# Evaluation of a Dialogue System Based on a Generic Model that Combines Robust Speech Understanding and Mixed-Initiative Control

**R. López-Cózar, A. J. Rubio, J. E. Díaz Verdejo, A. De la Torre**

Dpto. Electrónica y Tecnología de Computadores
Universidad de Granada, 18071 Granada, España (Spain)
Tel.: +34-958-243193, FAX: +34-958-243230, E-mail: {ramon,rubio,jedv,atv}@hal.ugr.es

**Abstract**

This paper presents a generic model to combine robust speech understanding and mixed-initiative dialogue control in spoken dialogue systems. It relies on the use of semantic frames to conceptually store user interactions, a frame-unification procedure to deal with partial information, and a stack structure to handle initiative control. This model has been successfully applied in a dialogue system being developed at our lab, named SAPLEN, which aims to deal with the telephone-based product orders and queries of fast food restaurants' clients. In this paper we present the dialogue system and describe the new model, together with the results of a preliminary evaluation of the system concerning recognition time, word accuracy, implicit recovery and speech understanding. Finally, we present the conclusions and indicate possibilities for future work.

## 1. Introduction

Computers have become indispensable in modern life. There has been a rapid growth of different kinds of remote services using computers, which has led to the development of new technologies to facilitate user access. As speech is the most natural and flexible means of communication between people, a new technology called *Spoken Dialogue Systems* appeared in the late 1980s, intended to improve speech-based communication with computers. Currently, it is possible to find spoken dialogue systems for a variety of languages and domains, working in almost real time using conventional workstations or personal computers (Gustafson et al., 1999; Asoh et al., 1999; Chao et al., 1999). These systems make use of several kinds of technology, mainly speech recognition, speech understanding and speech generation. Initially, the interaction in dialogue interfaces was carried out using written text. However, spoken dialogue systems must deal with phenomena that do not exist in written text, such as different user voices, false starts to sentences, background noise, cross-talk, out of vocabulary words, etc. Dialogue corpora and speech databases are essential in the development of these systems. On the one hand, it is necessary to determine the set of words and linguistic expressions concerning the application domain, whilst on the other, it is necessary to collect as much speech data as possible to develop the speaker-independent acoustic models (Ziegenhain et al., 1998; Pfitzinger, 1998). Although considerable advances have been made in recent years, there still remain several issues of interest which need to be addressed to make these systems acceptable for people who are not familiar with them.

## 2. The Dialogue System

We are developing a multimodal dialogue system for Spanish, termed SAPLEN, which aims to deal with the product orders and queries of fast food restaurants' clients (López-Cózar et al., 1999). We believe this system could be useful commercially in the near future, as it is designed to answer telephone calls 24 hours a day, keeping a record of product orders when the restaurant is closed to be delivered later. At the moment, the system can be used experimentally in our lab. It is implemented using two socket-interconnected machines: a SparcStation5 is used for voice signal sample acquisition and an UltraEnterprise3000 is used for recognition and text-to-speech synthesis (TTS). We initially obtained a corpus of more than 500 dialogues from recordings obtained in a fast food restaurant. This corpus was used to determine the vocabulary, the syntactic and semantic structures normally used by clients, the goals sought by clients and restaurant staff, and the set of sentences the system must generate as responses.

### 2.1. Speech recognition

The system uses a continuous-speech recogniser developed at our lab that uses context-independent phone-like units modelled by SCHMM (Semi-Continuous Hidden Markov Models). This recogniser is also currently used in another spoken dialogue system developed in our lab, called STACC, which informs students about their marks using the telephone (Rubio et al., 1997). The recogniser sampling frequency is 8 kHz, using 8bit and μ-law. It includes a voice activity detector, which is trained in a discriminative manner to distinguish between voice and background noise. The speech signal is pre-emphasised and segmented into frames 30 ms. long. The frames are overlapped and the resultant frame period is 10 ms. Every frame is analysed and represented by a vector including 14 Mel Frequency Cepstral Coefficients, the energy and the first and second derivatives. The language is modelled by class-based bigrams (Nasr et al., 1999). In total 53 classes are used. Currently, the vocabulary is about 2,000 words, including restaurant-product names, numbers, names of streets, avenues, squares, etc.

