A French Version of the MIT-ATIS System: Portability Issues [1]

H. Bonneau-Maynard†, J.L. Gauvain†, D. Goodine‡, L.F. Lamel†, J. Polifroni‡, and S. Seneff‡[2]

†Speech Communication Group
LIMSI - CNRS, B.P. 133
91403 Orsay, France

‡Spoken Language Systems Group
Laboratory for Computer Science
M.I.T., Cambridge, MA 02139 USA

ABSTRACT

This paper presents our recent research in developing L'ATIS, a French version of the MIT Air Travel Information Service (ATIS) system used to interrogate a database derived from the Official Airline Guide (OAG). We have adopted an approach of accessing the English based system at the level of the semantic frame, so as to produce a language-independent meaning representation. Thus the same core back-end component as in the English-based version can be used, with only the input and output modules replaced by French versions. In addition, the input module uses the same mechanism to convert an input sentence to a semantic frame, with the English grammar rules and constraints being replaced by corresponding French versions. A common approach for language generation is also used for both systems. Once a core system in French was operational, data were collected in the form of typed queries so as to expand the rules and vocabulary, as well as spoken queries using a wizard-of-Oz (WOZ) setup. Preliminary speech recognition error rates as well as an informal analysis of the performance of the NL component are provided.

INTRODUCTION

Spoken language systems are emerging as an important research area to enable speech recognizers to operate as a natural interface between the user and the computer. Through simple and natural dialogues, users can gain access to a rich body of useful data. Concurrently, computers are becoming much faster and more economical, and there is a rapid explosion of information services becoming available over networks and phone lines, (MINITEL is a prime example), such that a large number of potential applications for speech are reaching the point of marketability.

Developing any particular spoken language system is a time-consuming process. The system components to be developed include the speech recognizer, the natural language component, a discourse and dialogue model, as well as the database and tools for database access. As spoken language systems emerge, it will become increasingly important to address the issues of portability. It would be reassuring if porting in some new direction would involve less time and effort than the initial system development. In designing spoken language systems, one aim has been to choose methods that would make it relatively easy to separate out those parts of the system that are domain/language independent from those that are not, so that porting can be made more efficient. Our focus in this paper is on issues relating to porting from one natural language to another.

MIT began exploring multilingual systems within the context of their VOYAGER interactive spoken language system. [3]. Over the past two years, Japanese has been added as an input/ouput language for VOYAGER, where the user can freely mix and match Japanese or English as the input or output language through a simple switch. Given the success of MIT's efforts with VOYAGER, and the common desire of MIT and LIMSI to work on multilingual spoken language systems, it was decided to initiate a joint enterprise in porting the existing MIT ATIS (Air Travel Information Service) system to French. A prior European interest in an interactive travel-planning domain has been demonstrated in the ESPRIT project, SUNDIAL [6].

ATIS [7] has been designated as a common task for data collection and evaluation within the ARPA community. This domain has been under active development at five ARPA sites as well as several non-ARPA participants, over the last three years. As a consequence, the systems are generally capable of handling a rich repertoire of spoken English sentences restricted to the domain. ATIS allows users to acquire information about fares and flights available between a restricted set of cities within the United States and Canada. Some ancillary information, such as the meals served on the flight or the type of aircraft, is also available. There is a limited amount of information about ground transportation available as well.

PORTING TO FRENCH

As a first step in developing L'ATIS a set of several hundred English training sentences representative of the ATIS domain were translated into French. An initial grammar for parsing these derived French sentences was developed at MIT, using rules that corresponded to equivalent English rules as much as possible. Mapping the parsed sentences to an English-language semantic frame was relatively straightforward, and involved in many cases the same mappings from parse tree categories to semantic frame categories as were used for English. In addition, an initial French generation component was developed to allow the system to respond in French. The robust parsing capability [9] was extended to handle French, so that the system would generally try to answer even when it only had partial understanding. The discourse and dialogue components worked

[1] The portion of this research carried out at MIT was partially supported by DARPA under Contract N00014-89-J-1322, monitored through the Office of Naval Research.

[2] Names in alphabetic order.

directly from the semantic frame, with no knowledge of the language being spoken.

