Design and Evaluation of a SLDS for E-Mail Access through the Telephone

Nuria Bel[(1)], Javier Caminero[(2)], Luis Hernández[(3)], Montserrat Marimón[(1)], José F. Morlesín[(4)], Josep M. Otero[(5)], José Relaño[(2)], M. Carmen Rodríguez[(2)], Pedro M. Ruz[(6)], Daniel Tapias[(6)]

(1) Gilcub-Universitat de Barcelona; Adolf Florensa s/n, E-08028 Barcelona, Spain; {nuria, montse}@gilcub.es
(2) Telefónica Investigación y Desarrollo; Emilio Vargas 6, E-28043 Madrid, Spain; {fjcg, joserg, mcrg52}@tid.es
(3) E.T.S.I. Telecomunicación; Ciudad Universitaria s/n, E-28040 Madrid, Spain; luis@gaps.ssr.upm.es
(4) Terra Networks; Edificio Ática, Vía Dos Castillas 33, E-28224 Madrid, Spain; josefmorlesin@corp.terra.es
(5) Sail-Labs; Roger de Llúria 50, E-08009 Barcelona, Spain; josep.otero@sail-labs.es
(6) Telefónica Móviles España; Serrano Galvache 56, E-28033 Madrid, Spain; {tapias_d, ruth_pm}@tsm.es

Abstract

E-MATTER (E-Mail Access through the Telephone Using Speech Technology Resources) is a Trial EC project (IST-1999-21042) directed to make e-mail universally and seamlessly accessible to a broad population of potential users through an affordable telephone-based service. Thus, the main objective of E-MATTER was to develop a Spoken Language Dialogue System (SLDS) for an e-mail access service that uses a multilingual spoken language interface (both input and output) and that takes into account the cultural and the linguistic diversity nature of the e-mail messages.

This paper addresses the different linguistic resources involved in the design and evaluation of the E-MATTER prototype. In the first part of the paper, we describe the guidelines for the design of the principal linguistic technologies involved in the development of the system: multilingual speech recognition, multilingual text-to-speech conversion, semantic parsing, dialogue management, language identification and advanced text verification. Then we present an evaluation methodology we have followed to obtain a complete analysis of both module deficiencies and global system behaviour. This methodology has been used to show us how to improve the prototype system, and we hope it will be general enough to be useful for testing other similar SLDS's.

1. The E-Matter System

E-MATTER is a multilingual Spoken Language Dialogue System (SLDS) that provides e-mail access over the telephone by combining different technologies such as continuous speech recognition, text-to-speech conversion, semantic parsing, dialogue management, language identification and text verification (E-MATTER, 1999). The user is able to access his e-mail inbox using natural language over the telephone, select one of the incoming messages, listen to it, get it replayed and reply it with voice mail. The system's users are expected to be rather varied so the user profile should not be restricted to a specific user type but allow the user to configure his profile (native language, e-mail filters, etc). This is done through a windows-based Internet interface connected with a user database. The system can read e-mail messages in all the languages of Spain (Spanish, Catalan, Galician and Basque) and in two more European languages (English and French) and the system architecture has been designed to easily allow the addition of new languages. Users can interact with the system both in Spanish and Catalan.

1.1. General functionality

The dialogue follows a menu-like structure so that it is clear to the user how that structure is built and how to move up and down in the structure. Although it is similar to a menu structure, it does not have the restrictions that such a structure has, but it makes use of the possibilities natural language offers, i.e. allows shortcuts to traverse the menu structures.

Before starting the dialogue interaction and answering the incoming call, E-MATTER connects to a mail-server to access the user's e-mail account and download new e-mails from the inbox. These e-mails are sorted according to the filters that the user has predefined in the web interface. The system then passes these e-mails to the e-mail processing system described below to prepare them for the text-to-speech synthesiser. If the user is registered and the corresponding e-mail account has been successfully accessed, the system informs about the amount of messages stored in the user's inbox. If the user has activated some filters, the system also sums up the messages for each of these folders. The user is allowed to directly select one or several folders of messages ("I want to listen to all my job messages") or let the system handle the entire inbox by reading the e-mail headers and the corresponding text body one by one. The user is also able to jump to the next header before a whole header has been read by using barge-in and it is also possible to get a header repeated or to go back to the previous one.

E-mail Processing System

The e-mail processing system receives the e-mails downloaded from the inbox and prepares them for the text-to-speech conversion: First, e-mail is pre-processed so that it is divided in several parts, i.e. the header, the body and the attachments. Then the text body of the e-mail is converted into plain text. The following step in the procedure is the language detection by the language identification system which associates the e-mail with a particular language. Later, e-mail is filtered from smileys (such as ":-)"), forwards, lines etc. converting it into a readable text. Finally, the text is passed over to the text verification system for that language, which corrects spelling mistakes that may affect the reading of the e-mail.

