Acquisition and Evaluation of a Dialog Corpus through WOz and Dialog Simulation Techniques

David Griol, Lluís F. Hurtado, Encarna Segarra, Emilio Sanchis

Department de Sistemes Informàtics i Computació
Universitat Politècnica de València, E-46022 València, Spain
{dgriol,lhurtado,esegarra,esanchis}@dsic.upv.es

Abstract

In this paper, we present a comparison between two corpora acquired by means of two different techniques. The first corpus was acquired by means of the Wizard of Oz technique. A dialog simulation technique has been developed for the acquisition of the second corpus. A random selection of the user and system turns has been used, defining stop conditions for automatically deciding if the simulated dialog is successful or not. We use several evaluation measures proposed in previous research to compare between our two acquired corpora, and then discuss the similarities and differences between the two corpora with regard to these measures.

1. Introduction

Learning corpus-based approaches to model the different modules that compose a dialog system has reached a growing interest during the last decade (Minker et al., 1999), (Young, 2002), (Esteve et al., 2003), (He and Young, 2003), (Torres et al., 2005), (Georgila et al., 2006), (Williams and Young, 2007).

A considerable effort is necessary to acquire and label a corpus with the data necessary to train good models. Different techniques have been developed to carry out the acquisition process. The Wizard of Oz technique (WOz), in which a person simulates the behavior of the system, is a well-known approach for acquiring a dialog corpus. In this paper, we also propose an approach to acquire a labeled dialog corpus from the interaction of a user simulator and a dialog manager. In this approach, a random selection of the system and user answers is used. The only parameters that are needed for the acquisition are the definition of the semantics of the task (that is, the set of possible user and system turns), and a set of conditions to automatically discard unsuccessful dialogs.

Different studies have been carried out to compare corpora acquired by means of different techniques and to define the most suitable measures to carry out this evaluation (Schatzmann et al., 2005), (Turunen et al., 2006), (Ai et al., 2007a), (Ai and Litman, 2006), (Ai and Litman, 2007), (Ai et al., 2007b).

In this work, we have applied the WOz and dialog simulation techniques to acquire and compare two corpora within the domain of a Spanish project called DIHANA (Benedí et al., 2006). The task that we considered is the telephone access to information about train timetables and prices in Spanish.

2. Definition of the semantics of the task

As in many other dialog systems, the representation of the user and system turns is done in terms of dialog acts. We defined a set of dialog acts in order to describe the user turns and another set of dialog acts for the system turns. In particular, the semantic representation chosen for the DIHANA task is based on the concept of frame (Minsky, 1975). Frames are a way of representing semantic knowledge. A frame is a structure for representing a concept or situation. Each concept in a domain has usually associated a group of attributes (slots) and values (Fikes and Kehler, 1985).

2.1. User dialog acts

In the semantic representation defined for DIHANA, one or more concepts represent the intention of the utterance, and a sequence of attribute-value pairs contains the information about the values given by the user. Therefore, the natural language understanding (NLU) module takes the sentence supplied by the automatic speech recognizer (ASR) as input and generates one or more frames as output. We defined eight concepts and ten attributes for the DIHANA task. There are two kinds of concepts: Task-dependent concepts and task-independent concepts. A total of ten attributes were specified. Table 1 shows the concepts and attributes defined for the DIHANA task.

Task-dependent concepts
Hour, *Price*, *Train-Type*, *Trip-Time* and *Services*
Task-independent concepts
Affirmation, *Negation* and *Not-Understood*
Attributes
Origin, *Destination*, *Class*, *Departure-Date*, *Arrival-Date*, *Departure-Hour*, *Arrival-Hour*, *Train-Type*, *Services*, and *Order-Number*

Table 1: Concepts and attributes defined for the representation of the user turns in the DIHANA task

An example of the semantic interpretation of an input sentence is shown below:

Yes, I would like to know the price of the Talgo train from Madrid to Barcelona.
(Affirmation)
(Price)
Origin: Madrid
Destination: Barcelona
Train-Type: Talgo