### 2.2. Speech understanding

The literature describes various speech understanding techniques (Boros & Heisterkamp, 1999; Noeth at al., 1999; Schadle et al., 1999). Some of them are based on keywords obtained from the recogniser and on semantic rules by which meaningless words are ignored. The system we are developing uses semantic rules for sentence-comprehension written taking into account key semantic concepts. This kind of analysis presents two main advantages: on the one hand, many syntactic ambiguities can be avoided if they have no semantic meaning, whilst on the other, syntactic details that have no effect on semantic analysis can be ignored. The linguistic analyser of the system uses 45 rules. It can deal with anaphors, ellipsis, ambiguity, and tautology. The dialogue manager uses a previously developed adaptive confidence threshold technique to

decide which words might have been incorrectly recognised (López-Cózar et al., 1999). In order to enhance the robustness of the system against recognition errors, the system uses an implicit recovery strategy that, on some occasions, achieves the correct semantic interpretation even if some words in the recognised sentence are wrong.

### 2.3. Dialogue management

Several dialogue management techniques can be found in literature (Rosset et al., 1999; Relaño et al., 1999). Basically, they can be divided into user directed, system directed and mixed-initiative strategies. We have set up a dialogue management strategy based on a set of goals that the system must try to achieve during the conversation with the user. For example, one goal is concerned with attempting to sell restaurant products, another is concerned with obtaining the user's phone number, etc. The system uses predefined rules to decide which goals must be activated. For instance, the *phone number* goal is activated only if the user orders at least one product. The dialogue between the system and the user is based on the context of the conversation. For example, if the system queries the user for certain information and (s)he answers giving only partial information, the answer would generate one or more sub-goals for the system. Then, the system will try to achieve these sub-goals by getting the necessary information to perform a database query or to complete an uncompleted product order. Once the system has achieved these sub-goals, it will keep on trying to achieve its original goals. Thus, the next interaction of the system is decided using a mixed-initiative dialogue control, considering both the sub-goals generated by the user and the predefined goals of the system.

The next table sets out the different recognition tasks considered, as well as the number of words for every task that can be combined into the sentences.

| Recognition task | # Different words |
|---|---|
| (1) Guided confirmation | 6 |
| (2) Free confirmation | 24 |
| (3) Information needed | 8 |
| (4) Correction | 16 |
| (5) Post code | 116 |
| (6) Product order | 158 |
| (7) Telephone number | 116 |
| (8) Address | 1723 |
| (9) Query | 220 |

Table 1. Recognition tasks considered

We call *guided confirmations* those used to confirm critical data in the fast food domain (phone number, post code, address, products ordered, price payable and estimated delivery time). For these data, the system queries for yes/no answers from the user to obtain the highest possible word accuracy. For example, to confirm a telephone number the system generates confirmations such as *"¿Has dicho 9, 5, 8, 17, 13, 28? Por favor, responde sí o no" (Did you say 9, 5, 8, 17, 13, 28? Please answer yes or no).* We call *free confirmations* those used to confirm non-critical data. In this case, the user is free to utter different confirmation words or expressions without the system previously indicating possible words, such as "*vale*" (*ok*), "*claro*" (*sure*), *"de acuerdo" (all right), "en absoluto"* (*not at all*), etc.

