duration or number of user turns. The same happened with satisfaction (judgment) and task success (interaction parameter), which appeared to be highly correlated with duration measures in mixed-initiative interactions, but not in system-directed dialogues. The duration of these dialogues was sig-

nificantly correlated with user satisfaction, whereas in restricted interaction systems this was not considered so important by users. Besides, as can be observed in Figure 9, the average duration of the dialogues was shorter when the interaction was more flexible (mixed-initiative instead of system-directed).

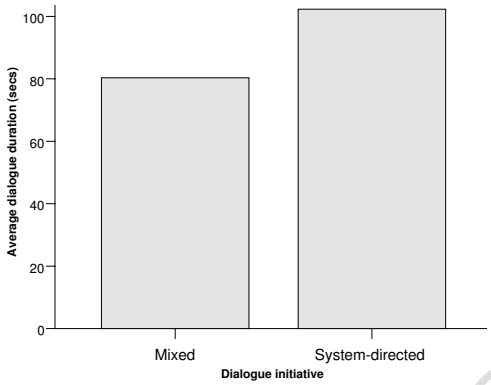


Fig. 9. Dialogue duration for each dialogue management strategy

Additionally, the perceived presence of errors was related in mixed-initiative dialogues to the user's knowledge of dialogue systems. This was not the case for system-directed initiative. Besides, it was not correlated with other measures such as user confidence, WER or number of confirmation turns, which were important factors in the system-directed dialogues.

Studies based on laboratory tests like, for example, Rajman et al. (2004) could not find clear quality perception variations with respect to predominance of system or user-driven dialogue management initiatives. Besides, some laboratory tests like the one conducted for the BoRIS system in (Möller, 2005) could not find any significant relationship between the initiative experienced by users and other interaction parameters. However, our results show that the significance of the relationships between the different evaluation criteria, including both interaction parameters and quality judgments, vary depending on the initiative used for dialogue management.

7 Conclusions and future work

In this paper we have presented a study of the relationships between several de-facto standard criteria for the evaluation of a telephone-based spoken dialogue system. Our experimental results are based on a field study using real interactions recorded from users who spontaneously telephoned the system to obtain information, without being recruited to do this.

To carry out our study we have calculated both interaction parameters (or objective measures) and quality judgments (subjective measures) by employing a corpus of real system-user interactions. Specifically, the quantitative criteria employed were: dialogue duration, dialogue completeness, task success, number of user turns, average recognition confidence, average WER, percentage of correctly understood utterances and number of confirmation turns. The qualitative measures were extracted from questionnaires that the users could optionally fill in. The criteria employed were: the extent to which the user felt correctly understood by the system, the extent to which the user understood the system messages, the perceived interaction speed, the perceived ease of error correction, the perceived presence of errors, the extent to which the user was sure about what he should do in every moment of the interaction, the extent to which the user believed the system's behaviour was human-like, and the level of user satisfaction with the interaction. Additionally information about users was also taken into account, namely: user knowledge about new technologies for information access, user knowledge about spoken dialogue systems, and number of times the user had already used the system.

Several statistical studies were developed from which significant relations

between all the criteria were extracted. This approach has not been sufficiently exploited in the literature, and some noteworthy empirical findings have been highlighted. Our empirical evidence shows that task success, perceived ease of obtaining information and perceived extent to which the system understands the user are very closely correlated with user satisfaction. These results suggest that obtaining the required information does not completely explain user satisfaction, as in some cases users judged successful dialogues as not satisfying because they found it difficult to obtain the information they are looking for. This is one of the implications derived from the usage of field tests, in which users are very concerned not only with obtaining the information they were looking for, but also with doing it easily. Furthermore, the relationship between the perceived ease of obtaining information and other criteria varies remarkably with the dialogue management strategy. Our experimental results show that, in the system-directed dialogues, perceived ease was related to the good functioning of the understanding module. On the contrary, in mixed-initiative dialogues, both user satisfaction and the perceived ease of obtaining the information seemed to be related to duration metrics. This had a strong implication in the quality judgments, as task success was highly correlated with user satisfaction in both initiatives. Thus, our results suggest that the prediction of user satisfaction also depends on the dialogue management initiative used. In the mixed-initiative dialogues it seemed to be more directly related to objective measures, such as dialogue duration. However, in more restricted dialogues, subjective measures such as the perceived extent to which the user feels that he is understood by the system, had a bigger impact. This is an important result that could indicate a need to tailor evaluation procedures to the type of interactions being analysed.