Reading and Replying E-mails

Once the user has chosen which e-mail he wants to listen to, the e-mail is sent to the appropriate text-to-speech synthesizer, i.e. a converter that handles the language that has been detected for that e-mail. The synthesizer then reads aloud the header and the text body. If the language detector has not been able to associate the e-mail with any of the supported languages, the e-mail is

read either in Castilian Spanish or in Catalan the interaction language selected by the user. Attachments are notified by giving the type of the attachment and the name e.g. "The attachment included is a Word document called file.doc". The only exception are the WAV files, which are played.

The user has two options to answer an e-mail: sending a predefined short text message, that indicates that the user has received and read the e-mail, or sending a voice message in WAV format.

1.2. Text and Speech Technology Resources

Figure 1 shows the block diagram of the system that illustrates the technologies involved in the project and the relation between them. The system is divided into two parts: the dialogue system and the e-mail processing system.

The dialogue system talks to the user in order to find out what the user wants, so that it is composed of a continuous speech recognizer; a semantic parser, that extracts the functions and arguments (i.e. the meaning) from the recognized sentence; a dialogue manager, which depending on the state of the dialogue, the knowledge about the task and the information obtained by the semantic parser, either interacts to the user to ask for more information, contacts the e-mail server or sends an e-mail message. Finally, the text to speech conversion system reads aloud the system messages. The e-mail processing identifies the language of the incoming e-mail message, automatically detects and corrects the misspelling errors and reads aloud the e-mail message.

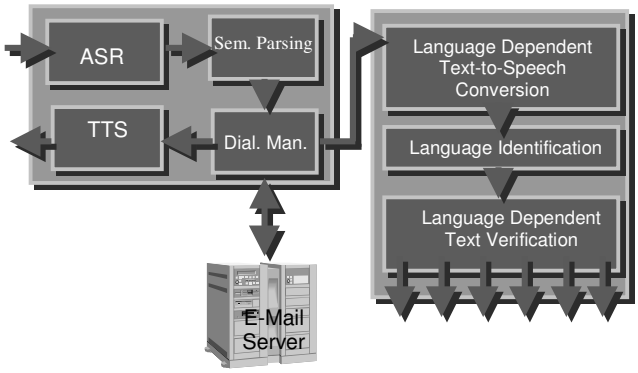

***Figure 1**: Block diagram of the technologies involved.*

Next, we briefly describe all the modules involved in the E-MATTER system:

Multilingual Speech Recognition and Text-to-Speech Conversion

E-MATTER has been implemented using the Speech Recognition and Text-to-Speech technologies developed by Telefónica I+D. The Natural Language Speech Recognition module allows users to communicate with the e-mail server by voice both in Castilian Spanish and in Catalan. It is based on context-dependent tri-phones, represented through Hidden Markov Models (Cortázar et al., 2002), and statistical language modelling with tri-grams working on parts-of-speech clustering. The TTS system is a concatenative speech synthesis system, that reads text both in Castilian Spanish and in Catalan.

Semantic Parsing

E-MATTER is a multilingual SLDS that must be able to hold dialogues in Spanish as well as in Catalan. This bilingual feature implies that we need two different parsers that should be able to produce the same semantic content. The Semantic Parser we used is an adapted version of the Phoenix parser developed at CMU (Álvarez et al., 1996), that can be described as a frame-based *concept-spotting* semantic parser. During the configuration of the semantic parser, as we have to develop grammars in two different languages, we designed a grammar structure that easily reflects the type of sentences that it should accept. We do that by specifying grammar rules reflecting what kind of specific dialogue moves or dialogue functions the user will be allowed to perform. Thus abstract things from language, and grammars in different languages were constructed separately. For example, a dialogue move such as "yanswer" was represented through a grammar rule YANSWER for all possible ways of answering yes in this kind of dialogue in one particular language.

Dialogue Management

The Dialogue Manager is rule-based and its design is based on a collaborative dialogue model. According to the classification of SLDSs proposed by J. Allen (Allen, 1997), it could be described as a system with topic-based performance capabilities, adaptive single task, a minimal pair clarification/correction dialogue model and fixed mixed-initiative (Relaño et al., 1999).

Language Identification

The automatic language identification module detects the language of the e-mail text so that the appropriate text-to-speech synthesiser can be used. This is an important issue as E-MATTER users in the multilingual environment can receive e-mails in many different languages.

The languages that the system is able to detect are all the official languages of Spain (Castilian Spanish, Catalan, Galician and Basque) and initially two other European languages (English and French). However, the identification system is developed in a way that the adaptation to new languages is easily done. The detecting algorithms used in the Language Identifier are two: the first one is based on a language dictionary (with word frequencies), the another one is based on a bigram-based language statistical model. The identification system classifies the text with a language and with a confidence value that estimates the credibility of that language classification.

