Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA

José-Miguel Benedí[†], Eduardo Lleida[*], Amparo Varona[‡], María-José Castro[†], Isabel Galiano[†], Raquel Justo[‡], Iñigo López de Letona[‡] , Antonio Miguel[*]

[†] Universidad Politécnica de Valencia. 46022-Valencia, Spain
{jbenedi, mcastro, mgaliano}@dsic.upv.es
[*] Universidad de Zaragoza. 50015-Zaragoza, Spain
{lleida, amiguel}@posta.unizar.es
[‡] Universidad del País Vasco. 48080-Bilbao, Spain
{amparo.varona, webjublr, webloori}@we.lc.ehu.es

Abstract

In the framework of the DIHANA project, we present the acquisition process of a spontaneous speech dialogue corpus in Spanish. The selected application consists of information retrieval by telephone for nationwide trains. A total of 900 dialogues from 225 users were acquired using the *Wizard of Oz* technique. In this work, we present the design and planning of the dialogue scenes and the wizard strategy used for the acquisition of the corpus. Then, we also present the acquisition tools and a description of the acquisition process.

1. Introduction

In the last few decades, the development of speech technologies has led to speech-based solutions for several tasks. Dialogue systems (Gorin et al., 1997; Kuppevelt and Smith, 2003) are examples of these solutions where a computer interacts with a human user to solve a problem using speech dialogues. Although the current state of speech technologies does not allow the construction of general dialogue systems, domain-restricted dialogue systems have become more feasible in the last decade. Tasks such as ticket reservation or timetable consultation (Aust et al., 1995; Seneff and Polifroni, 2000; Lamel et al., 2000) have usually been considered appropriate for these systems.

In this work, we present the design and acquisition of a telephone spontaneous-speech dialogue corpus in Spanish for the DIHANA project (Benedí et al., 2004). DIHANA is aimed at the construction of a modular speech dialogue system that handles train services queries. In order to limit the domain, the queries are restricted to timetables and prices for long-distance, nationwide trains.

The acquisition of the DIHANA corpus was carried out by means of an initial prototype (DIHANA, 2005) using the *Wizard of Oz* technique (Fraser and Gilbert, 1991). This acquisition was only restricted at the semantical level (i.e., the acquired dialogues are related to a specific task domain) and was not restricted at the lexical and syntactical level (spontaneous speech). In our acquisition, this semantic control was provided by the definition of scenarios that the user must accomplish and by the wizard strategy, which defines the behavior of the acquisition system.

In Section 2, we present the design and development of both the dialogue scenarios and the wizard strategy. Moreover, a new *Wizard of Oz* server module was added to the DIHANA distributed dialog architecture, replacing the dialog management server module. This new server module allows the wizard to control the entire dialog system, to listen to the user, to receive the output from the speech recognition and understanding modules, and to send speech outputs to the user. In Section 3, we also describe the way in which the user interacts with the system-wizard in the acquisition process of the dialogues. Finally, in Section 4, we present the main characteristics of the spontaneous-speech dialogue corpus acquired.

This work has been partially supported by the Spanish MCyT under the contract (TIN2005-08660-C04-02).

2. Design of the corpus

A careful design of the acquisition is crucial to arrange the obtaining of spontaneous-speech dialogues and, at the same time, to obtain dialogues semantically restricted (domain-restricted dialogues). To accomplish both goals, we designed several types of scenarios together with a strategy and an acquisition blackboard for the wizard.

2.1. The DIHANA corpus

The DIHANA task consists of the retrieval of information about Spanish nationwide trains by telephone (DIHANA, 2005). Several types of scenarios were defined in order to control the interaction of the user with the system. A scenario is defined by:

- an objective: the information needed by the user,
- a situation: the specific circumstances related to the trip request, and
- the specific requirements of the trip: type of trip, departure city, destination city, and one or more restrictions.