Additionally, we have studied the reasons that made some users answer the optional subjective test from which we obtained the quality judgments. We found that it was explained mainly in terms of dialogue completion and task success. Thus, the experiments that we carried out by including the users' perceptions about the quality of the system, corresponded mainly to successful dialogues, in which the users obtained the information they were looking for. This could be one of the reasons why we found that these users were very cooperative, which yielded high dialogue completion rates rarely reported in previous field test studies. Besides, contrary to what generally happens in laboratory studies, these measures consider that even when the user obtains information from the system, the dialogue cannot be considered successful if the provided information is not correct. Finally, we did not find any evidence of the effect of the users' previous experience employing the system on system performance or task success.

We believe that statistical analyses such as the ones presented in this study can lead to interesting empirical relationships that can be taken into account to enhance system development and evaluation. Besides, such studies can serve to evaluate systems as a whole instead of individual components. Future work will focus on adding factor analysis studies to group criteria and obtain the major trends that have to be taken into account. For this purpose a more extensive list of criteria will be compiled. Once the factors are computed, they will be analysed to obtain dependencies between them and build a criteria taxonomy that can then be compared with other state-of-the-art taxonomies, for example the Quality-Of-Service proposed by (Möller, 2002).

8 Acknowledgements

The authors wish to thank the reviewers for their valuable comments, which have enhanced this paper.

References

Baggia, P., Castagneri, G., Danieli, M., 2000. Field trials of the Italian ARISE train timetable system. Speech communication 11, 355–367.

Becker, T., Gerstenberger, C., Kruijff-Korbayova, I., Korthauer, A., Pinkal, M., Pitz, M., Poller, P., Schehl, J., 2006. Natural and intuitive multimodal dialogue for In-Car Applications: The SAMMIE System. In: Proc. of the 4th European Conference of Prestigious Applications of Intelligent Systems (PAIS'06). Riva del Garda, Italy, pp. 612–616.

Beringer, N., Kartal, U., Louka, K., Schiel, F., Tük, U., 2002. PROMISE: A Procedure for Multimodal Interactive System Evaluation. In: Proc. of the Workshop on Multimodal Resources and Multimodal Systems Evaluation. Las Palmas de Gran Canaria, Spain, pp. 77–80.

Bernsen, N. O., Dybkjaer, L., 2000. A methodology for evaluating spoken language dialogue systems and their components. In: Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC'00). Athens, Greece, pp. 183–188.

Bickmore, T., Giorgino, T., 2006. Health dialog systems for patients and consumers. Journal of Biomedical Informatics 39, 556–571.

Callejas, Z., López-Cózar, R., 2005. Implementing modular dialogue systems: a case study. In: Proc. of Applied Spoken Language Interaction in Distributed Environments (ASIDE'05). Aalborg, Denmark.

Callejas, Z., López-Cózar, R., 2007. Automatic creation of asr grammar rules for unknown vocabulary applications. In: Proc. of the 8th International workshop on Electronics, Control, Modelling, Measurement and Signals (ECMS'07). Liberec, Czech Republic.

Degerstedt, L., Jönsson, A., 2006. LinTest, A development tool for testing dialogue systems. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech/ICSLP). Pittsburgh, USA, pp. 489–492.

den Os, E., Boves, L., Lamel, L., Baggia, P., 1999. Overview of the ARISE project. In: Proc. of the European Conference on Speech Technology (Eurospeech'99). Budapest, Hungary, pp. 1527–1530.

Devillers, L., Maynard, H., Rosset, S., 2004. The French Media/Evalda project: the evaluation of the understanding capability of spoken language dialog systems. In: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC'04). Vol. 6. Lisbon, Portugal, pp. 2131–2134.

