4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

DEVELOPMENT OF SPOKEN LANGUAGE CORPORA FOR TRAVEL INFORMATION*

L. Lamel, S. Rosset, S. Bennacef, H. Bonneau-Maynard, L. Devillers, J.L. Gauvain
LIMSI - CNRS, B.P. 133, 91403 Orsay, France

10.21437/Eurospeech.1995-478

ABSTRACT

In this paper we report on our ongoing work in developing spoken language corpora in the context of information access in two travel domain tasks, L'ATIS and MASK. The collection of spoken language corpora remains an important research area and represents a significant portion of work in the development of spoken language systems. The use of additional acoustic and language model training data has been shown to almost systematically improve performance in continuous speech recognition. Similarly, progress in spoken language understanding is closely linked to the availability of spoken language corpora. We record subjects on a regular basis using development versions of the spoken language systems for both tasks, obtaining over 1000 queries/month from 20 subjects. To help assess our progress in system development, each subject since March'95 completes a questionnaire addressing the user-friendliness, reliability, ease-of-use of the MASK data collection system.

INTRODUCTION

The collection of spoken language corpora is an important research area and represents a significant portion of the work in developing a spoken language system. The use of additional acoustic and language model training data has been shown to almost systematically improve performance in continuous speech recognition[3]. Progress in spoken language understanding is also closely linked to the availability of spoken language corpora.

In this paper we report on our ongoing work in developing spoken language corpora for access to travel information for two travel domain tasks, an air travel information task (ATIS) and a train travel information task (MASK). L'ATIS is a French version of the ARPA ATIS task which has been used as a common task for data collection and evaluation[5] within the ARPA Speech and Natural Language program. L'ATIS allows users to acquire information about fares and flights available between a restricted set of cities within the United States and Canada as well as some ancillary information. In the MASK task users can ask for rail travel information such as timetables, tickets and reservations for train travel among 500 cities in France. Data collection for this task is carried out in the context of the ESPRIT project MASK (Multimodal-Multimedia Automated Service Kiosk) in which we are developing a spoken language system interface for an automated service kiosk. The aim of the MASK project is to enhance the effectiveness of automated public service provision by enabling interaction through the coordinated use of multi-modal inputs (speech and touch) and multi-media output (sound, video, text, and graphics).

DATA COLLECTION SYSTEM

An overview of the spoken language system for information retrieval is shown in Figure 1. The main components are the speech recognizer, the natural language component which includes a semantic analyzer and a dialog manager, and an information retrieval component that includes database access and response generation. While our goal is to develop underlying technology that is speaker, task and language independent, any spoken language system will necessarily have some dependence of the chosen task and on the languages known to the system. The spoken query is decoded by a speaker independent, continuous speech recognizer, whose output is then passed to the natural language component. In our current implementation the output of the speech recognizer is the best word sequence, however, the recognizer is also able to provide a word lattice. The semantic analyzer carries out a caseframe analysis to determine the meaning of the query, and builds an appropriate semantic frame representation[2]. Keywords are used to select an appropriate case structure for the sentence, which is then used to construct an appropriate semantic frame representation. The dialog history and default values generated from task knowledge are used to complete missing information in the semantic frame. The cases specifying the case-frame grammar, and the trigger keywords are also to a large part task- (or at least domain) dependent. For the L'ATIS task we currently have concepts corresponding to flight-time, fare, stop, type, and reservation. For MASK we replaced the concepts flight-time with train-time, stop with changes, and added new concepts for reductions and services. The cases associated with the concepts may have different trigger keywords as in the case of class specification for fares. The response generator uses the semantic frame to generate a request to the database management system, and presents the result of the database query and an accompanying natural language response to the user. A vocal response is optionally provided.

DATA COLLECTION SCENARIOS

For L'ATIS we currently use a set of 9 scenarios, with different complexities. Each subject solves about 7 different scenarios. Two example scenarios are given in Figure 2(a). The first scenario is relatively simple, and is usually the first one solved by the subject. At the start of this scenario subjects typically formulate queries close to what is written. A typical start of the interaction with the system is:

U: "I want to go from Boston to San Francisco monday morning."
S: <List of flights from Boston to San Francisco>

*This work is partially funded by the ESPRIT project 9075 MASK.

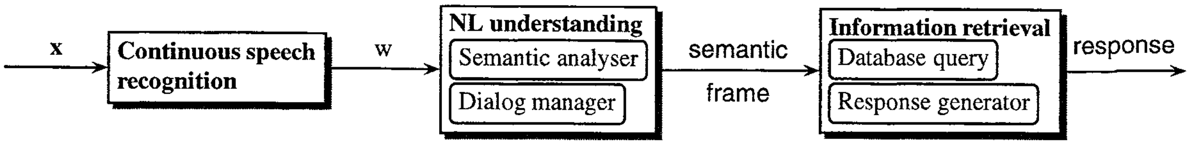

Figure 1: Overview of the spoken language information retrieval system. **x** input speech, **w** word sequence output by speech recognizer.