### 2.4. Speech generation

The SAPLEN system uses 41 sentence patterns for response generation. Various grammatical rules are used to handle gender, number and pronouns. The patterns consist of different concepts, expressions and vacant gaps. During response generation, the system expands the concepts and expressions, and fills the gaps with the appropriate words. A variety of responses can be generated for the same goal in order to enhance a friendly interaction. Text responses are generated in almost real time, in such a way that this process does not mean a delay in the overall response time. To carry out the TTS process, we use the multilingual FESTIVAL speech synthesis system developed at the CSTR of Edinburgh University (Black & Taylor, 1997). The dialogue system supports both speech and text interaction to make it accessible by clients by telephone and by Internet. It is also possible (only for developers) to interchange the interaction mode during a dialogue to check a particular response of the system at a given state of the dialogue.

## 3. The Generic Model

The generic model proposed in this paper combines robust speech understanding and mixed-initiative dialogue control. It relies on the use of semantic frames to conceptually store user interactions, a frame-unification procedure to deal with partial information, and a stack structure to handle initiative control (Dahlbäck & Jönsson, 1999).

### 3.1. Semantic frames

Each frame represents a class of elements, and is a compound of a slot set. Each slot has associated values and possible value restrictions. When the necessary slots in one frame are filled, this represents a class instance. The system we are developing uses three types of frames: one is concerned with the user's address, and the other two are concerned with fast food product orders and queries. In the frames concerned with products, one slot represents the user's intention, another represents the time when the frame is created, and several slots are concerned with the product information (amount, type, size, flavour, etc.). We say a slot is *primary* if it cannot be empty in a frame. Otherwise, we say it is *secondary*. We say a frame is *complete* if all its primary slots are filled, otherwise, we say it is *incomplete*.

### 3.2. Frame-unification procedure

The frame-unification procedure is used to add new information to the incomplete frames by filling the corresponding slots. To carry out the unification procedure, the system takes into account the speech act types (user intentions) and the four slot compatibility criteria described below. Using this procedure, it is possible to create frames from the user's interaction, even though the information (s)he provides to the system is partial or is only partially recognised. Partial information can be generated as a wrong output by the recogniser; for example, the user may order *"a ham sandwich"* and the recogniser output might be *"one sandwich"*. Partial information can also be generated by the user; for example, (s)he may utter *"I want one sandwich"* without indicating the type of sandwich (*ham*, *cheese*, etc.). The unification procedure considers that two or more frames can be unified if the following compatibility criteria are satisfied:

- **Speech act compatibility**. Two frames are considered *compatible* if both share the same speech act type. For example, the sentences *"One sandwich, please"* and *"What is the price of a ham sandwich?"* are incompatible and therefore the corresponding frames cannot be unified. However, the sentences *"I want a sandwich"* and *"I want ham"* are considered compatible and therefore it is possible to unify the corresponding frames.

- **Incremental compatibility**. The aim of unification is to fill empty slots by considering the time when they are created. If the most recent frames do not include new data, the unification process does not take place. For example, if in one interaction the user says s(he) wants a *sandwich* and in the next interaction s(he) says again that s(he) wants a *sandwich*, then the corresponding frames cannot be unified as the latest one does not include any new information.

- **Structural compatibility**. The frames that can be unified must have the same slot structure. For example, if one frame contains two slots and another contains three slots, then the two frames cannot be unified. This criterion enables the frame-unification procedure to operate simultaneously with different types of frame, considering at each step only the frames with the same structure.

- **Content compatibility**. The content of non-empty slots in the frames that can be unified must be the same. For example, if the user says (s)he wants *a beer* and *a coffee*, the two corresponding frames cannot be unified.

The use of conceptual frames in addition to the frame-unification procedure sets up a speech understanding strategy that is robust against some kinds of recognition errors. For example, the semantic frames allow the system to understand correctly sentences with errors in Spanish gender. So, the sentence *"un bocadillo de jamón" –a ham sandwich-* can be correctly understood even though the recogniser produces a gender recognition error (*"una bocadillo de jamón"*) as output. The system is also robust against hesitation expressions or repetition of some words generally generated by the user when (s)he is thinking what to say (for example, *"uh ...one ... uh ... one ham sandwich"*). This type of robustness against recognition errors is called *implicit recovery*, as the system itself is able to repair the error without the intervention of the user.