Once we had an operational shell system, it was set up at both sites. LIMSI has continued to expand the rule coverage, and MIT has provided know-how in augmenting the grammar. Close contact has been maintained via E-mail correspondence, and system updates as well as new versions of the grammar rules are exchanged via FTP between the two sites, ensuring consistency on both sides of the Atlantic.

In order to expand the coverage of the system, LIMSI researchers first began collecting queries from native French speakers *typing* sentences to the system. As new language patterns were detected, these were added to the input grammar, thus increasing the coverage. The current version of the system includes a toggle switch allowing alternation between English and French as the input/output language, which greatly aids the process of adding rules since the more fully developed English analysis is easily accessible for comparison. In the second phase of data-collection, spoken queries were recorded from native French users, with a hidden wizard typing a paraphrase of each query to the NL component.

The LIMSI continuous speech recognizer [4, 2] has been ported to this task by defining a task-dependent lexicon and bigram language model. The acoustic models are the same as used for large vocabulary speaker-independent recognition and have been trained on a subset of the BREF corpus [5]. The final step will be to integrate the recognizer with the back-end to have a complete spoken language system.

ISSUES IN PORTING

Parsing: French is in some respects more difficult to parse than English. For English, it is possible to ignore gender, as it plays a minor role restricted to +ANIMATE personal pronouns. Gender is pervasive in French, however. In addition to all nouns and pronouns, adjectives, articles and quantifiers also carry gender and number, which must agree with the noun being modified, even when in a predicate position detached from the main noun phrase. French also has many more inflectional forms on verbs, and a large number of contractions such as "*l'*avion", "les tarifs *des* vols", and "l'heure *d'*arrivée", as well as liaison phenomena as in "Quand arrive-*t*-il?"

One approach would be to ignore many feature constraints and allow the grammar to overgenerate, assuming that the input is well-formed. However, since we would like the grammar to act as a constrained language model for the recognizer, in addition to providing a meaning representation for the sentence, we want it to be able to predict the correct form as much as possible. The TINA grammar formalism already has in place a mechanism for unifying syntactic (and semantic) features during the parse process, making both local and long-distance agreement possible (e.g., "*Quel* est *le tarif le* moins *cher*?"). It was straightforward to add gender as a feature to be unified, and to add gender and number as features defined for adjectives, articles, and quantifiers. Each terminal node in the parse tree is required to unify features provided by the left sibling with any features associated with the vocabulary item. In addition, certain nonterminals can set (or clear) certain features, during either the top down or the bottom up cycle. These features are propagated through the parse tree in an orderly fashion, as described in [8].

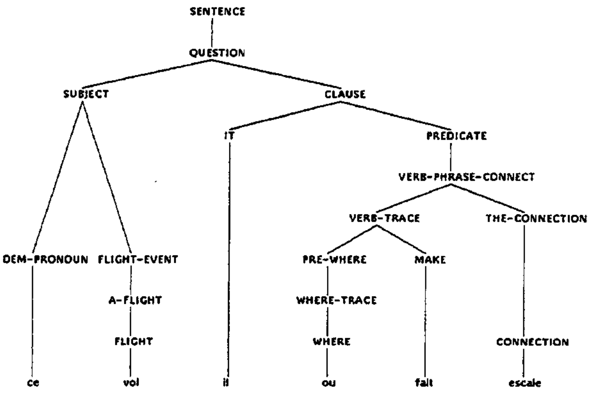

Figure 1: Parse tree for the sentence, "Où ce vol fait-il escale?" showing double movement phenomenon.

French is also challenging from the standpoint of movement phenomena. For example, the sentence, "Où ce vol fait-il escale?" would be easier to interpret if transformed to the underlying "deep structure" form, "ce vol il où fait escale?", placing the verb adjacent to both its object ("escale") and its modifier ("où"). This makes it easier to produce a semantic frame that identifies the request for *stop location*.[3] In the case of the above sentence, there are two movements – the wh-phrase to the front and the verb to the left of the pronoun. TINA's gap mechanism can handle this situation by first moving the wh-phrase to be just before the verb, and then moving the entire verb phrase to the underlying predicate position. The resulting parse tree is shown in Figure 1. We are unaware of any analogous "double movements" in the current database of English utterances.