Advanced Text Verification

An automatic text verification system for Spanish, Catalan, and English, is responsible for processing and correcting the text of the e-mail message. The correction of spelling errors and typical mistakes improve the quality of the reading performed by the speech converter. Additionally, as e-mails normally do not include only readable characters, the text need to be filtered to clean the text from typical e-mail garbage such as lines, repeated exclamation marks (e.g. "!!!!!!"), forwarding marks, automatic signs at the bottom, e-mail symbols such as smileys and other characters (e.g. ">>", "→").

Finally, when the text only consists of readable

characters the text is sent to the verification system. As far as English is concerned, a word checker will detect and correct most common mistakes. For Spanish and Catalan, a more advanced spelling checker has been designed. In those cases system proceeds as follows: First, each word is examined against a fullform database. When finding non-existing fullforms, the system provides a list of candidates for substituting the wrong form, together with their morpho-syntactic information in the form of tags or labels. This list takes into account both typographical errors (i.e. character deletion, insertion, substitution and transposition) and orthographic misspellings (i.e., accents, dieresis, graphie alternations). Then, a decision mechanism determines the substitution form. In case of ambiguity, i.e. more than one substitution candidate is supplied, the decision mechanism takes into consideration contextual information for candidates belonging to different categories, as well as error frequency derived from the corpus study. In Catalan, for instance, 100% of ambiguous cases are due to accentuation misspellings. The system also uses contextual information for checking that homographs that belong to different categories are appropriate in a given context.

2. E-Matter Evaluation

This section describes the different evaluation techniques we have followed to test both the particular performance of individual modules and the usability of the whole system. The usability of the E-MATTER prototype is measured from two subsets of users: users that follow two simulated but realistic testing scenarios and real users that use the system for accessing their e-mail. In this paper we only describe the evaluation carried out with the first subset of users. This evaluation phase has been done in order to have a preliminary systematic evaluation of the dialogue system. The results from this evaluation provide important information to detect and correct deficiencies both in some particular modules and in the design of the discourse structure and control strategies implemented by the prototype dialogue manager. This first diagnosis of the system has allowed us to tune it before a second evaluation of the global performance with real users.

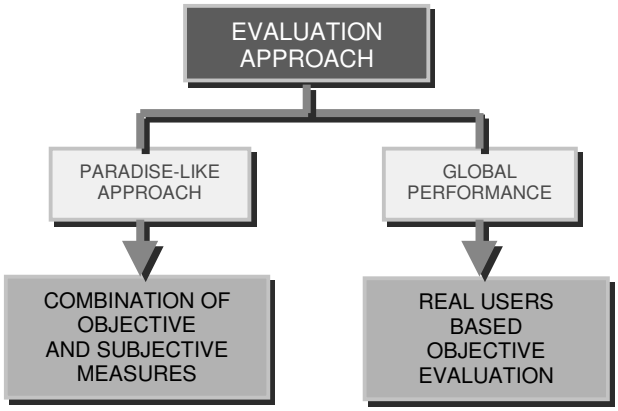

***Figure 2**: Evaluation Approach.*

An exhaustive evaluation of a SLDS is very costly and requires to have access to log-files and annotated dialogues, thus it is generally accomplished only during the first steps of the design cycle of a new system. In spite of this, during the deployment of a real system it would be highly advisable to have some simple procedures for continuously checking its usability and for discovering the impact of the performance of its different modules. Therefore, the evaluation methodology we present here has also been designed to provide a better insight on how to track the usability of the real system. This is carried out using metrics related to the user's perceived performance, that are collected from simple questionnaires filled-up by a small amount of users.

The evaluation of SLDSs is the subject of active research due to the lack of a general framework to test and compare the performance of different systems (Walker et al., 1998a; Price et al., 1992; Minker, 1998). However, SLDSs such as E-MATTER, integrate several highly-specialized Natural Language Processing (NLP) modules for most of which well-known evaluation procedures exist. During the evaluation of E-MATTER, we defined a two-level methodology that we expected to be appropriate to generalize results for similar systems or environments. These two levels correspond to module evaluation and global Spoken Dialogue evaluation respectively.

2.1. Module Evaluation Level

At this level, each NLP module integrated under E-MATTER, was tested in isolation following its corresponding standard evaluation procedure. As we explain later, depending on the module characteristics, testing data was selected from two different sources:

- General task-related data like a corpus of e-mails for the language identification system or the text verification module.
- Data extracted from recorded and/or annotated dialogues collected under the scenarios designed for testing the usability of the whole system.

The module evaluation level mainly corresponds to what is usually referred to as "glass box" evaluation (Simpson et al., 1993). In our case, we have tested the performance of the following individual modules:

- Multilingual Speech Recognition System.
- Multilingual Text-to-Speech Conversion System.
- Semantic Parser.
- Language Identification System.
- Multilingual Advanced Text Verification System.