DISC, 1999. DISC Final Report covering the period from 1.6.98 to 28.2.99. Deliverable D5.2. Tech. rep., The DISC Consortium.

Dybkjaer, L., Bernsen, N. O., 2000. Usability issues in spoken language dialogue systems. Natural Language Engineering 6, 243–271.

Dybkjaer, L., Bernsen, N. O., Minker, W., 2004. Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication 43, 33–54.

EAGLES, 1996. Evaluation of Natural Language Processing Systems. Final report. Document EAG-EWG-PR2. Tech. rep., Center for Sprogetknologi, Copenhague, Denmark.

Farzanfar, R., Frishkopf, S., Migneault, J., Friedman, R., 2004. Telephone-linked care for physical activity: A qualitative evaluation of the use patterns of an information technology program for patients. Journal of Biomedical

Informatics 38 (3), 220–228.

Geutner, P., Steffens, F., Manstetten, D., 2002. Design of the VICO Spoken Dialogue System: Evaluation of User Expectations by Wizard-of-Oz experiments. In: Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC'02). Las Palmas de Gran Canaria, Spain.

Hartikainen, M., Salonen, E.-P., Turunen, M., 2004. Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. In: Proc. of the 7th International Conference on Spoken Language Processing (ICSLP'04). Jeju Island, Korea, pp. 2273–2276.

Hurtig, T., 2004. Visualization and multimodality: a mobile multimodal dialogue system for public transportation navigation evaluated. In: Proc. of the 8th Conference on Human-computer interaction with mobile devices and services (MobileHCI'06). Helsinki, Finland, pp. 251–254.

Lamel, L., Bennacef, S., Gauvain, J., Dartigues, H., Temem, J., 2002. User evaluation of the MASK kiosk. Speech Communication 38 (1-2), 131–139.

Larsen, L. B., 2003. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. In: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03). St. Thomas, U.S. Virgin Islands, USA, pp. 209–214.

Legris, P., Ingham, J., Collerette, P., 2003. Why do people use information technology: A critical review of the technology acceptance model. Information and Management 40, 191–204.

Litman, D. J., Pan, S., 2002. Designing and evaluating an adaptive spoken dialogue system. User Modelling and User-Adapted Interaction 12, 111–137.

López-Cózar, R., Araki, M., 2005. Spoken, Multilingual and Multimodal Dialogue Systems. Development and Assessment. John Wiley and Sons.

McTear, M. F., 2004. Spoken dialogue technology. Springer.

Minker, W., Haiber, U., Heisterkamp, P., Scheible, S., 2004. The SENECA spoken language dialogue system. Speech Communication 43, 89–102.

Möller, S., 2002. A new taxonomy for the quality of telephone services based on spoken dialogue systems. In: Proc. of the 3rd Workshop on Discourse and Dialogue (SIGDial'02). Philadelphia, USA, pp. 142–153.

Möller, S., 2005. Quality of telephone-based spoken dialogue systems. Springer.

Möller, S., Smeele, P., Boland, H., Krebber, J., 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study. Computer Speech and Language 21, 26–53.

Paek, T., 2001. Empirical methods for evaluating dialog systems. In: Proc. of the Workshop on Evaluation for Language and Dialogue Systems. Vol. 9. Toulouse, France, pp. 1–8.

Park, W., Han, S. H., Park, Y. S., Park, J., Yang, H., 2007. A framework for evaluating the usability of spoken language dialogue systems (SLDSs). Lecture Notes on Computer Science 4559, 398–404.

Rajman, M., Bui, T. H., Rajman, A., Seydoux, F., Trutnev, A., Quarteroni, S., 2004. Assessing the usability of a dialogue management system designed in the framework of a rapid dialogue prototyping methodology. Acta acustica united with acustica 90, 1906–1111.

Raux, A., Bohus, D., Black, A. W., Eskenazi, M., 2006. Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech/ICSLP). Pittsburgh, USA, pp. 65–68.