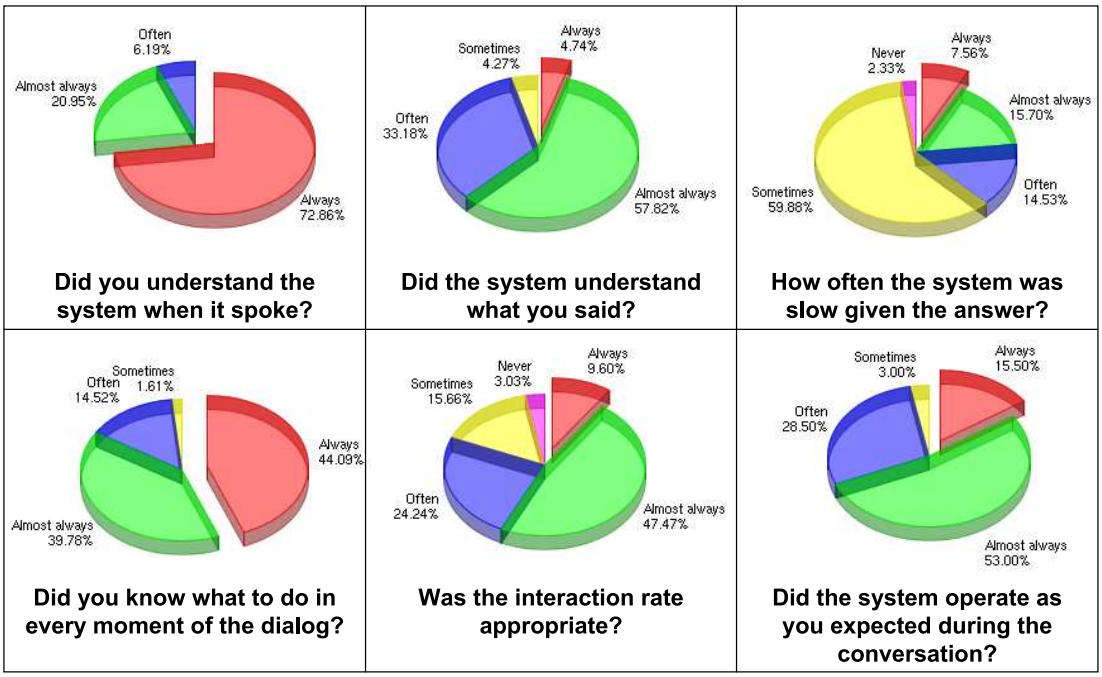

Figure 1: Results of the statistics provided by the users during the WOz acquisition

2.2. System dialog acts

In order to represent the system turns, three levels of labeling of the dialog acts were defined. The first level describes general acts of any dialog and is independent of the task. The second level represents the concepts and attributes involved in the turn and is specific to the task. The third level represents the values of the attributes given in the turn. Each system turn of the dialogs was labeled with one or more dialog acts. The labels defined for these three levels are shown in Table 2.

First level labeling
Opening, *Closing*, *Undefined*, *Not-Understood*, *Waiting*, *New-Query*, *Acceptance*, *Rejection*, *Question*, *Confirmation* and *Answer*
Second and third levels labeling
Origin, *Destination*, *Date*, *Departure-Hour*, *Arrival-Hour*, *Price*, *Train-Type*, *Services*, *Order-Number Number-Trains*, *Class*, *Trip-Type*, *Trip-Time* and *Nil*

Table 2: Set of labels defined for the representation of the system answers in the DIHANA task

Two examples of the dialog act labeling of the system turns are shown below:

Welcome to the railway information system. How can I help you?
(Opening:Nil:Nil)
Do you want to go to Valencia?
(Confirmation:Origin:Valencia)

3. Acquisition of the dialog corpus

As stated in the introduction, two different techniques have been used to acquire two dialog corpora for the DIHANA project. A set of 300 scenarios was defined to carry out the acquisition. These scenarios can be classified into four categories depending on the number of objectives and the default information that they provided. Type S1 and S2 defined only one objective for the dialog and Type S3 and S4 defined two objectives. Type S2 and S4 incorporate into their definition attributes not specified as mandatory for the objective of the dialog. The objective of both acquisition processes was the acquisition of 900 dialogs (three for each scenario).