The first three modules, Speech Recognizer, Text-to-Speech System and Semantic Parser, were tested in the context of data collected from the evaluation scenarios defined during the evaluation at dialogue level. These modules are general ones, but their performance is highly dependent on the task and their particular configuration in E-MATTER. This is specially important for the Speech Recognizer and the Semantic Parser. The remainder modules, Language Identification and Text Verification, although being also general ones, have to work with unrestricted domain data. Therefore, we decided to test them using a corpus extracted from a real e-mail database collected at Telefónica I+D.

2.2. Evaluation of the E-MATTER SLDS

This evaluation was defined at the level of the global dialogue. In this case, we have followed the PARADISE framework proposed by (Walker et al., 1998a). PARADISE tries to include and combine most of the proposed Spoken Dialogue evaluation procedures both from the efficiency and performance point of view. Therefore, several dialogue metrics for task success and both objective and subjective dialogue costs are

combined. The final aim of the evaluation procedure is to statistically describe the correlation between objective and subjective metrics in order to measure the usability of the system in terns of user satisfaction. Objective metrics are mainly derived from log-files generated while the system is working. They include number of turns, average system response time and parser and speech recogniser errors. Subjective metrics require a human evaluator to categorise an utterance or a dialogue section within the whole dialogue into two different qualitative dimensions: inappropriate system response and incorrect speech act interpretation.

For this evaluation, we relied on our previous experiences on SLDSs evaluation for ATOS and Voice Portal systems developed at Telefónica I+D using the AGORA spoken dialogue system (Relaño et al., 1999). These two agents were developed as prototypes for which we designed an experimental evaluation procedure to organize the information collected in log-files during their evaluation. This experimental evaluation framework has been already presented in (Charfuelán et al., 2000) and can be summarized in three main aspects: an annotation schema, an annotation tool and an automatic extraction of dialogue metrics from annotated corpora. Thus, from annotated dialogue databases we extract metrics and statistics like average of user and system turns, number of tasks completed, usability or user satisfaction, etc. These metrics are then used to obtain a predictive function of usability as it is proposed in the PARASISE framework (Walker et al., 1998b).

For the evaluation of E-MATTER we extended our previous dialogue annotation scheme, annotation tools and processing methodology for the evaluation metrics as it is detailed in (Charfuelán et al., 2002). Moreover, this evaluation also focussed on extending our previous results applying PARADISE, in three new directions:

1. We have tested the possibilities of using PARADISE as a model for predicting a SLDSs usability in the multilingual context of E-MATTER. There are previous research (Walker et al., 1998b) suggesting that PARADISE is able to learn a performance function on data for one system and use that as the performance function for another system (Walker et al., 2000). Then we had the opportunity to test if the usability predictive model of a system can be used to generalize across a multilingual system.
2. We have applied the multivariate analysis of PARADISE to obtain a weighted linear combination of module-related subjective metrics, obtained from user's perceptions, as a predictor of the perceived usability of the system. We hope that in some very simple "black box" evaluations, where the users only provide a single value of usability or user satisfaction, it could be useful to apply the multivariate analysis of PARADISE as predictor of a single user satisfaction metric based only on a set of user's perceived performances for the different modules of the system (Speech Recognizer, Dialogue Manager, etc.)
3. We also try to get a better insight into the correlation between objective metrics related to different modules, and user's perceptions on the performance of the same modules. We have analyzed the correlation between subjective user's perception on the behaviour of different modules and the detailed module-specific evaluation metrics. This analysis is useful because, as we have pointed out, we plan to use data from simple questionnaires filled out by some users as an indicator that reveals which module(s) need(s) to be improved during the system development.

3. Experimental Design

The evaluation was performed as a controlled test in which a set of subjects interacted with E-MATTER to complete two realistic scenarios of similar complexity. We recruited 42 subjects, each one called E-MATTER twice and used dialogue interaction to look for a particular e-mail according to 2 scenarios:

- In the first one, the user had to find an e-mail from his/her office to obtain the date of a future meeting. Then, the user was requested to reply it using a predefined text message provided by E-MATTER.
- In the second scenario, the user had to find an e-mail where his/her friends informed about the date and place of a party and then, the user had to reply it with a voiced-recorded message.

All the selected users were unexperienced on the E-MATTER system and executed 84 tasks. The testing population was divided into four different subsets in order to evaluate some features of E-MATTER such as multilinguality and text verification of e-mails:

- Subset A: 16 users who tested the E-MATTER system in Spanish and with the text-verification module enabled.
- Subset B: 10 users (dialogues in Spanish) which tested an E-MATTER version where barge-in was not allowed while reading the e-mail headers.
- Subset C: 6 users which tested E-MATTER in Spanish but the text-verification module was disabled, so that there were misspelling errors in the e-mails to be read by the TTS system.
- Subset D: 10 users which tested E-MATTER in Catalan Language and with the text-verification module active.