Meaning Representation: A table-driven procedure is used to convert French or English parse trees to a common semantic frame. The functions that carry out the conversion are essentially language independent, with the language-dependent information being stored in separate tables for each language. The semantic mappings for French show very strong correspondences with those for English, which might be expected since the two languages are related.

Semantic encoding is defined at the level of the grammatical category, identified with each node in the parse tree, rather than at the level of an entire rule. All of the semantic encoding instructions are entered in the form of simple association lists. Each semantically active category (preterminal or nonterminal) in the parse tree is associated with a corresponding semantic name, each of which is in turn associated with a functional type, defining what function to call when this node is encountered in the parse tree during the stage of converting the parse to a semantic frame. There are only a few distinct functional types.

The function that converts a parse tree to a semantic frame visits each node once in a top-down left-to-right fashion, calling the appropriate functions as dictated by the mappings. Figure 2 gives category correspondences required in order to produce a semantic frame from the parse tree in Figure 1. Many of the nodes in the parse tree are ignored. Entries in the frame are

[3] Associating moved wh-phrases with their underlying original position also eliminates potential overgeneration problems and allows more general sharing of grammar rules.

Parse Category	Semantic Category	Function
question	wh-query	set-sentype
subject	topic	set-topic
verb-phrase-connect	connect	predicate
a-flight	flight	topic-name
connection	flight-mode	predicate
pre-where	in	predicate
where	where	quantifier
dem-pronoun	dem	quantifier

Figure 2: The control table required to convert from the parse tree of Figure 1 to the semantic frame of Figure 3, defining mappings from parse tree categories to semantic categories to functional types.

```
[clause-name: wh-query
 Topic:       [name: flight; quant: dem]
 Predicate:   [name: connect; flight-mode: stop;
               in: [name: NIL; quant: where ]]]
```

Figure 3: Schematic of semantic frame produced by parse tree of Figure 1 using mappings defined in Figure 2.

order-independent, and the same or a nearly equivalent semantic frame is produced from a large pool of questions with different phrasings but equivalent meanings. A schematized semantic frame for the example sentence is shown in Figure 3.

Generation: Once a semantic frame has been created, there is a straightforward procedure that generates a text response, given the output from the database. This response, which is, in many cases, a paraphrase of what the system understood the user to say, may be spoken by a speech synthesizer and provides a mechanism to keep the user in synch with the system. Generation is generally easier than understanding, because only a single way to say a given concept must be developed. Generation uses a procedure which is analagous to the procedure used by analysis to convert a sentence to a semantic frame. The top level clause name maps to a high level generation function. The table returned by the database is taken into account to determine the number of the answer or to check for presupposition failure. Individual noun phrases are constructed by consulting language-dependent ordering tables to assign each modifier its correct position within a noun phrase. Features such as number and gender are propagated from a noun to its modifiers, as well as from the subject noun phrase up to the main clause.

Considering the example from Figure 1, the subject referred to "ce vol" which needs to be interpreted in context. The discourse mechanism fills in the semantic frame, replacing "ce vol" with the appropriate flight from the history, and the complete semantic frame is passed on to the generation component. Thus, the response to our example query might be: "Le vol le plus tard de Boston à Denver fait l'escale suivante:" with the stop location identifed in the table.

In addition to sentences that are constructed in response to individual user queries, certain "canned" phrases are also generated by the system when, for example, further information is needed before a complete database query can be constructed. The control structure that determines when these phrases are needed is language independent. Separate tables for English and French control the actual productions.

DATA COLLECTION

The data collected in French contain both *typed* and *spoken* queries posed by native French speakers. As in English ATIS data collection, the subjects were asked to solve a set of task-specific scenarios selected from among 11 scenarios which were translated into French. Each session (written or spoken) lasted about 50 minutes, during which the subject solved on average 6 scenarios. In the case of typed input, the subjects interrogated the system themselves, typing in their queries with the response appearing on the terminal. Nine subjects provided a total 505 queries, with an average length of 8 words/query.

style	*s'il vous plaît*	*donc*	*okay*	*bonjour*
written	2	1	0	1
spoken	55	16	7	11

Table 1: Occurrences of certain interjections and politeness forms in written and spoken utterances.