An example of a scenario follows:

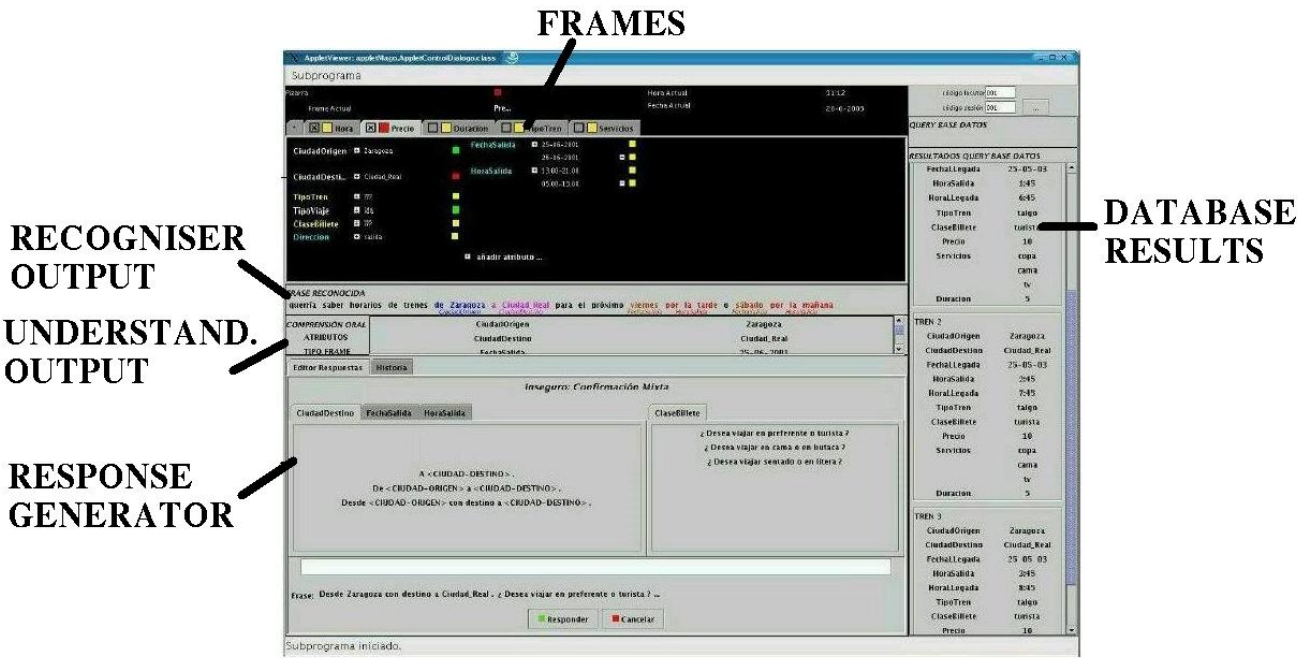

Figure 1: Acquisition platform of DIHANA project.

Goal:	obtain timetable; obtain price;
Type of trip:	one-way trip;
Situation:	go to a wedding;
Departure city:	Valencia;
Destination city:	Barcelona;
Restrictions:	leaving on Friday; arriving on Saturday; traveling by Euromed.

Three types of scenarios based on the objective to be obtained were defined: to obtain timetables for one-way trips, to obtain timetables (and, optionally, prices) for one-way trips, and to obtain timetables (and, optionally, prices) for return trips. Different situations with different restrictions were also defined. In this way, 300 different scenarios were designed, which were acquired by 225 users who performed 4 scenarios each. A total of 900 dialogues were acquired.
A fragment of a real dialogue of the task is shown below (only the English translation). The first column indicates the speaker: Machine (M) or User (U) turn :

M:	Welcome to the information system for nationwide trains, what information would you like?
U:	I would like to know the timetables of the Euromed train from Barcelona to Valencia.
M:	Do you wish to travel from Barcelona to Valencia?
U:	Yes.
M:	Do you wish to travel today?
U:	No, next Thursday.
M:	I am looking for timetables from Barcelona to Valencia for the 15th of July. One moment, please.

2.2. The Wizard of Oz strategy

The task of the wizard is to help the user get the information required by interacting with the user following a given strategy. To do this, a blackboard structure is used to provide all the information to the wizard. The output of the speech recognition and understanding servers is written in the blackboard, and its content is updated in each dialogue turn. The wizard makes use of an oral answer generation server and a text-to-speech conversion server to interact with the user. Figure 1 shows the acquisition platform with all the knowledges sources.
The wizard supervises the information generated during the dialogue process, modifying it if necessary. It can interacts with the user to do the following:

- to complete the needed information to give an answer,
- to confirm information or to clarify misunderstandings if it necessary,
- to validate information, and
- to consult the database to give an answer.