This testing environment allowed us to evaluate the following characteristics of the system:

- Results from users in Subset A were taken as a basic reference for the PARADISE predictive model. Evaluation for Subset B tested the impact of a particular barge-in policy.
- Results from Subset B were mainly directed to evaluate the deviation of the system's performance related to the value in Subset A, due to the presence of misspelling errors in the e-mails.
- Finally, results from Subset C were used to test how effective the predictive modeling of PARADISE is in a bilingual environment.

Before starting the test, each user was required to use the web-based enrolment interface of E-MATTER, to learn about its functionality. From here we provided a link to a web page where the user was instructed about the evaluation scenarios. Web page forms were also used to collect all the perceived metrics from the user.

So far, at the end of each call, each user had to fill out a simple yes/no questionnaire to provide information

about his/her perceived completion of the tasks. At the end of the test the user was also requested to fill out a general evaluation survey that is described in Section 4.2. During the test, the dialogues were recorded and the dialogue manager produced a log-file that, together with the results from the questionnaires, were included in our global annotation process (Charfuelán et al., 2002).

Finally, we also tested how easy a user can understand, learn and handle the control of the E-MATTER system (understandability, learnability and operatibility (Hulstijn, 2000). Therefore, we included an additional survey, using a 5 point multiple choice Likert scale. Results from this survey were used in combination with other metrics, that are shown in Section 4.2.1, to test how it can influence the usability of the system.

4. Evaluation Results

4.1. Module Evaluation

4.1.1. Speech Recognizer and Semantic Parser Evaluation

The quality of a Speech Recognizer is usually calculated in terms of word or/and sentence error rates. However, in the context of a SLDS such as E-MATTER, it is interesting to use the following metrics:

- The concept accuracy (**CA**) (Walker et al., 1998b) is, for a SLDS, a more valuable metric than word error rate since it is related to the capability of the system to understand the user's utterances. Thus, during the annotation process, we also labeled concept accuracy.
- To better characterize the recognition environment, we calculated several metrics, in terms of average number of turns for some important events: ASR rejections, no speech detection, recognition under barge-in conditions (which is very common since users interrupt the system as it is reading e-mails), user's attempts to interrupt the system when the recognizer is disabled. Noise environment, was not considered since the tests were made under moderate noise conditions.

Concerning the Semantic Parser, we evaluated its quality by means of the rate of parser errors at utterance level obtained from manual information stored in the annotated dialogues. It is important to notice that our *concept-spotting* parser can be considered as a robust filter, able to process an input sentence with recognition errors and provide a correct semantic information to the system. Then, although we measured the semantic parser errors, to properly interpret the results, it is important to notice the improvements obtained between the word error rates (**WER**) and the concept accuracy (**CA**), which are mainly due to the robustness of our semantic parser.

Table 4.1 shows the evaluation results from annotated dialogue databases (see Section 4.2.2) for Castilian Spanish (894 user turns) and Catalan (256 user turns) The high number of *barge-in* is a clear characteristic of the system (users frequently interrupt the e-mail reading process). The percentage of turns where the user speaks but the recognizer is disabled, corresponds to the population of Subset B. The differences in terms of **WER** and **CA** for different languages are due to the use of more elaborated language models and parser grammars for Spanish and only preliminary versions of them in Catalan.

***Table 4.1: S**peech Recognizer and Parser Evaluation*

ASR & Parser metrics	**Castilian Spanish**	**Catalan**
WER	45	54
PER	2.5	3.0
CA	70.50	59.60
Av. no. Turns:		
BargeIn	48	37
FalseRec	14	8.3
NoDetect	7.2	13
RecDisabled	6.3	0

WER: Word Error Rate
PER: Parser Error Rate
CA: Concept Accuracy.

Metrics as average number of turns:
BargeIn: recognition under barge-in
FalseRec: false recognition.
NoDetect: no speech detection.
RecDisabled: user's attempts to interrupt the system when the recognizer disabled.

4.1.2. Language Identification System evaluation

For the evaluation of the Language Identification system we collected a database composed of about 200 real e-mails per language: Castilian Spanish, Catalan, Galician, Basque, English and French.

The most important parameter when evaluating the Language Identifier is the Accuracy Rate in the Identification. The following table shows the accuracy rate for each language.

Table 4.2:** Accuracy of the language identification for the **combined method (dictionary & Language Model)

Language	Correct	Error
Castilian Spanish	**200/202 (99%)**	**2/202 (0.99%)**
Catalan	**198/206 (96.1%)**	**8/206 (3.89%)**
Galician	**204/204 (100%)**	**0/204 (0%)**
Basque	**203/205 (99.2%)**	**2/205 (0.8%)**
English	**249/249 (100%)**	**0/249 (0%)**
French	**205/205 (100%)**	**0/205 (0%)**
TOTAL	**1259/1271 (99.1%)**	**12/1271 (0.9%)**

It can be observed, that the accuracy of the language identification is very high (99.1% in average). However, this is not the unique relevant evaluation metric. It is also important to analyze the distribution of the confidence values of the Language Identifier. This distribution helps us to correctly interpret the meaning of these values.