### 3.3. Stack structure

The stack structure is used to handle mixed-initiative dialogue control. Each new frame is created at the top of the stack. As the frames in the stack direct the focus of the conversation, starting with the frame at the top, the current focus is determined by the most recent frame. If this frame is incomplete, the system asks the corresponding questions in order to fill the primary slots of the frame. If the frame at the top is complete, the system continues searching down the stack for an incomplete frame. If one is found, the system queries the user to fill its primary empty slots. If no incomplete frames are found, the system continues the conversation in order to reach its next goal.

This procedure permits a mixed-initiative dialogue control. The system generally takes the initiative to reach its goals, as well as the sub-goals generated by the user. However, the user can answer the system's questions unexpectedly, taking the initiative in the conversation. For example, (s)he may answer a system query with a question, which may change the focus of the conversation. This question would create a new frame(s) at the top of the stack that becomes a new sub-goal(s) for the system. When the new sub-goals are reached, the system returns to its previous goals, going back to the previous context in the conversation.

## 4. Evaluation

In order to carry out a preliminary evaluation of the system, 6 male speakers who were familiar with the system recorded (in our lab) a test sentence corpus containing 100 sentences per recognition task. 126 different bigrams were created for the sentence recognition. The bigrams were compiled from sets of sentences concerning the different recognition tasks. These sentences were automatically created by combining all the words in the word classes, following the syntactic and semantic structures in the dialogue corpus obtained from the recordings made in a fast food restaurant. 109 bigrams were used to recognise the user data (post code, telephone number and address) and the 17 remaining bigrams were used for the other recognition tasks (see Table 1).

Word error rate and recognition time are inversely-related measures. Generally, a low word error rate requires the exploration of a considerable number of candidate sentences, which tends to a high recognition time. In order to enhance the performance of the recogniser, the dialogue manager of the system selects the most adequate bigram according to the state of the dialogue. A pruning threshold (Pt) was experimentally associated to every task, considering word error rate and recognition time restrictions. The effects of 6 different pruning thresholds on recognition time, word accuracy, implicit recovery and sentence understanding were measured. The following table sets out the recognition times obtained for the lab sentences.

| Recog. Task | Pt=10 | Pt=20 | Pt=30 | Pt=40 | Pt=50 | Pt=60 |
|---|---|---|---|---|---|---|
| (1) | 3.04 | 3.05 | 3.05 | 3.05 | 3.05 | 3.06 |
| (2) | 3.17 | 3.18 | 3.20 | 3.21 | 3.23 | 3.24 |
| (3) | 3.68 | 3.70 | 3.79 | 4.02 | 4.33 | 4.54 |
| (4) | 3.65 | 3.64 | 3.71 | 3.76 | 3.94 | 4.24 |
| (5) | 3.50 | 3.55 | 3.65 | 3.81 | 4.0 | 4.31 |
| (6) | 3.65 | 3.65 | 3.75 | 3.92 | 4.05 | 4.32 |
| (7) | 4.02 | 4.12 | 4.91 | 4.93 | 5.84 | 7.05 |
| (8) | 5.35 | 5.82 | 6.25 | 7.14 | 8.35 | 10.30 |
| (9) | 4.35 | 6.35 | 10.14 | 15.55 | 24.15 | 38.20 |

Table 2. Recognition time (seconds) for several Pt

As can be observed in Table 2, the recognition of tasks (1)-(6) took less than 5 sec. on average; the recognition of telephone numbers (7) took 5.22 sec., the recognition of addresses (8) took 7.20 sec. and the recognition of queries (9) took 16.45 sec. The next table sets out the word accuracy (WA) results obtained for the same sentences.