The selection of one of these modes is related to the confidence measure given by the speech recognition/understanding server and the wizard's own experience. The semantic errors from the automatic semantic module are evaluated by the wizard. If the information is not sufficient, the wizard asks the user for new information until the wizard considers that there is enough reliable information to give the required answer to the user.

3. Corpus acquisition

3.1. Acquisition platform

The corpus acquisition architecture is composed of eight conceptual components (Figure 2): an audio server (AUDIO), an automatic speech recognition server (RAH), a

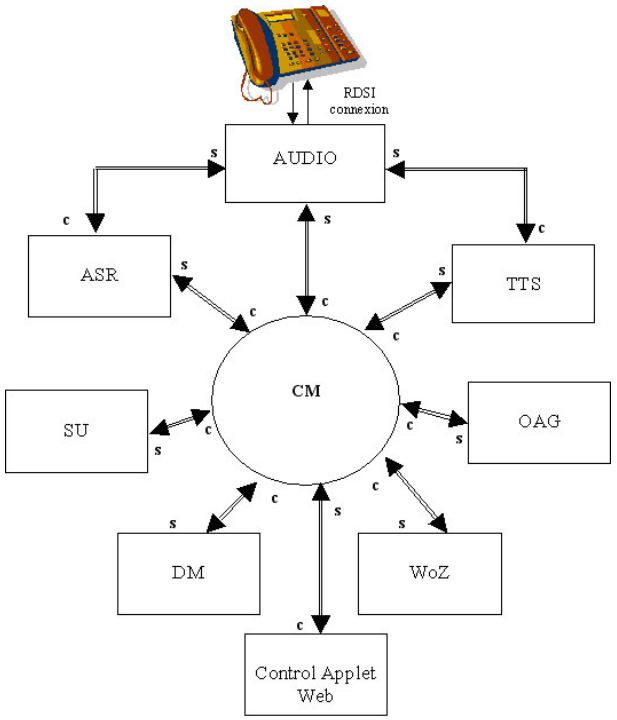

Figure 2: DIHANA architecture (c: client side, s: server side)

speech understanding server (CH), a Wizard of Oz server (MO), a dialogue manager server (GD), an oral answer generation server (GRO), a text-to-speech conversion server (CTV), and finally, a communications management client (GC). The communication between modules is performed by means of packets that are sent from one module to another using a TCP/IP protocol. All the packages of information, with the exception of the audio packages, are sent from one module to another through the communications manager using a XML format. The communications manager is controlled by an applet that is executed in a web browser.

During the acquisition process, the dialogue manager server is replaced by the Wizard of Oz server which simulates the dialogue manager behavior. The system works in the following manner. Once an incoming call is detected by the audio server, a message of "incoming call" is sent to the Wizard of Oz server to start the dialogue. A welcome message is synthesized to the user, which starts the dialogue. The audio signal is sent to both the automatic speech recognition server and the Wizard of Oz. The automatic speech recognition server sends the recognized sentence to both the speech understanding server and the Wizard of Oz server through the communication management client. The speech recognition system used during the acquisition makes use of context-dependent acoustic models (continuous Hidden Markov Models) and a stochastic language model trained with speech material obtained in a previous project (BASURDE, 2001). The speech understanding server gives an understanding frame as output, which is sent to the Wizard of Oz server. The Wizard of Oz hears the user utterance and uses the recognized sentence and the understanding frame to make the decision about the next dialogue turn. An output dialogue frame is sent to the oral answer generation server, which generates the sentence to be synthesized by the text-to-speech conversion server.

3.2. Acquisition process

The way in which the user interacts with the system-wizard in the acquisition process of the dialogues is as follows:

1. The user places a call to interact with the system. The audio signal is redirected towards the speech recognition server and to the Wizard of Oz server so that the wizard will hear the caller.

2. The speech recognition server sends the recognized string of words and the confidence measure to the speech understanding server.

3. The speech understanding server extracts the relevant information, fills the frame that is associated to the scenario, and sends the frame to the Wizard of Oz server.

4. If the frame requires more information, the wizard asks the user for more information in a new dialog turn. The process is repeated until the information is sufficient to place a query to the database.