3.1. Using the Wizard of Oz technique

For the acquisition process using the WOz technique, 225 volunteers were recruited, each of them acquiring four scenarios. The acquisition process resulted in a spontaneous Spanish speech dialog corpus with 225 different speakers (153 male and 72 female), with small dialectal variants. The total number of user turns was 6280, with an average of 7.7 words per user turn. The vocabulary size was 823 words. The total amount of speech signal was about 10.8 hours. This corpus has been recently used in order to develop new stochastic strategies for the dialog management, as well as for the design of the models of the main modules in the DIHANA dialog system. More information about the acquisition process can be found in (Benedí et al., 2006).
Each volunteer was requested to give their opinion about the system operation and their feelings after using it. The data collection was made using surveys that the users complete after their interaction with the system. Figure 1 shows the results of the evaluation of the surveys. The questions included in the survey were:

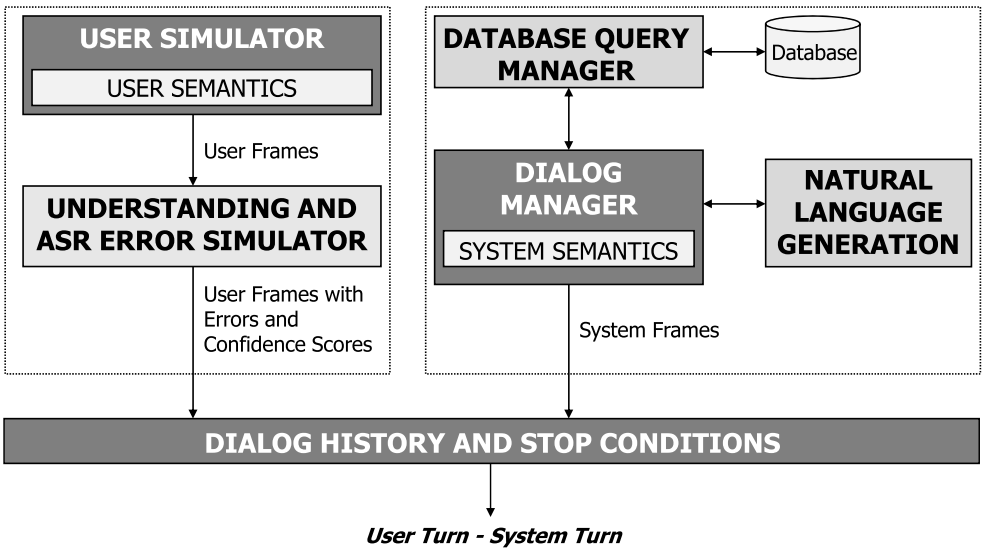

Figure 2: Graphical scheme of the architecture defined for the automatic dialog generation

- Did you understand the system when it spoke?
- Did the system understand what you said?
- Was the interaction rate appropriate?
- Did you know what to do in every moment of the dialog?
- How often the system was slow given the answer?
- Did the system operate as you expected during the conversation?

Figure 3 shows an example of a dialog extracted from the WOz corpus. Turns with S refer to system turns, and turns with U refer to user turns.

S1: Welcome to the railway information system. How can I help you?
U1: I want to know timetables from Valencia to barcelona for next Monday.
S2: Do you want timetables leaving from Madrid?
U2: No, I want to leave from Valencia.
S3: There are only two trains. Anything else?
U3: The price of the first train
S4: Do you want to know the price leaving from Valencia?
U4: Yes
S5: The price is 52.5 euros. Anything else?
U6: No, thank you.
S7: Thank you for using the system.

Figure 3: An example of a dialog acquired by means of the WOz technique

3.2. Using the simulation technique

As stated in the introduction, our approach for acquiring a dialog corpus is based on the interaction of a user simulator and a dialog manager simulator. Both modules use a random selection of one of the possible turn defined for the semantic of the task (user and system dialog acts).

The user simulation simulates the user intention level, that is, the simulator provides concepts and attributes values that represent the intention of the user utterance. Therefore, the user simulator carries out the functions of the ASR and NLU modules.

An error simulator module has been designed to perform error generation and the addition of confidence measures. This information modifies the frames generated by the user simulator and also incorporates confidence measures for the different concepts and attributes. The error generation can be modified to adapt the error simulator module to the operation of any ASR and NLU modules.