Collection of the spoken data used a WOZ setup, where a wizard typed a paraphrased version of the spoken question to the system. The subject saw only the response of the system on the screen. The recordings were made in an acoustically isolated room, simultaneously with a close-talking, noise cancelling Shure SM10 and a table-top Crown PCC160 microphone. The wizard also monitored the recordings and could communicate with the subject via a microphone. The subject typed a key to start and stop the recordings. Whenever the subject forgot to signal the end, the system could use an automatic endpoint detector to terminate recording. Ten subjects were recorded, providing a total of 508 sentences. The average number of words per sentence is 13, including hesitations, false starts and reparations. If these spontaneous speech phenomena are excluded, the average number of words is about 11.

While data collection of typed and spoken input gave about the same number of queries in the same amount of time (about 50 queries in a 50 minute session), there were a number of differences between the spoken and written utterances. Spoken inputs are substantially longer than the written ones (13/11 vs. 8), presumably since it is easier to talk than to type. This increased length can be attributed in part to an increased occurrence of politeness forms and interjections, such as "*Bon ben tant pis* je prend le vol..." and "Donnez moi *je vous prie* des tarifs..." Some commonly occurring examples from the data are tabulated in Table 1. In addition, the variance on sentence length as a function of speaker was much higher for the spoken utterances. In particular, one subject had an average sentence length of 18 without hesitations.

Hesitations, false starts and reparations are phenomena specific to spoken queries. Hesitations occurred in 25% of the queries, and 67 of the 508 sentences had false starts. Reparations are less frequent, appearing in 29 of the sentences. An example sentence with such spontaneous speech phenomena is: "[euh] okay je veux je veux commander je veux réserver un vol le vol enfin une place sur le vol numéro trois cent trente neuf de la compagnie [euh] delta."

SPEECH RECOGNITION

The recognizer uses continuous density HMM (CDHMM) for acoustic modeling and bigram statistics estimated on the texts of typed and spoken queries for language modeling. Acoustic modeling uses 48 cepstrum-based features derived from a Bark frequency spectrum, estimated on the 0-8kHz band every 10ms. A set of 428 speaker-independent (SI), context-dependent (CD) acoustic phone models which are position and word independent [1] are used. Each phone model is a left-to-right CDHMM with

Gaussian mixture observation densities. Phone durations are modeled with gamma distributions. The recognizer uses a one pass time-synchronous Viterbi decoder [4] which includes intra- and inter-word phonological rules. The mechanism developed to handle the phonological rules can also handle the liaisons and mute-e in French.[4] The acoustic training is based on 2770 sentences from 57 speakers (28m/29f) taken from the BREF corpus [5]. Although task-specific speech data have been collected, no acoustic data from the ATIS domain have been used for training.

The recognition lexicon contains 728 words including filler words such as "euh", "hum", "bon", "enfin", and "ben". No attempt has been made to acoustically model nonspeech events. The pronunciations for the lexical entries use a set of 35 phonemes [1], and a pronunciation graph is associated with each word so as to allow alternate pronunciations, including optional phones. The recognizer uses a bigram backoff language model with probabilities estimated on the small amount of training data available. These data consist of a total of 1753 queries, 790 translated from English ATIS queries, 505 written and 458 spoken. A set of 50 spoken sentences answerable without history were reserved for test data. The perplexity of this test set is 21.

INFORMAL EVALUATION

Parsing: The initial grammar, developed from the translated queries, gave a full-parse coverage of 45% on the typed queries. After augmenting the rules and vocabulary based on phenomena observed in 350 of the queries, the coverage improved to 75%. This enhanced system provided only 48% full-parse coverage on the *spoken* data[5]. Since our parser had been trained only on written forms, it was quite deficient in handling politeness forms and interjections such as "bon ben" and "donc", a problem which accounted for a number of the parse failures. The performance would improve substantially with the addition of some simple rules to account for these phenomena.