4. Results

The designed corpus was acquired in the laboratories of the three research teams that participate in the DIHANA (2005) project in different cities of Spain (Valencia, Zaragoza and Bilbao). This corpus is composed of 900 dialogues from 225 users (153 males and 72 females). A careful manual transcription of the speech material was carried out. The main characteristic of the corpus are shown in Table 1. There are a total number of 6,278 user turns and 9,129 wizard turns; therefore, there are an average of 7 user turns and 10 wizard turns per dialogue. The average of words per user turn is 7.74.

Table 1: Summary of the main characteristics of the DIHANA database.

Environment	office
Channel	telephone line
Dialogues	900
Speakers	225
User turns	6,278
Wizard turns	9,129
Speech duration (hours)	5.3
Words	48,243
Vocabulary size	839
Phonemes	193,076

In a second stage, spontaneous-speech phenomena were labeled from an acoustic, lexical, and syntactic point of view, following a well-established annotation scheme (Rodríguez et al., 2001). The labeled spontaneous-speech events included non-speech acoustic events (noises, filled pauses,

etc.), events that distort the lexical content (cut off and mispronounced words), events that affect the grammatical structure of the utterances (speech repairs), and also discourse markers. The absolute accounts of acoustic disfluencies that appear in the database are shown in Table 2. In

Table 2: Summary of the acoustic phenomena of spontaneous speech annotated in the DIHANA database.

Noises	2,056
Silent pauses	1,847
Filled pauses	1,157
Lengthenings	1,640

the corpus, there were 499 lexical and 545 syntactical spontaneous events detected and annotated.
The corpus was divided into a training corpus that consists of 720 dialogues uttered by 180 speakers and a test corpus consisting of 135 dialogues uttered by 45 different speakers. This division was carried out in accordance with several factors, such as number of user turns, number of words, and speech duration (see Table 3).

Table 3: Characteristics of the divisions of training and test in the DIHANA database.

	Training	Test
male/female	122/58	31/14
user turns	4,929	1,350
speech duration (sec.)	30,542.20	8,361.09
number of words	38,015	10,616

Finally, in order to improve the acoustic models, two additional corpora were acquired. Each speaker read 16 different sentences (8 referred to the task, and 8 were phonetically balanced sentences); there was a total of 3600 sentences. There was a total of 10,8 hours of human voice recorded, which included 6,278 user turns and 3,600 read sentences.

5. Conclusions

In this paper, we have provided a description of the acquisition process of a spontaneous-speech dialogue corpus in Spanish. This corpus was acquired in the framework of the DIHANA (2005) project. We presented the design and planning of the dialogue scenes and the wizard strategy used for the acquisition of the corpus. We also presented the acquisition system and described the way in which the user interacts with the system-wizard in the acquisition process of the dialogues. Finally, we reported the results and the main characteristics of the corpus.

6. References

H. Aust, M. Oerder, F. Seide, and V. Steinbiss. 1995. The Philips automatic train timetable information system. *Speech Communication*, 17:249–263.

BASURDE. 2001. BASURDE Proyect (TIC98-0423-C06) http://gps-tsc.upc.es/veu/basurde/Home.htm.

J. M. Benedí, A. Varona, and E. Lleida. 2004. DIHANA: Dialogue system for information access using spontaneous speech in several environments. In *Reports of TIN Program, MEC*, pages 128–139.

DIHANA. 2005. DIHANA Proyect (TIC2002-04103-C03) http://www.dihana.upv.es.

M. Fraser and G. Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, 5:81–99.

A. Gorin, G. Riccardi, and J. Wright. 1997. How may I help you? *Speech Communication*, 23:113–127.

J. Van Kuppevelt and R. W. Smith. 2003. *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*. Springer.

L. Lamel, S. Rosset, J.L. Gauvain, S. Bennacef, M. Garnier-Rizet, and B. Prouts. 2000. The LIMSI ARISE system. *Speech Communication*, 4(31):339–353.

L.J. Rodríguez, I. Torres, and A. Varona. 2001. Annotation and analysis of disfluencies in a spontaneous speech corpus in spanish. In *Disfluency in Spontaneous Speech. An ISCA Tutorial and research workshop*, pages 1–4.

S. Seneff and J. Polifroni. 2000. Dialogue management in mercury flight reservation system. In *ANLP-NAACL*, pages 1–6.