Month	before Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
#speakers	20	36	80	98	117	146	163	210	242	247
#queries	1115	1907	3802	4376	5629	7731	8611	10255	11728	12007
total #words	12.1k	18.8k	39.6k	46.2k	52.8k	71.7k	80.6k	96.3k	108k	110.7k
#words no frag.	12.1k	18.7k	39.6	46.1k	52.7k	71.5k	80.4k	95.9k	108.3K	110.3k
#distinct words	562	673	985	1058	1058	1374	1548	1742	1852	1864
#distinct no frag.	558	658	938	998	998	1252	1403	1561	1649	1657
#new words	-	100	280	60	0	254	151	158	88	8

Table 1: Status of data collection – L’ATIS

(a) You want to travel from Boston to San Francisco Monday morning. You would like to arrive before 10:30 am. Reserve the cheapest flight with a stopover.

You live in Dallas and need to go to Oakland. You can only afford to pay $575, and you do not like to fly with Continental Airlines.

(b) You want to go from Grenoble to Paris next Friday as late as possible. You want to take a direct TGV and to pay a reduced fare. Reserve a non-smoking place.

You are traveling from Bordeaux to Avignon next Sunday. You have a reduction **Carissimo.** Your dog is traveling with you. Reserve an aisle seat in a second class, smoking car. Will you need to change trains?

Figure 2: Example scenarios used for data collection: (a) L’ATIS scenarios, (b) MASK scenarios.

U: “I would like to arrive before 10:30 am.”

S: There are no flights from Boston to San Francisco arriving before 10:30am.

Given the unavailability of flights arriving early in the morning, subjects then proceed to ask about the earliest flight, or about leaving the evening before. The second example scenario is presented less precisely, so as to invoke a wider variety of query formulations.

Similarly we have designated a set of 10 MASK scenarios with a range of complexities. Two example scenarios are given in Figure 2(b). The scenarios for both tasks are periodically modified to ellicit a wider variety of vocabulary items, such as city names, dates and times of travel. We also include specific scenarios in which users need to find out information about concepts not yet handled by the system, to see how they react in order to help us develop ways to detect such situations and to guide the user accordingly.

DATA COLLECTION STATUS

We started our data collection for the L’ATIS task using WOZ setup, where a wizard typed a paraphrased version of the spoken query to the system.[1] 20 subjects solved task-specific scenarios translated from English. Each session lasted about 50 minutes, during which the subject produced on average 50 queries, for a total of 1115 queries. We have since greatly expanded our data collection effort, and are recording subjects on a regular basis using up-to-date versions of our spoken language systems for L’ATIS and MASK. The recordings are made in office environement, simultaneously with a close-talking, noise cancelling Shure SM10 and a table-top Crown PCC160 microphone. We collect speech data at the rate of over 1000 queries per month from at least 20 speakers.

The cumulative number of subjects and number of queries recorded thus far for L’ATIS are shown in Table 1. There are an average 10 words per query. The total number of words, and the number of distinct words are shown, with and without the inclusion of word fragments. There are about 200 different word fragments, accounting for less than 0.5% of all word occurrences. The new word rate is seen to decrease from Sep. through Dec. and then to increase again in Jan. In mid December a new version of the L’ATIS data collection system was installed. This version corrected some problems in the maintenance of the dialog history which had caused the subjects to repeat several times the same query, and integrated a new version of the speech recognizer. The combined improvements changed significantly the user’s interaction with the system. With the more performant speech recognizer, subjects speak more easily and use longer and more varied sentences. This also leads to the occurrence of more new words in the queries. They are also more likely to perceive that the recognition errors are their fault, rather than the system’s. As a result they continue to speak relatively naturally to the system, enabling us to record more representative spontaneous speech. These new recordings will then be used to improve the system, leading once again to a more flexible, performant system.

For MASK we have recorded 113 speakers for a total of 6853 queries. The cumulative number of subjects and queries recorded are shown in Table 2. The average sentence length is 8 words, slightly shorter than for L’ATIS. The shorter sentences are probably linked to the performance of the speech recognizer which for now has a higher word error than for L’ATIS due to the limited amount of training data. This difference is also due to the need for the user to provide

[1] For this initial setup, the NL understanding component was developed in collaboration with colleagues at MIT[1].