The following tables (Table 4.3 and Table 4.4) represent the confidence parameter distribution, its mean and its standard deviation.

Table 4.3: *Percentage of Emails identified in each range of confidence values.*

Lang.	Cast.	Catal.	Galic.	Basq.	Engl.	Fren.
0-9	0	1.5	0.5	0.49	0	0
10-19	0	0.49	0.49	0.49	0	0
20-29	0.5	0	0.49	0	0	0.49
30-39	0.99	0.97	0.49	0	0	0
40-49	0.99	0	0	0.98	0	0.49
50-59	3	1.5	2	0.49	0	0
60-69	7.4	1.5	3.4	0	0	0.98
70-79	21	3.4	11	2	0	0
80-89	29	2.4	19	0.98	0	0
90-99	30	9.2	40	1.5	0	1.5
100	7.4	79	24	93	100	97

Table 4.4: *Confidence distribution: statistics.*

Lang..	Num. of examples	Confidence		
		Mean	Std. deviation	variance
Castilian	202	83.24	13.51	182.41
Catalan	206	94.70	16.24	263.832
Galician	204	89.33	13.28	176.38
Basque	205	97.58	11.66	135.91
English	249	100	0	0
French	205	99.01	7.19	51.63

Observing the tables there are several remarks that could be considered. Firstly, we can see that in most cases the distributions are concentrated around high confidence values. As a matter of fact, in most cases more than 50% of the e-mails were identified with the maximum value of the confidence parameter (100). There are two cases where the confidence values are lower: the ones that identify Castilian and Galician. The reason for this, is the amount of similarities between these two languages. Therefore, the decision becomes more complex and the confidence values are worse.

In general, the distribution of the confidence values changes considerably according to the considered language. In fact, the confidence values depend on many parameters, among them:

- Representativeness degree of the language model. This idea can be clearly explained with the algorithm used by the dictionaries. For instance, let's suppose that a dictionary covers 80% of the words of a language and another one covers only 40%: the first one will obtain better (i.e. higher) confidence values than the second one. Unfortunately, it is very difficult to control up to which level a dictionary covers the words of a language. Here, dictionaries with the same amount of information (about 6 MB) have been generated for the different languages. Despite our efforts, the coverage degree of each dictionary will vary according to the number of morphological variants that can be found in each language (English will be more easily covered because it has fewer morphological variants, whereas covering Castilian, Catalan, French…will be harder).
- The similarities among languages. When a language is very similar to another one (for example: Castilian and Galician) the decision will not be so clear and the confidence values will be lower.
- The set of chosen languages. When a new language is included in this set, the final decision becomes harder. In our case, the set of chosen languages does not vary.
- The set of texts used in the test. The most important variations of the confidence values are due to the features of the analyzed text. In this respect the test is a bit biased because you can find English words in e-mails written in many different languages. Obviously in the case of English emails identification that is not a problem. So the maximum value of the confidence (100) obtained for the English case is based on this feature and on its lack of similarities with other languages considered in the test.

From the experience acquired performing this test we have made out the following table. It could be a guide to use the confidence values into the E-MATTER system:

- Values from 0 to 49:
  - Interpretation: There is no confidence in the identification. Possibly the text was too short, the text was written in two or more different languages, there are many common words in two different languages, or the language of the text is not included in the set of languages used by the Language Identifier.
  - Possible actions: Try again with a longer piece of text. If the text was written in different languages each piece of text can be identified separately with the Language Identifier.
- Values from 50 to 69:
  - Interpretation: The identification is probably correct. The reasons for the low confidence values are the same than in the previous case.
  - Possible actions: The same actions as in the previous range can be performed. It is also possible to accept the result as the correct one. In this case the error probability should be considered.
- Values from 70 to 100:
  - Interpretation: Very high confidence in the result.
  - Possible actions: The result of the identification can be assumed as the correct one.

Remark: This information is illustrative. The values of the confidence parameter depend on the factors previously described. Unfortunately these factors are not easily quantifiable, and the values of the confidence parameter will vary according to these factors. For instance, Castilian and Galician texts will produce low confidence values because both languages are alike.

4.1.3. Text Verification System Evaluation.

Our first prototype of the text verification module has been evaluated on a corpus consisting of 100 e-mails (of 9,000 words approximately).

The text verification module consists of two sub-modules: the spell-checker, in charge of identifying misspelling errors and suggesting possible substitution words, and the text verifier, in charge of choosing one of the suggesting variants on the basis of contextual information and the error type.

Misspelling errors dealt with the verification module include the following two types of errors:

1. Orthographic errors: missing accents and orthographic errors due to phonetic similarity.
2. Typographical errors; i.e. character deletion, character insertion, transposition of two adjacent characters and character substitution due to proximity of keys.