Understanding: It was unclear how to evaluate the performance of the NL component, since no official annotations were available. Furthermore, we did not have any data that were suitable as a test set in the strict sense. As might have been expected, the initial data collection effort brought to our attention some minor system problems within the robust parser with pervasive consequences, such as the article 'une' being misinterpreted as the number one. The evaluations reported below are based on a system with such overgeneration problems fixed.

Due to time limitations, we did not examine in detail the subset that received a full analysis. To assess the remaining utterances, we decided to examine the semantic frames produced from a robust parse. The resulting semantic frames were classified into five categories: "correct", "partial", "incorrect", "no answer", and "class X" (unanswerable). Based on this subjective analysis of the 52% of the sentences not having a full parse, about 60% of the semantic frames were judged correct. Another 9% were considered partially correct; this usually meant that the system displayed a superset of what the user requested. About 15% of the semantic frames were incorrect or had no answer. 14% of the utterances were judged to be class X, meaning that they were out of the task domain. Clearly, these results are very preliminary, and we expect to have more data in the future that will serve as a true test set.

Recognition: In order to provide an initial performance measure, the speech recognizer was evaluated on 50 test utterances selected from the subset of sentences judged to be answerable without context. There are on average 5 sentences from each of the 10 recorded speakers, with a maximum of 6 from any given speaker. These recognition results are cross-task in that no task-specific acoustic data have been used to train the acoustic models. The overall word error rate is 18.2%, with a corresponding 74% sentence error rate. One speaker has a word error more than double the average. This speaker inserted many interjections and also had a large number of reparations. The limited amount of training data available to train the language models does not adaquately cover such spontaneous speech phenomena.

SUMMARY AND FUTURE PLANS

We are encouraged by our progress in creating a prototype French version of ATIS, building on a preexisting English version. Clearly, the system is still incomplete. We plan to collect more spoken data, and to expand the grammar to cover spoken phenomena more thoroughly. In the near future we plan to integrate the recognizer into the system. Improvements in recognition accuracy can be expected by using task-specific acoustic training data and by explicitly modeling spontaneous speech as well as non-speech events. The lexicon will be expanded to include additional pronunciations and phonological rules. The language model will be extended from a simple bigram to a model based on grammatical categories. This approach is better suited to French with the large number of homophones and word forms, as well as to the ATIS task where it cannot be expected to observe all occurrences of numerical items such as flights, times, and dates. Backend improvements continue to be made based on data collected in both English and French. Because the two systems share a common backend, any improvements in one immediately carry over to the other. After further development of the system and the recording of more spoken data for training and test, the performance of the components as well as the complete system will be evaluated.

REFERENCES

[1] J.L. Gauvain, L.F. Lamel, "Speaker-Independent Phone Recognition Using BREF," *Proc. DARPA Speech and Natural Language Workshop*, Arden House, Feb. 1992.

[2] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, "Speaker Independent Continuous Speech Dictation," *Proc. EUROSPEECH-93*, Berlin, Sept. 1993.

[3] J. Glass, D. Goodine, M. Phillips, S. Sakai, S. Seneff, and V. Zue, "A Bilingual VOYAGER System," *Proc. EUROSPEECH-93*, Berlin, Sept. 1993.

[4] L.F. Lamel, J.L. Gauvain, "Continuous Speech Recognition at LIMSI," Final review of the *DARPA Artificial Neural Network Technology (ANNT) Speech Program*, Stanford, CA, Sept. 21-22.

[5] L.F. Lamel, J.L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," *Proc. EUROSPEECH-91*, Genova, Sept. 1991.

[6] J. Peckham, "Speech Understanding and Dialogue over the Telephone: an Overview of the ESPRIT SUNDIAL Project," *Proceedings, DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, Feb. 1991.

[7] P. Price, "Evaluation of Spoken Language Systems: The ATIS Domain," *Proceedings, DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, June, 1990.

[8] S. Seneff, "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, Vol. 18, No. 1, pp. 61-86, 1992.

[9] S. Seneff, "Robust Parsing for Spoken Language Systems," *Proceedings, International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 189-192, March 23-26, 1992.

[4] Liaisons are phonemes inserted at the junction of two words.

[5] As input to the parser, we used a version of the spoken utterances with false starts and filled pauses removed.