Month	Jan	Feb	Mar	Apr	May
#speakers	12	42	78	106	113
#queries	208	1603	3825	6219	6853
total #words	1.6k	12.3k	29.1k	44.5k	48.6k
#words no frag.	1.6k	12.2k	29.0k	44.4k	48.5k
#distinct words	273	737	975	1120	1168
#distinct no frag.	271	691	902	1015	1049
#new words	-	420	211	113	34

Table 2: Status of data collection – Mask

	Classification (%)				
Corpus	A	D	P	XT	X
L'Atis	47	48	4	-	1
Mask	13	67	<1	19	<1

Table 3: Classification of query types.

additional information in order to make a train reservation such as smoking or non-smoking, and window or aisle seat. There are 1049 distinct words (excluding fragments) in the Mask queries, with only 373 not in the L'Atis word list.

The most frequent words in both corpora are je, le, de, à, voudrais, heures, vol/train. Je alone accounts for 6% of all word occurrences. The relative frequencies of the words in the corpora are shown in Figure 3. 10% of the words account for almost 90% and 70% of all word occurrences in the L'Atis and Mask corpora respectively. For L'Atis a lexicon with 30% (about 600) of the observed words essentially covers all of the data. For Mask about 50% (500) of the observed words gives about the same coverage. It should be noted that for both systems, the number of city names is rather small, with 10 cities for L'Atis and 35 cities/stations for Mask.[2]

In order to obtain the maximum benefit of the corpus, we must have a good statistical understanding of it. Each query is transcribed and classified as "answerable without context" (A), "answerable given context" (D), "politeness forms" (P), "out of domain" (X), and "temporarily out of domain" (XT). This latter category refers to queries which are not treated in the version of the system used to collect the data, but will be treated in future versions. The classification of query types is summarized in Table 3. Politeness forms occur within sentences more frequently for L'Atis than for Mask: please (4.2% vs 1.5%), hello (2.9% vs 1.5%), thank you (2.2% vs 0.5%). Other interjections such as then, well, and okay occur in about 3% of the utterances. Spontaneous speech phenomena such as hesitations, false starts and reparations occur in about 25% of the queries. The filler word euh occurs 423 times (9.4% of the queries) in the Mask data and 1849 times (15% of the queries) in the L'Atis data. Breath noises (inspiration and expiration) were marked in about 11% of the transcriptions.

The data collection system uses a mixed-initiative dialog strategy, where the user is free to ask any question, at any time. However, in order to aid the user to obtain a reservation, the system prompts the user for any missing information needed for database access. The average number of queries to solve a scenario is 9 for L'Atis and 14 Mask. The longer Mask sessions are in fact linked to the need to specify more information in order to solve the scenario. This also has the consequence that the system more often takes the initiative asking the user to provide specific information. As shown in Table 4, 33% of the system responses for Mask involve requests for additional information as compared to only 12% for L'Atis. For L'Atis these direct requests involve the class (31%), departure city (21%), date (16%), time (16%), and arrival city (14%). For Mask these direct requests involve the class (24%), date (25%), time (14%), smoking (10%), departure city (20%) and arrival city (7%). Due to this dialog strategy, the length of the first query of each scenario is twice as long as the average length of the remaining queries.

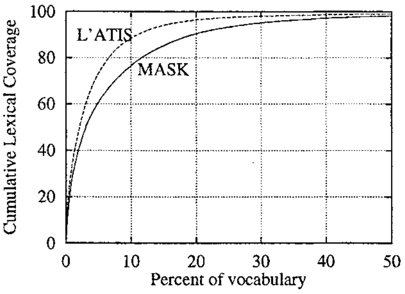

Figure 3: Percentage of transcription covered as a function of the percentage of words.

Corpus	#Scenarios	#Queries	Precisions	Null
L'Atis	1310	12007	12%	4%
Mask	461	6853	33%	7%

Table 4: Analysis of dialogs.

We analysed about 15% of the dialogs to see how subjects respond to the system's initiatives. Subjects provided a direct response to these requests for precision about 50% of the time for both tasks. For L'Atis half of the responses were complete sentences, compared to one-third for Mask. In 30% of the L'Atis user responses, additional information was also given. For example, when asked the date of travel, in 40% of the responses the user also provided time information, such as "next Monday, after 5 p.m.". For Mask additional information was given in less often, in only 12% of the responses, most often for date of travel where the time was also given (20%) and for class where the smoking zone preference was given (12%). For L'Atis 5% of the user's responses to direct questions were requests for clarification and 10% were on a totally different concept. In the latter case, the subject was clearly following their train of thought, ignoring output of the system. When the system asked for a repetition, 25% were due to recognition errors, 5% to understanding errors, and 70% because the user did not provide the requested information. For Mask half of the requested repetitions were due to system errors (recognition or understanding), and the remaining because the user ignored the system's request for precision. The higher percentage of precision requests in Mask is also partly because the system performs less well than the L'Atis system (see

[2] The Mask kiosk will use 500 cities.