Raux, A., Langner, B., Black, A. W., Eskenazi, M., 2003. LET'S GO: Improving Spoken Dialog Systems for the Elderly and Non-natives. In: Proc. of the European Conference on Speech Technology (Eurospeech'03). Geneva, Switzerland, pp. 753–756.

Robinson, S. M., Roque, A., Vaswani, A., Traum, D., 2006. Evaluation of a spoken dialogue system for virtual reality call for fire training. In: Proc. of the 25th Army Science Conference. Orlando, USA.

Sanderman, A., Sturm, J., den Os, E., Boves, L., Cremers, A., 1998. Evaluation of the Dutch train timetable information system developed in the ARISE project. In: Proc. of IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA'98). Torino, Italy, pp. 91–96.

Schiel, F., 2006. Evaluation of multimodal dialogue systems. In: Wahlster, W. (Ed.), SmartKom: Foundations of Multimodal Dialogue Systems. Springer, pp. 617–643.

Sturm, J., Cranen, B., Terken, J., Bakx, I., 2005. Effects of prolonged use on the usability of a multimodal form-filling interface. In: Minker, W., Bühler, D., Dybkjaer, L. (Eds.), Spoken multimodal human-computer dialogue in mobile environments. Springer, pp. 329–348.

Turunen, M., Hakulinen, J., 2001. Agent-based adaptive interaction and dialogue management architecture for speech applications. In: Proc. of the 4th International Conference on Text, Speech and Dialogue. Pilsen, Czech Republic, pp. 357–364.

Turunen, M., Hakulinen, J., Kainulainen, A., 2006. Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: Similarities and differences. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech/ICSLP). Pittsburgh, USA, pp. 1057–1060.

Wahlster, W. (Ed.), 2006. SmartKom: Foundations of Multimodal Dialogue Systems. Springer.

Walker, M., Kamm, C. A., Litman, D. J., 2000a. Towards developing general models of usability with PARADISE. Natural Language Engineering, 363–377.

Walker, M., Langkilde, I., Wright, J., Gorin, A., Litman, D., 2000b. Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You. In: Proc. of the North American Meeting of the Association for Computational Linguistics. Seattle, USA, pp. 210–217.

Walker, M., Litman, D., Kamm, C., Abella, A., 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. Computer Speech and Language 12, 317–347.

Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E. O., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D., 2002a. DARPA Communicator: Cross-System Results for The 2001 Evaluation. In: Proc. of the 7th International Conference on Spoken Language Processing (ICSLP'02). Vol. 1. Denver, USA, pp. 269–272.

Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E. O., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D., 2002b. DARPA Communicator Evaluation: Progress from 2000 to 2001. In: Proc. of the 7th International Conference on Spoken Language Processing (ICSLP'02). Vol. 1. Denver, USA, pp. 273–276.

A APPENDIX

A.1 Questionnaire in Spanish

Q1. Puntúe de 1 a 5 su conocimiento de las nuevas tecnologías de acceso a la información (1="Bajo", 5="Alto").

Q2. Puntúe de 1 a 5 su uso previo de sistemas automáticos de diálogo telefónico (1="Bajo", 5="Alto").

Q3. ¿Cuántas veces había utilizado el sistema UAH con anterioridad?

- No lo había usado antes.
- ..... veces.

Q4. ¿Cómo le entendía el sistema a usted?

- Muy mal.
- Mal.
- Aceptablemente.
- Bien.
- Muy bien.

Q5. ¿Cómo entendía usted los mensajes que generaba el sistema?

- Muy mal.
- Mal.
- Aceptablemente.
- Bien.
- Muy bien.

Q6. La conversación le ha parecido:

- Muy lenta.
- Lenta.
- Adecuada.
- Rápida.
- Muy rápida.

Q7. Corregir los errores que quizás haya cometido el sistema le ha parecido:

- Muy difícil.
- Difícil.
- Fácil.
- Muy fácil.
- El sistema no ha cometido errores.

Q8. ¿Ha sido fácil averiguar la información que necesitaba conocer?

- No, ha sido totalmente imposible.
- Sí, pero con gran dificultad.
- Sí, pero con dificultad.
- Sí, ha sido fácil.
- Sí, ha sido muy fácil.