A maximum number of nine user turns per dialog was defined for acquiring a corpus. A user request for closing the dialog is selected once the system has provided the information defined in the objective(s) of the dialog. The dialogs that fulfill this condition before the maximum number of turns are considered successful. The dialog manager considers that the dialog is unsuccessful and decides to abort it when the following conditions take place:

- The dialog exceeds the maximum number of user turns.
- The answer selected by the dialog manager corresponds with a query not required by the user simulator.
- The database query module provides an error warning because the user simulator has not provided the mandatory information needed to carry out the query.
- The answer generator provides a error warning when the selected answer involves the use of a data not provided by the user simulator.

The semantics selected for the dialog manager is represented through the 51 possible system answers defined for the task. The selection of the possible user answers is carried out using the semantics defined for the user in the NLU module.

A total of two million simulated dialogs were necessary in order to acquire a dialog corpus of 900 successful dialogs following the same set of scenarios defined for the WOz acquisition.

Figure 4 shows an example of a dialog corresponding to a scenario with two objectives (to know timetables and prices from Valencia to Barcelona for May the 30th). The confidence scores generated by the error simulator are shown between brackets. A sentence in natural language, that is equivalent to the labeled system turn selected by the dialog manager, has been incorporated to clarify the explanation.

S1: (Opening:Nil:Nil)
Welcome to the railway information system. How can I help you?
U1: Departure-Hour [0.9]
Origin: Madrid [0.9]
Destination: Barcelona [0.9]
Departure-Date: 05-30-2007 [0.9]
Train-Type: Euromed [0.9]
S2: (Confirmation:Departure-Hour:Nil)
(Confirmation: Origin:Origin)
Do you want timetables leaving from Madrid?
U2: Negation[0.9]
Origin:Valencia [0.9]
S3: (Answer:Departure-Hour:Departure-Hour)
(New-Query:Nil:Nil)
{*Timetables information*} Anything else?
U3: Price [0.1]
Origin: Valencia [[0.9]
Destination: Barcelona [0.9]
Departure-Hour: 08.00-23.00 [0.9]
S4: (Confirmation:Price:Nil)(Confirmation:Origin:Origin)
Do you want to know the price leaving from Valencia?
U4: Affirmation [0.9]
S5: (Confirmation:Date:Date)
U5: Affirmation [0.9]
S6: (Answer:Price:Price) (New-Query:Nil:Nil)
{*Prices information*} Anything else?
U6: Negation
S7: (Closing:Nil:Nil)
Thank you for using the system.

Figure 4: An example of a dialog acquired by means of the simulation technique

In the first turn, the user simulator selects to make a query about timetables and provides the *Origin*, *Destination* and the *Departure-Date*. In addition it provides additional information defined in the scenario (the *Train-Type* attribute). For this turn, the error simulator introduces an error in the *Origin* (it changes Valencia by Madrid) and assigns to this value a high confidence.

In the following system turn (S2), the dialog manager asks the simulated user to consult timetables leaving from Madrid. In the following turn (U2), the user simulator consults the objective of the scenario and provides again the *Origin*. After this turn, the system makes a query about timetables to the database (S3).

The user simulator verifies in the U3 turn that the objective of the dialog has not been completed. In this turn it selects to make a query about prices, providing again the *Origin* and the *Destination*. It also incorporates the *Departure-Hour* as additional information. In the following system turn (S4), the dialog manager makes a confirmation about prices leaving from Valencia. Verified the objective of the dialog, the user simulator selects *Affirmation* (U4). In the following turn, the system selects to confirm the *Departure-Date*. The user simulator confirms this information according to the objective of the dialog (U5). Then, the system selects to carry out a database query about prices (S6). As the necessary information is available, the database query module carries out the query and the dialog manager completes the objectives for the dialog. Having this information, the user simulator selects a request for closing the dialog in the following turn (U6).

4. Evaluation of both corpora

We used a set of measures to carry out the evaluation of the acquired corpora based on prior work in the dialog literature. (Schatzmann et al., 2005) proposed a comprehensive set of quantitative evaluation measures to compare two dialog corpora. These measures were adapted for use in our comparisons, based on the information available in our corpora. This set of proposed measures are divided into three types:

- High-level dialog features: These features evaluate how long the dialogues last, how much information is transmitted in individual turns, and how active the dialogue participants are.
- Dialog style/cooperativeness measures: These measures try to analyze the frequency of different speech acts and study what proportion of actions is goal-directed, what part is taken up by dialogue formalities, etc.
- Task success/efficiency measures: These measures study the goal achievement rates and goal completion times.