| Recog. Task | Pt=10 | Pt=20 | Pt=30 | Pt=40 | Pt=50 | Pt=60 |
|---|---|---|---|---|---|---|
| (1) | 60.0 | 85.0 | 93.0 | 100 | 100 | 100 |
| (2) | 20.0 | 48.0 | 55.0 | 60.0 | 61.0 | 61.0 |
| (3) | 5.0 | 10.0 | 12.0 | 41.0 | 58.0 | 65.0 |
| (4) | 48.0 | 72.0 | 82.0 | 85.0 | 90.0 | 91.0 |
| (5) | 76.63 | 92.42 | 98.62 | 98.62 | 98.96 | 98.96 |
| (6) | 67.16 | 79.10 | 83.58 | 85.07 | 86.56 | 93.41 |
| (7) | 27.53 | 52.17 | 80.67 | 90.33 | 90.82 | 91.30 |
| (8) | 51.35 | 80.99 | 81.08 | 90.99 | 91.20 | 93.72 |
| (9) | 45.97 | 57.47 | 80.45 | 81.60 | 83.90 | 85.05 |

Table 3. Word accuracy for several Pt

Table 3 shows that word accuracy increased as the pruning threshold was raised. The greatest score corresponded to the recognition of the guided confirmations (1), as only two confirmation words: "*sí*" (*yes)* and "*no*" (*no*) and some other words (such as "*uh*") were possible (Lavelle et al., 1999). The word accuracy for the *free confirmations* (2) and for the *information needed* sentences (3) was very low. Concerning the free confirmations, many words were changed for other acoustically similar (like "*no*" and "*not*"). Moreover, there were many insertions of hesitation words, which were included in the bigrams. Nevertheless, a large proportion of these errors was implicitly recovered. Concerning the *information needed* sentences, no specific bigram was defined for this task, as users may ask for information at any moment during the dialogue. The corresponding score was calculated when the *product order* bigram was the active one. In consequence, many insertions occurred as the recogniser outputs tended to follow the syntactic structures of the *product order* sentences. Many of these errors were also implicitly recovered. The next table sets out the implicit recovery results obtained.

| Recog. Task | Pt=10 | Pt=20 | Pt=30 | Pt=40 | Pt=50 | Pt=60 |
|---|---|---|---|---|---|---|
| (1) | 17.50 | 26.66 | 42.85 | 0 | 0 | 0 |
| (2) | 43.54 | 65.0 | 83.78 | 78.12 | 81.81 | 81.81 |
| (3) | - | - | - | - | - | - |
| (4) | 51.85 | 33.33 | 0 | 0 | 0 | 0 |
| (5) | - | - | - | - | - | - |
| (6) | 36.36 | 42.85 | 50.0 | 40.0 | 50.0 | 36.36 |
| (7) | - | - | - | - | - | - |
| (8) | 6.25 | 12.50 | 16.66 | 33.33 | 33.33 | 33.33 |
| (9) | 13.33 | 20.0 | 55.55 | 66.66 | 75.0 | 66.0 |

Table 4. Implicit recovery for several Pt

As can be observed in Table 4, there was no implicit recovery for the understanding of post codes (5) nor for the understanding of telephone numbers (7), since an error in a digit (or pair of digits) made the corresponding semantic interpretation wrong. No implicit recovery was found in the experiments for the *information needed* task (3). The next table sets out the sentence understanding (SU) results obtained.

| Recog. Task | Pt=10 | Pt=20 | Pt=30 | Pt=40 | Pt=50 | Pt=60 |
|---|---|---|---|---|---|---|
| (1) | 67.0 | 89.0 | 96.0 | 100 | 100 | 100 |
| (2) | 65.0 | 86.0 | 93.0 | 94.0 | 94.0 | 94.0 |
| (3) | 48.0 | 55.0 | 75.0 | 81.0 | 86.0 | 90.0 |
| (4) | 87.0 | 90.0 | 91.0 | 94.0 | 95.0 | 96.0 |
| (5) | 48.0 | 80.0 | 88.0 | 88.0 | 96.0 | 96.0 |
| (6) | 42.0 | 65.0 | 83.0 | 87.0 | 90.0 | 91.0 |
| (7) | 4.0 | 34.0 | 51.0 | 64.0 | 67.0 | 67.0 |
| (8) | 25.0 | 65.0 | 75.0 | 90.0 | 90.0 | 91.0 |
| (9) | 60.0 | 65.0 | 80.0 | 85.0 | 88.0 | 90.0 |