On running the text verification system on the corpus, we achieved the following results:

- 64.65% of errors were detected and corrected.
- 35.34% of errors neither detected nor corrected due to ambiguity
- 0.61% of words in the testing corpus were wrongly identified as misspelling errors and corrected by the system. This was due to errors of the spell-checker module, which did not have such forms or which could not deal properly with some of the enclitics.

4.2. Spoken Dialogue System Evaluation

4.2.1. Evaluation Metrics

The metrics we used in our evaluation were chosen to provide the necessary data for applying PARADISE to obtain a model of the relationship between a representative set of objective metrics and the system usability measured through user satisfaction (Walker, 1998b). At the end of the test, each user was requested to fill out a web-based survey to obtain the Perceived Task Success and system usability. Usability was measured through questions about different aspects of the users' perceptions during the system test. We used a 5 point multiple choice Likert scale for 12 questions related to:

- TTS Performance (was the system ease to understand?)
- ASR Performance (did the system understand you?)
- Task Ease (was it easy to find the e-mail?)
- -Dialogue behavior (did the dialogue progress the way you expected?)
- Potential future use of the system.

Therefore the user satisfaction measure for each dialogue ranged from 5 to 60. The rest of the metrics were dialogue costs, related to efficiency and quality, which are necessary to apply PARADISE. We also used them to set objectives for re-designing and assessing the relative importance of different problem types in E-MATTER. These measures were derived from dialogue recordings, log-files and hand-labelling, and are summarised next (see (Charfuelán et al., 2002) for more information):

Dialogue Efficiency. We used the total elapsed time in seconds (**ET**) and the number of system turns (**ST**) and user turns (**UT**).

Dialogue Quality measures: time out prompts (**Timeouts**), number of user helps (**Helpu**), number of system helps (**Helps**), number of turns in which user is lost in the dialogue (**UserLost**), user barge in (**BargeIn**), concept accuracy (**CA**), false recognition (**FalseRec**), no speech detection (**NoDetect**) and user's attempts to interrupt the system when the recognizer is disabled (**RecDisabled**). Instead of raw counts we normalized the quality metrics by dividing the raw counts by the number of utterances in the dialogue.

4.2.2. Spoken Dialogue Evaluation Results

Table 4.5 summarizes the metrics, described above, we collected from two testing corpora: a) 64 dialogues in Castilian Spanish (Subsets A to C of Section 3, approx. 3 hours of recording, 894 user turns), and b) 20 dialogues in Catalan (Subset D, approx. 1 hour of recording, 256 user turns). Results for **CA**, **BargeIn**, **FalseRec**, **NoDetect and RecDisabled,** were given in Table 4.1.

Table 4.5 *Performance measures means*

Metrics	*Castilian Spanish*	*Catalan*
ET (seg)	168.6	158.3
UT	14.0	12.8
ST	10.4	10.6
Timeouts%	3.5	5.2
Helpu%	0.6	1.1
Helps%	2.7	1.6
UserLost%	5.4	0.7
Comp	52	15
US	39.48	37.8

In Table 4.5 we have also included the subjective metrics: user perception of task success (**Comp**) and user satisfaction (**SAT**).

Following the PARDISE framework we trained several models for the set of dialogues corresponding to the populations in Subsets A, B and D (described in Section 3). We performed stepwise multivariate linear regressions with user satisfaction as the dependent variable and the independent variables shown in Tables 4.5 and 4.1. An overall summary of our results is presented in Table 4.6, where we show which factors were found to be significant predictor of user satisfaction, ordered by degree of contribution. Table 4.6 also presents the variance in R^2, which gives an idea of the contribution of the combined factors to the variance of **US**, and is a descriptive measure of how strong is the linear association between metrics and user satisfaction.

Table 4.6: *Significant prediction factors*

Testing Population	*Factors*	R^2
Subset A	UserLost% BargeIn% ET	0.46
Subset B	UserLost%, ET, RecDisabled%	0.54
Subset D	CA, ET, Comp	0.56

For Subsets A and B, average numbers of turns where the user can not follow the dialogue (**UserLost%**) and elapsed time (**ET**), are always between the largest contributions to user satisfaction. These results reflects that the desirable characteristic of *easy of use* of the E-MATTER prototype is not accomplished. We will see this again in our analysis of the subjective metrics, and it is mainly due to the complexity for the definition and use of e-mail filters to arrange the incoming e-mails in different folders. An interesting difference between Subsets A and B is that, while the model for Subset B (no barge-in was allowed while reading the e-mail headers) reveals the negative effect of the number of turns where users try to speak and the speech recognizer is disabled (**RecDisabled%**), the model for Subset A shows that the percentage of barge-ins (**BargeIn%**) is a significant predictor of user satisfaction.

For the Subset D model (users speaking Catalan), concept accuracy (**CA**), elapsed time (**ET**) and task success (**Comp**) were the largest contributors to user satisfaction. We have to remark that the language model of the speech recognizer and the parser grammars were only preliminary versions in the Catalan E-MATTER prototype, and can not be compared to the more elaborated versions used for Castilian Spanish. This fact can explain the major impact of **CA** in Subset D compared to the results for the Castilian Spanish, where a higher and more homogeneous **CA** presented a less impact on the user satisfaction.