We defined six high-level dialog features for the evaluation of the dialogs: the average number of turns per dialog, the percentage of different dialogs withouth considering the attribute values, the number of repetitions of the most seen dialog, the number of turns of the most seen dialog, the number of turns of the shortest dialog, and the number of turns of the longest dialog. Using these measures, we tried to evaluate the success of the simulated dialogs as well as its efficiency and variability with regard to the different objectives.

For dialog style features, we define and count a set of system/user dialog acts. On the system side, we have measured the confirmation of concepts and attributes, questions to require information, and system answers generated after a database query. We have not take into account the opening and closing system turns. On the user side, we have measured the percentage of turns in which the user carries out a request to the system, provide information, confirms a concept or attribute, Yes/No answers, and other answers not included in the previous categories.

WOz Corpus	Type S1	Type S2	Type S3	Type S4
Average number of user turns per dialog	4.9	5.2	5.8	6.4
Percentage of different dialogs	99.2%	99.2%	99.5%	99.5%
Number of repetitions of the most seen dialog	2	3	2	2
Number of turns of the most seen dialog	7	9	9	9
Number of turns of the shortest dialog	5	5	7	7
Number of turns of the longest dialog	21	25	25	27
Simulated Corpus	**Type S1**	**Type S2**	**Type S3**	**Type S4**
Average number of user turns per dialog	4.0	4.4	5.7	5.9
Percentage of different dialogs	91.3%	94.4%	100%	100%
Number of repetitions of the most seen dialog	5	4	1	1
Number of turns of the most seen dialog	5	7	7	11
Number of turns of the shortest dialog	5	5	7	7
Number of turns of the longest dialog	17	17	17	19

Table 3: Results of the high-level dialog features defined for the comparison of both corpora

We have not considered task success/efficiency measures in our evaluation, since only the dialogs that fulfill the objectives predefined in the scenarios have been incorporated into our corpora. We have considered successful dialogs those that fulfill the complete list of objectives defined in the corresponding scenario.
Figure 5 summarizes the complete set of measures used in the evaluation.

High-level dialog features
Average number of turns per dialog
Percentage of different dialogs
Number of repetitions of the most seen dialog
Number of turns of the most seen dialog
Number of turns of the shortest dialog
Number of turns of the longest dialog
Dialog style/cooperativeness measures
System dialog acts: Confirmation of concepts and attributes, Questions to require information, and Answers generated after a database query.
User dialog acts: Request to the system, Provide information, Confirmation, Yes/No answers, and Other answers.

Figure 5: Evaluation measures used to compare the acquired corpora

Table 3 shows the results of the comparison of the high-level dialog features. It can be seen that all measures have similar values in both corpora. The more significant difference is the average number of user turns. In the four types of scenarios, the dialogs acquired using the simulation technique are shorter than those acquired using the WOz technique. This can be explained because of the fact that there is a set of dialogs acquired using the WOz technique in which the user asked for additional information not included in the definition of the corresponding scenario once the dialog objectives had been achieved.
Finally, we have compared the percentage of the most significant types of dialog acts in both corpora (confirmations of concepts and attributes, questions to require information to the user, and answers obtained after a query to the database). Table 4 shows the results of this comparison for the system dialog acts. It can be observed that there are also only slightly differences between the values obtained for both corpora. There is a higher percentage of confirmations and questions in the WOz corpus due the higher average number of turns per dialog in this corpus. Table 5 shows the results of this comparison for the user dialog acts. The most significant difference between both corpora is the percentage of turns in which the user makes a request to the system. The percentage of these kind answers is lower in the WOz corpus. This can be explained by the fact that it is less probable that simulated users provide useless information, as it is shown in the lower percentage of the users turns classified as Other answers.