Table 5. Sentence understanding for several Pt

Table 5 shows that the *telephone number* task (7) achieved the smallest understanding percentage, since the bigram permitted any digit or pair of digits to be followed by one hundred words (digits or pairs of digits). Even if the word accuracy is very high, an error in one word means that the telephone number would be incorrectly understood.

The next table sets out the pruning thresholds that could be the most appropriate for every task, considering the results obtained and the real time restrictions of a typical interactive system.

| Recog. Task | Selected Pt | Time | WA | SU |
|---|---|---|---|---|
| (1) | 50 | 3.05 | 100 | 100 |
| (2) | 60 | 3.24 | 61.0 | 94.0 |
| (3) | 60 | 4.54 | 65.0 | 90.0 |
| (4) | 60 | 4.24 | 91.0 | 96.0 |
| (5) | 60 | 4.31 | 98.96 | 96.0 |
| (6) | 60 | 4.32 | 93.41 | 91.0 |
| (7) | 50 | 5.84 | 90.82 | 67.0 |
| (8) | 50 | 8.35 | 90.99 | 90.0 |
| (9) | 40 | 15.55 | 81.60 | 85.0 |

Table 6. Selected Pt for every recognition task

As can be observed in Table 6, tasks (1)-(6) reported acceptable scores. Concerning the telephone numbers (7), better results would have been obtained if only digit recognition had been performed. Concerning the *address* task (8), improvements in sentence understanding and recognition time would be desirable (note that a few more seconds were necessary to generate the synthesised output). Clearly, the scores obtained for the queries (9) were not acceptable as the sentence understanding rate was too low and the time required was excessive for an interactive application. Therefore, it is necessary to set up another strategy for this task.

## 5. Future Work

There are several issues that need to be addressed before the system under development can be set up in the real world. On the one hand, it would be possible to divide the *query* task into subtasks considering the different states of the dialogue and creating the corresponding bigrams. Then, in order to recognise queries the dialogue manager could select a smaller bigram according to the state of the dialogue. This would lead to a reduction in recognition time and an enhancement in speech understanding. Concerning the telephone numbers, it would be preferable to use only single-digit recognition as the sentence understanding rate obtained was too low.

Another research issue is concerned with the use of wordspotting instead of continuous speech recognition for some tasks, such as telephone numbers and post codes. In the near future we plan to set up a wordspotting recogniser and to compare the scores obtained. The results presented in this paper were obtained from sentences recorded in our lab. Clearly, attention must be paid to the performance of the system under noisy conditions. We plan to obtain results from recordings made in a fast food restaurant.

Concerning the TTS synthesis, in our current system architecture the TTS server takes about 9.5 sec. on average to generate the waveforms corresponding to the text responses. If we assume the sentence recognition to be performed takes about 5 sec., this makes a total delay of about 15 sec. for the system to generate its responses, which is too long for an interactive application. To overcome this drawback we plan to embed the TTS C++ source code with the rest of the dialogue system in order to save some seconds.

Concerning the type of interaction, the user cannot speak again until the whole response has been completely synthesised. A more natural interaction would be achieved if the output of the system could be cancelled if the user starts to speak again before the output is finished. Additionally, this would lead to a reduction in the time required for the dialogues.