Although we have trained our models with only a relatively small number of dialogues, trying to corroborate these differences between Subset A and D, we made and experiment to test how the model trained on Spanish could predict user satisfaction for the Catalan. This test showed that the Spanish model only accounts for 11% of the variance in user satisfaction in the Catalonian population, and the correlation of the predicted values to the actual ones is only 0.30.

For the population in Subset C we evaluated the possible impact of uncorrected misspelling errors in the e-mails (an average of two misspelling errors per sentence) to be read by the TTS. Results showed only a moderate reduction in average user satisfaction (37.82 compared to 39.48) and on the punctuation given to the survey directly related with the quality of the synthetic voice (3.5 compared to 3.75). We think that this impact should be higher, but we assume that in this evaluation of E-MATTER it has been masked due to he major impact of other deficiencies.

Finally we have applied the multivariate linear regression analysis to the subjective metrics (12 surveys) designed to obtain users' perception of the system components. The results, both for Castilian Spanish and Catalan, reflected a larger contribution of surveys related with *Task Ease* (was it easy to find the e-mail?) *Dialogue behavior* (did you get lost in he dialogue?) and *Dialogue transparency* (how the system informed you about what it understands?). But additionally for Catalan the survey directly related to the understanding process (did the system understand you?) was found to be relevant. These results are correlated to the ones obtained with the objective metrics, specially with **UserLost** and **CA**. So we will take these results from simple subjective surveys as indicators of which module(s) need(s) to be improved in future versions of the system.

5. Acknowledgments

We are grateful to Marcela Charfuelán, Rafael Nuche, Guillermo Álvaro and Javier Molina from ETSIT-UPM (collaboration under the Spanish project TIC-1669-C04-4) for their assistance during the evaluations of the SLDS.

6. Conclusions

This work has addressed the major topics related to the design and evaluation of all he linguistic resources involved in the E-MATTER project: a multilingual Spoken Dialogue System that will provide e-mail access over the telephone.

We have described the global architecture of the system, stressing on the integration of different linguistic technologies: multilingual speech recognition, multilingual text-to-speech conversion, semantic parsing, dialogue management, language identification and advanced text verification.

We have presented a detailed and complete evaluation methodology for the E-MATTER prototype working in Castilian Spanish and Catalan. Experimental results have been obtained to provide a broad analysis of both module deficiencies and global system behavior. The methodology we have followed is general enough to be useful for obtaining a complete diagnosis of the main improvements needed for other similar SLDS's prototypes.

7. References

Álvarez J, Caminero J, Crespo C, and Tapias D. (1996) *The natural language processing module for a voice assisted operator at Telefónica I+D*, In Proceedings of ICLSP 1996, Philadelphia, USA.

Allen J. (1997) *Tutorial: Dialogue Modelling,* In ACL/EACL Workshop on Spoken Dialogue Systems, Madrid, Spain, 1997.

Charfuelán M., Esteban C. and Hernández L. (2002) *A XML –based tool for evaluation of SLDS*, In Proceedings LREC 2002.

Cortázar I, Rodríguez M, Garrido JM, Caminero J, Bernat J, Relaño J, Garijo F and Hernández L. (2002) *Últimos desarrollos en tecnologías de voz y del lenguaje*, Comunicaciones de Telefónica I+D, Enero 2002. http://www.tid.es/presencia/publicaciones/comsid/esp/24/art2.pdf

E-MATTER (1999) EC project (IST-1999-21042): E-Mail Access through the Telephone Using Speech Technology Resources, http://www.ub.es/gilcub/e-matter/

Hulstijn J. (2000) *Modelling Usability Development Methods for Dialogue Systems*, Natural Language Engineering 1 (1):1-16. 2000.

Minker W. (1998) *Evaluation methodologies for interactive speech systems*, In Proceedings LREC 1998, Granada Spain.

Price P, Hirschman L, Shriberg E and Wade. (1992), *Subject-based evaluation measures for interactive spoken language systems*, In DARPA Proceedings of Speech and Natural Language Workshop, 1992.

Relaño J, Tapias D, Rodríguez M, Charfuelán M, and Hernández L. (1999) *Robust and flexible mixed-initiative dialogue for telephone servicves*, In Proceedings EACL 1999, Bergen, Norway.

Simpson A, Fraser N. (1993) *Black box and glass box evaluation of the SUNDIAL system*, In Proceedings EUROSPEECH 1993.

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

Walker M, Candace K and Litman D. (1998b) *Towards Developing General Models of Usability with PARADISE*, Natural Language Engineering.

Walker M, Hirschman L and Aberdeen J. (2000) *Evaluation for DARPA Communicator spoken dialogue systems*, In proceedings LREC 2000.