Corpus	#Sents	WAcc	NL	SLS
L'ATIS Oct94	225	90.2%	85%	84%
L'ATIS Jan95	225	93.7%	89%	88%
MASK Jan95	205	78.0%	'85%	60%
MASK Apr95	205	85.1%	93%	79%

Table 5: Evaluation results for L'ATIS and MASK.

Table 5).

EVALUATION

The development of a spoken language system is incremental, where errors are analysed and the system is further refined. Periodic evaluation allows us to continually monitor progress through objective performance measures. The speech recognizer is evaluated in terms of speed and recognition accuracy (word and sentence error). An analysis of the recognition errors is carried out to determine their effect on the understanding performance. The understanding component is evaluated using typed versions of the exact transcriptions of spoken queries including all spontaneous speech effects, such as hesitations or repetitions, (so as to evaluate the understanding component without intrusion of errors made by the speech recognizer) and on the recognized word string. In order to evaluate the understanding component we have developed a semi-automatic method which makes use of reference semantic frames for each test query. In Table 5 evaluation results of the data collection system is given for 225 L'ATIS queries and 205 MASK queries from 10 speakers. On L'ATIS, the speech recognition word accuracy improved from 90.2% in Oct94 to 93.7% in Jan95. This improvement is directly linked to the availability of additional training data, particularly for the LM. The word accuracy is lower for MASK but has improved from 78% in Jan95 to 85% in Apr95. For the Jan95 L'ATIS system we observe that there is only a slight degradation in performance for the complete spoken language system relative to the NL component. Since the performance of the speech recognizer for MASK has not yet reached the same level as that for L'ATIS, a larger degradation occurs.

The dialog process formally consists of transitions between 5 different dialog states: opening formalities, information, stagnation, confirmation sub-dialogues and closing formalities. The evaluation of the dialog focuses on the two middle states and includes quantitative measures such as (1) the number of speech turns generated by the user and by the system to complete the task, (2) the number of system requests for precisions, (3) the number of times the user reformulates the same query, (4) the number of dialog errors, and (5) the number of understanding errors.

It is also important to assess the overall performance of the system from the point of view of the subjects. Since March'95 all subjects have completed a questionnaire (Figure 4) addressing the user-friendliness, reliability, ease-of-use of the MASK system. Subjects are also asked what are the good aspects of the system, how it should be improved, and if they would use such a potential system. Information about the subject includes how often they travel by train, how they obtain their tickets, and their computer experience.

1. Is it easy to speak to the system?
2. Is the system easy to understand?
3. Does the system respond fast enough?
4. Are you confident in the information given by the system?
5. Did you get the information you wanted?
6. Are you satisfied with the information?
7. Did the system recognize what you said?
8. Did the system understand what you said?
9. If the system did not understand you, was it easy to reformulate your question?

Figure 4: User questionnaire

Experience	Ease-of-use	Reliability	Friendliness
Questions	1 2 3	4 5 6	7 8 9
Expert (16)	7.7	7.6	7.7
Novice (9)	7.2	6.4	6.7

Table 6: User responses to questionnaire.

The responses of 25 speakers (independent of their age which ranged from 18-60 years, and of their travel habits) were analysed. The results are shown in Table 6 an a scale of 10. Users were classed as novices if they had difficulties speaking with the system or using the computer. In general "expert" users (no difficulty speaking with the system and used to working with computers) were more at ease with the system, and judged it to be more user-friendly, easier to use, and more reliable than the novices. The novices were more likely to critique the reliability of information obtained from the system, whereas the experts criticized problems in understanding or dialog. Overall users express an interest in using such types of systems, and often ask to come back to participate in future experiments.

SUMMARY

In this paper we have reported on our ongoing data collection efforts using development versions of spoken language systems for travel domain tasks L'ATIS and MASK. We record over 1000 queries a month from over 20 subjects. This data is analyzed and used to improve the system. We evaluate the performance of the individual system components as well as the overall system. We are currently experimenting with different dialog strategies and with the presentation of the returned information. As more data is recorded we expect to be able to improve the performance of the data collection systems, which in turn should enable us to record spontaneous speech representative of how users interacts with a fully automated system.

REFERENCES

[1] H. Bonneau-Maynard et al., "A French Version of the MIT-ATIS System: Portability Issues," Eurospeech-93.

[2] S. Bennacef et al., "A Spoken Language System For Information Retrieval," ICSLP-94.

[3] J.L. Gauvain et al., "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task," ICASSP-94.

[4] L. Lamel et al., "Recent Developments in Spoken Language Sytems for Information Retrieval," ESCA Workshop on Spoken Dialog Systems, Vigsø, Denmark, Spring 1995.

[5] P. Price, "Evaluation of Spoken Language Systems: The ATIS Domain," DARPA Workshop on Speech & NL, 1990.