Q9. ¿Está satisfecho/a con el funcionamiento del sistema?

- No, nada.
- Casi nada.
- Indiferente.
- Satisfecho/a.
- Muy satisfecho/a.

Q10. ¿Tenía claro lo que debía hacer en cada momento del diálogo?

- No, nunca.
- Casi nunca.
- A medias.
- Casi siempre.
- Sí, siempre.

Q11. ¿Cree que el sistema se ha comportado de forma "similar" a como lo haría un ser humano en esta tarea?

- Nunca.
- Casi nunca.
- A medias.
- Casi siempre.
- Siempre.

A.2 Summary of the main experimental results

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. NT knowledge	1																				
2. DS knowledge	0.619 0.000	1																			
3. UAH usage	0.116 0.506	0.226 0.191	1																		
4. UAH understands user	0.277 0.107	0.265 0.124	0.116 0.508	1																	
5. User understands UAH	0.149 0.394	0.034 0.846	0.288 0.093	0.343 0.044	1																
6. Interaction speed	0.113 0.519	-0.063 0.719	0.222 0.200	0.334 0.050	0.323 0.058	1															
7. Easy to correct errors	0.139 0.570	-0.190 0.435	0.095 0.699	0.809 0.000	0.267 0.269	0.387 0.101	1														
8. Perc. presence of errors	-0.034 0.848	-0.197 0.258	-0.131 0.453	-0.511 0.002	-0.258 0.135	-0.147 0.399		1													
9. Perc. easiness to get info.	0.197 0.257	0.128 0.464	0.074 0.672	0.784 0.000	0.628 0.000	0.353 0.038	0.706 0.001	-0.326 0.056	1												
10. User satisfaction	0.286 0.096	0.105 0.550	0.147 0.399	0.753 0.00	0.582 0.000	0.274 0.112	0.722 0.000	-0.318 0.063	0.844 0.000	1											
11. User sure	0.124 0.479	0.018 0.917	0.093 0.597	0.476 0.004	0.539 0.001	0.204 0.239	0.234 0.335	-0.419 0.012	0.488 0.003	0.385 0.022	1										
12. UAH human behaviour	0.238 0.168	0.208 0.230	0.284 0.098	0.684 0.000	0.764 0.000	0.443 0.008	0.601 0.006	-0.385 0.022	0.761 0.000	0.690 0.000	0.580 0.000	1									
13. Dialogue completion	0.113 0.518	0.003 0.987	0.212 0.222	0.485 0.003	0.321 0.060	0.152 0.382	0.623 0.004	-0.146 0.402	0.524 0.001	0.589 0.000	0.147 0.505	0.329 0.054	1								
14. Task success	0.253 0.143	0.098 0.575	0.052 0.768	0.742 0.000	0.498 0.002	0.301 0.079	0.623 0.004	-0.226 0.191	0.912 0.000	0.827 0.000	0.331 0.052	0.637 0.000	0.614 0.000	1							
15. Dialogue duration	-0.037 0.834	-0.072 0.682	0.104 0.554	0.433 0.009	0.155 0.375	0.215 0.215	0.421 0.073	-0.177 0.310	0.348 0.040	0.375 0.026	0.245 0.157	0.226 0.191	0.462 0.005	0.475 0.004	1						
16. # user turns	0.054 0.758	0.134 0.442	-0.173 0.320	0.340 0.046	-0.046 0.792	0.030 0.865	0.160 0.513	-0.193 0.266	0.249 0.149	0.257 0.137	0.183 0.293	0.027 0.878	0.354 0.037	0.328 0.054	0.744 0.000	1					
17. Recog. confidence	0.152 0.384	0.149 0.394	-0.039 0.822	0.110 0.528	0.114 0.516	-0.032 0.855	0.133 0.587	-0.059 0.736	0.092 0.599	-0.024 0.891	0.315 0.065	0.246 0.154	0.070 0.688	-0.028 0.874	-0.027 0.876	0.098 0.576	1				
18. WER	-0.168 0.336	-0.034 0.845	-0.279 0.104	-0.157 0.369	-0.254 0.141	-0.134 0.443	0.081 0.742	0.191 0.273	-0.093 0.595	0.014 0.936	-0.426 0.011	-0.228 0.187	-0.163 0.349	-0.017 0.925	0.040 0.821	0.244 0.158	-0.526 0.001	1			
19. Words per turn	0.154 0.377	0.015 0.931	-0.057 0.745	-0.103 0.555	-0.107 0.542	-0.150 0.389	0.242 0.318	0.215 0.215	0.036 0.840	0.213 0.219	-0.158 0.364	-0.110 0.528	0.198 0.253	0.168 0.333	0.027 0.880	0.096 0.585	-0.323 0.058	0.539 0.001	1		
20. % understood utt.	0.197 0.256	0.080 0.647	0.212 0.221	0.469 0.004	0.249 0.150	-0.009 0.959	0.312 0.194	-0.341 0.045	0.350 0.040	0.495 0.002	0.398 0.018	0.379 0.025	0.291 0.089	0.329 0.054	0.115 0.510	0.027 0.878	0.440 0.008	-0.507 0.002	-0.065 0.713	1	
21. # confirmation turns	0.057 0.744	0.094 0.591	-0.297 0.083	0.363 0.032	-0.027 0.878	0.000 1.000	0.227 0.350	-0.478 0.004	0.181 0.297	0.188 0.179	0.312 0.068	0.090 0.608	0.126 0.471	0.153 0.381	0.598 0.000	0.749 0.000	0.305 0.075	0.034 0.847	-0.142 0.416	0.150 0.390	1