## 6. Conclusions

In this paper we present a preliminary evaluation of a spoken dialogue system under development for the fast food domain. The system is based on a generic model that combines robust speech understanding and mixed-initiative dialogue control. This model uses semantic frames to conceptually store user interactions, a frame-unification procedure to deal with partial information, and a stack structure to handle initiative control. The evaluation carried out was concerned with recognition time, word accuracy, implicit recovery and sentence understanding measures. We considered nine different recognition tasks and experimentally assigned a pruning threshold to each one, taking these measures into consideration. The results obtained showed that the strategies adopted to handle user queries and telephone numbers were not appropriate as the sentence understanding rates were too low. Moreover, the recognition time for the user queries was excessive for an interactive system. Finally, some possibilities for future work to enhance the performance of the system were mentioned. These were mainly focused on the management of user queries and telephone numbers, word recognition, text-to-speech synthesis and user interaction.

## 7. References


Asoh H., Matsui T., Fry J., Asano F. Hayamizu S. (1999). A Spoken Dialog System for a Mobile Office Robot. Eurospeech ’99, pp. 1139-1142

Black A., Taylor P. (1997). Assigning Phrase Breaks from Part-of-Speech Sequences. Eurospeech ‘97, pp. 601-604

Boros M., Heisterkamp P. (1999). Linguistic Phrase Spotting in a Single Application Spoken Dialogue System. Eurospeech ’99, pp. 1983-1986

Chao H., Xu P., Zhang X., Zhao S., Huang T., Xu B. (1999). LODESTAR: A Mandarin Spoken Dialogue System for Travel Information Retrieval. Eurospeech ’99, pp. 1159-1162

Dahlbäck N., Jönsson A. (1999). Knowledge Sources in Spoken Dialog Systems. Eurospeech ’99, pp. 1523-1526

Gustafson J., Lindberg N., Lundeberg M. (1999). The August Spoken Dialogue System. Eurospeech ’99, pp. 1151-1154

Lavelle C. A., De Calmès M., Pérennou G. (1999). Confirmations Strategies to Improve Correction Rates in Telephonic Inquiry Dialogue System. Eurospeech ’99, pp. 1399-1402

López-Cózar R., Rubio A. J., García P., Segura J. C. (1999). A New Word-Confidence Threshold Technique to Enhance the Performance of Spoken Dialogue Systems. Eurospeech ’99, pp. 1395-1398

Nasr A., Esteve Y., Béchet F., Spriet T., de Mori R. (1999). A Language Model Combining N-grams and Stochastic Finite State Automata. Eurospeech ’99, pp. 2175-2178

Noeth E., Boros M., Haas J., Warnke V., Gallwitz F. (1999). A Hybrid Approach to Spoken Dialogue Understanding: Prosody, Statistics and Partial Parsing. Eurospeech ’99, pp. 2019-2022

Pfitzinger Hartmut R. (1998). The Collection of Spoken Language Resources in Car Environment. First International Conference on Language Resources and Evaluation, pp. 1097-1100

Relaño Gil J., Tapias D., Villar J. M., Gancedo M. C., Hernández L. A. (1999). Flexible Mixed-Initiative Dialogue for Telephone Services. Eurospeech ’99, pp. 1179-1182

Rosset S., Bennacef S., Lamel L. (1999). Design Strategies for Spoken Language Dialog Systems. Eurospeech ’99, pp. 1535-1538

Rubio A.J., García P., De la Torre A., Segura J.C., Díaz-Verdejo J.E., Benítez M.C., Sánchez V., Peinado A.M., López-Soler J.M., Pérez-Córdoba J.L. (1997). STACC: An Automatic Service for Information Access Using Continuous Speech Recognition Through Telephone Line. Eurospeech ’97, pp. 1779-1782

Schadle I., Antoine J. Y., Memmi D. (1999). Connectionist Language Models for Speech Understanding: The Problem of Word Order Variation. Eurospeech ’99, pp. 2035-2038

Ziegenhain U., Harengel S., Kaiser J., Wilhem R. (1998). Creating Large Pronunciation Lexica for Speech Applications. First International Conference on Language Resources and Evaluation, pp. 1039-1043