5. Conclusions

In this paper, we have presented a comparison between two corpora acquired using two different techniques. First, we acquired a dialog corpus using the Wizard of Oz technique. Second, we have developed a dialog simulation technique based on the random selection of the user and system answers. A set of stop conditions were defined to decide automatically if the dialog has to be considered successful. A set of measures have been used to compare both corpora. The results of this comparison show that the two corpora have similar characteristics.
We are currently adapting the dialog simulator technique to acquire a corpus within the framework of a new project called EDECAN. One of the main objectives of the EDECAN project is to develop a dialog system for booking sports facilities in our university. Users can ask for the availability, the booking or cancellation of a facility, and the information about his/her current bookings. Using our simulation approach, we want to acquire a corpus that makes the learning of a dialog manager possible for the domain of the EDECAN project. This dialog manager will be used in a supervised acquisition of a dialog corpus with real users.

6. References

H. Ai and D. Litman. 2006. Comparing Real-Real, Simulated-Simulated, and Simulated-Real Spoken Dialogue Corpora. In *Procs. of AAAI Workshop Statistical and Empirical Approaches for Spoken Dialogue Systems*, Boston, USA.

	WOz Corpus	Simulated Corpus
Confirmation of concepts and attributes	49.3%	46.8%
Questions to require information	13.9%	11.6%
Answers generated after a database query	36.8%	41.6%

Table 4: Percentages of the different types of system dialog acts in both corpora

	WOz Corpus	Simulated Corpus
Request to the system	19.9%	25.4%
Provide information	26.8%	25.5%
Confirmation	11.6%	9.4%
Yes/No answers	31.6%	31.7%
Other answers	10.1%	7.9%

Table 5: Percentages of the different types of user dialog acts in both corpora

H. Ai and D.J. Litman. 2007. Knowledge Consistent User Simulations for Dialog Systems. In *Proc. of Interspeech'07,*, Antwerp, Belgium.

H. Ai, J.R. Tetreault, and D.J. Litman. 2007a. Comparing User Simulation Models For Dialog Strategy Learning. In *Proc. of NAACL HLT'07,*, pages 1–4, Rochester, NY, USA.

Hua Ai, Antoine Raux, Dan Bohus, Maxine Eskenazi, and Diane Litman. 2007b. Comparing Spoken Dialog Corpora Collected with Recruited Subjects versus Real Users. In *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 124–131, Antwerp, Belgium.

J.M. Benedí, E. Lleida, A. Varona, M.J. Castro, I. Galiano, R. Justo, I. López, and A. Miguel. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 1636–1639, Genoa, Italy.

Y. Esteve, C. Raymond, F. Bechet, and R. De Mori. 2003. Conceptual Decoding for Spoken Dialog systems. In *Proc. of European Conference on Speech Communications and Technology (EuroSpeech'03)*, pages 617–620, Geneva (Switzerland).

R.E. Fikes and T. Kehler. 1985. The role of frame-based representation in knowledge representation and reasoning. In *Communications of the ACM*, volume 28, pages 904–920.

K. Georgila, J. Henderson, and O. Lemon. 2006. User Simulation for Spoken Dialogue Systems: Learning and Evaluation. In *Proc. of the 9th International Conference on Spoken Language Processing (Interspeech/ICSLP)*, pages 1065–1068, Pittsburgh (USA).

Y. He and S. Young. 2003. A data-driven spoken language understanding system. In *Proc. of ASRU'03*, pages 583–588.

W. Minker, A. Waibel, and J. Mariani. 1999. Stocastically-based semantic analysis. In *Kluwer Academic Publishers*, Boston, USA.

M. Minsky, 1975. *The Psychology of Computer Vision*, chapter A Framework for Representing Knowledge, pages 211–277. McGraw-Hill.

J. Schatzmann, K. Georgila, and S. Young. 2005. Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems. In *Proc. of SIGDial Workshop'05*, pages 45–54, Lisbon (Portugal).

F. Torres, L.F. Hurtado, F. García, E. Sanchis, and E. Segarra. 2005. Error handling in a stochastic dialog system through confidence measures. In *Speech Communication*, pages (45):211–229.

Markku Turunen, Jaakko Hakulinen, and Anssi Kainulainen. 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)*, pages 1057–1060, Pittsburgh, USA.

J. Williams and S. Young. 2007. Partially Observable Markov Decision Processes for Spoken Dialog Systems. In *Computer Speech and Language 21(2)*, pages 393–422.

S. Young. 2002. The Statistical Approach to the Design of Spoken Dialogue Systems. Technical report, Cambridge University Engineering Department.