Table A.1
Correlations between the criteria used

Criteria relationship		Partial correlation tables in which it was significant
Perc. ease of obtaining the requested information	Perc. extent to which UAH understands the user	17
Perc. human-like behaviour of the UAH system	Perc. extent to which the user understands UAH	17
Knowledge about dialogue systems	Knowledge about new technologies for information access	17
Dialogue duration	Number of user turns	16
Number of confirmation turns	Number of user turns	16
% correctly understood utt.	WER	16
Avg. recognition confidence	WER	16
Task success	Perc. ease of obtaining the requested information	16
Perc. ease of obtaining the requested information	User satisfaction	15
Perc. human-like behaviour of the UAH system	User satisfaction	15
Avg. recognition confidence	% correctly understood utt.	15
Task success	User satisfaction	15
Dialogue completion	Task success	14
Dialogue completion	User satisfaction	14
Perc. ease of UAH error correction	Perc. ease of obtaining the requested information	14
Perc. ease of UAH error correction	Perc. extent to which UAH understands the user	14
Task success	Perc. extent to which UAH understands the user	14
Perc. extent to which UAH understands the user	User satisfaction	14
WER	Avg. words per turn	14
Perc. ease of UAH error correction	User satisfaction	13
Perc. ease of obtaining the requested information	Dialogue completion	13
Perc. ease of obtaining the requested information	Perc. human-like behaviour of the UAH system	13
Dialogue completion	Perc. ease of UAH error correction	12
Dialogue completion	Perc. extent to which UAH understands the user	12
% correctly understood utt.	User satisfaction	12
Perc. ease of UAH error correction	Task success	12
Perc. human-like behaviour of the UAH system	Task success	12
Perc. human-like behaviour of the UAH system	Perc. extent to which UAH understands the user	12
Perc. ease of UAH error correction	Perc. human-like behaviour of the UAH system	11
Dialogue duration	Perc. ease of obtaining the requested information	10
Dialogue duration	Task success	10
Dialogue duration	Perc. extent to which UAH understands the user	10
Dialogue duration	User satisfaction	10
Task success	Number of user turns	10
User satisfaction	Perc. extent to which the user understands UAH	10
Dialogue completion	Dialogue duration	9
Number of user turns	User satisfaction	7
Dialogue completion	Perc. human-like behaviour of the UAH system	6
Number of confirmation turns	Dialogue duration	5
Perc. ease of obtaining the requested information	Perc. extent to which the user understands UAH	5
Number of user turns	Perc. ease of obtaining the requested information	4
User satisfaction	Avg. words per turn	4
Number of confirmation turns	Avg. recognition confidence	3
Dialogue duration	UAH usage	2
Dialogue completion	Number of user turns	1
Dialogue completion	UAH usage	1
Dialogue duration	Perc. ease of UAH error correction	1
Dialogue duration	Perc. human-like behaviour of the UAH system	1
Number of confirmation turns	UAH usage	1
Number of confirmation turns	WER	1
Number of user turns	WER	1
Perc. ease of obtaining the requested information	% correctly understood utt.	1
Perc. ease of obtaining the requested information	Avg. recognition confidence	1
Avg. recognition confidence	Task success	1
Avg. recognition confidence	UAH usage	1
Avg. recognition confidence	User sure	1
Task success	% correctly understood utt.	1
Perc. extent to which UAH understands the user	Number of user turns	1
Perc. extent to which UAH understands the user	% correctly understood utt.	1
UAH usage	WER	1
Knowledge about dialogue systems	Perc. ease of obtaining the requested information	1
Knowledge about dialogue systems	Perc. extent to which UAH understands the user	1
User satisfaction	WER	1

Table A.2
Significant partial correlations

Criterion 1	Criterion 2	*Pearson*	*Tau-b*	*Rho*
Perc. extent to which UAH understands the user	DS knowledge	0.265 0.124	0.276 0.057	0.336 0.049
Perc. extent to which UAH understands the user	Perc. interaction speed	0.334 0.050	0.268 0.077	0.299 0.081
Perc. extent to which UAH understands the user	Dialogue completion	0.485 0.003	0.390 0.013	0.426 0.011
Perc. extent to which UAH understands the user	Dialogue duration	0.433 0.009	0.209 0.111	0.278 0.105
Perc. extent to which UAH understands the user	Number of user turns	0.340 0.046	0.157 0.255	0.197 0.257
Perc. extent to which UAH understands the user	Number of confirmation turns	0.363 0.032	0.291 0.054	0.335 0.049
Perc. extent to which the user understands UAH	Task success	0.498 0.002	0.408 0.013	0.424 0.011
Perc. human-like behaviour of the UAH system	Perc. interaction speed	0.443 0.008	0.355 0.019	0.389 0.021
Perc. human-like behaviour of the UAH system	Perc. ease of UAH error correction	0.601 0.006	0.474 0.018	0.523 0.022
Dialogue completion	Perc. ease of UAH error correction	0.623 0.004	0.559 0.011	0.602 0.006
Task success	Perc. ease of UAH error correction	0.623 0.004	0.559 0.011	0.602 0.006
Perc. easy of obtaining the required information	Perc. presence of errors made by UAH	-0.326 0.056	-0.337 0.033	-0.365 0.031
User sure	Perc. presence of errors made by UAH	-0.419 0.012	-0.429 0.008	-0.454 0.006
User sure	User satisfaction	0.385 0.022	0.291 0.054	0.316 0.064
Dialogue duration	User satisfaction	0.375 0.026	0.245 0.065	0.310 0.070
% correctly understood utt.	User satisfaction	0.495 0.002	0.223 0.151	0,248 0.151
Perc. easy of obtaining the requested information	Dialogue completion	0.524 0.001	0.384 0.015	0.416 0.013
Dialogue duration	Dialogue completion	0.462 0.005	0.350 0.014	0.421 0.012
Number of user turns	Dialogue completion	0.354 0.037	0.274 0.068	0.313 0.067
Perc. easy of obtaining the requested information	Dialogue duration	0.348 0.040	0.151 0.253	0.225 0.194
Dialogue duration	Task success	0.475 0.004	0.362 0.011	0.435 0.009
Dialogue completion	Number of user turns	0.354 0.037	0.274 0.068	0.313 0.067
User sure	WER	-0.426 0.011	-0.337 0.017	-0.388 0.021
Perc. easy of obtaining the requested information	% correctly understood utt.	0.350 0.040	0.244 0.113	0.262 0.129

Table A.3
Significance variations between *Pearson*, *Chramer's Tau-b* and *Spearman's Rho*