MULTILINGUAL HUMAN-COMPUTER INTERACTIONS: FROM INFORMATION ACCESS TO LANGUAGE LEARNING[1]

Victor Zue, Stephanie Seneff, Joseph Polifroni, Helen Meng, and James Glass

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

ABSTRACT

This paper describes our recent work in developing multilingual conversational systems that support human-computer interactions. Our approach is based on the premise that a common semantic representation can be extracted from the input for all languages, at least within the context of restricted domains. In our design of such systems, language dependent information is separated from the system kernel as much as possible, and encoded in external data structures. The internal system manager, discourse and dialogue component, and database are all maintained in a language transparent form. We will describe two possible application areas for such multilingual capabilities: on-line information access using multilingual spoken dialogue, and the learning and maintenance of a foreign language using a multilingual conversational system.

1. INTRODUCTION

Since 1989, our group has been conducting research leading to the development of conversational systems – systems that can converse with users in a spoken dialogue in order to fulfill their needs. This line of research is motivated by our belief that many aspects of human computer interactions that lend themselves to spoken input – making travel arrangements or finding a relevant document – are in fact exercises in interactive problem solving. The solution is often built up incrementally, with both the user and the computer playing active roles in the "conversation." Therefore, several language-based input and output technologies must be developed and integrated to reach this goal. Regarding the former, speech recognition must be combined with natural language processing so the computer can understand spoken commands (often in the context of previous parts of the dialogue). On the output side, some of the information provided by the computer – and any of the computer's requests for clarification – must be converted to natural sentences, perhaps delivered verbally.

Over the years, we have developed a series of conversational systems with increasing complexity, including VOYAGER for urban navigation and exploration [1], PEGASUS for air travel planning [2], and GALAXY for on-line, multi-domain information access [3]. A

cornerstone of our research effort throughout this period has been the development of multilingual conversational systems. As illustrated in this special session, there are several ongoing international spoken language *translation* projects whose goal is to enable humans to communicate with one another in their native tongues. Our objective, however, is somewhat different. Specifically, we are interested in developing multilingual human-*computer* interfaces, such that the information stored in the database can be accessed and received in multiple spoken languages. We believe that there is great utility in having such systems, since information is fast becoming globally accessible. Furthermore, we suspect that this type of multilingual system may be easier to develop than speech translation systems, since the system only needs to anticipate the diversity of one side of the conversation, i.e., the human side, and the topic of conversation is typically quite focused.

This paper summarizes our work on developing multilingual conversational systems over the past four years. Rather than delving into the details of the implementation, we will first outline our approach, with particularly emphasis on the language-independent meaning representation. We will then briefly describe our ongoing work in providing multilingual capabilities for GALAXY, followed by our use of multilingual conversational systems for language learning and maintenance. The readers are referred to our other publications for more detailed and technical descriptions.

2. GENERAL DESCRIPTION

2.1. Architecture

Figure 1 shows the architecture of our conversational system, emphasizing its multilingual nature. The language-independent aspects of the system components are described in more detail in the following subsections. This is followed by a discussion of some multilingual issues.

Speech Recognition For conversion from signal to words, we use the segment-based SUMMIT system developed in our group, which has been ported to fourteen domains and three languages. Detailed description of SUMMIT can be found in the literature. We have recently incorporated the notion of *anti*-phones into SUMMIT, modelling both the positive as well as negative examples of phones. This has resulted in significant improvement in its performance [4].
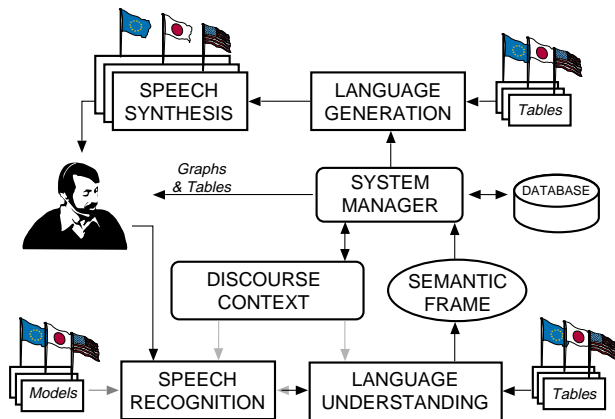
Figure 1: Architecture of MIT's multilingual conversational system.

```
INPUT: WHERE IS THE LIBRARY NEAR CENTRAL SQUARE
FRAME:
      Clause: LOCATE
           Topic: PUBLIC-BUILDING
                Quantifier: DEF
                Name: library
                Predicate: NEAR
                       Topic: SQUARE
                            Name: Central
```

Figure 2: Semantic frame for the sentence, "Where is the library near Central Square?

Language Understanding The language understanding component makes use of a probabilistic natural language system developed in our group called TINA, which has been ported to eleven domains and five languages. A detailed description of TINA can be found in [5]. Since its introduction in 1989, TINA has undergone significant refinement. A "robust parsing" strategy that attempts to piece together an understanding of the utterance from analyses of fragments has been introduced [6]. This strategy significantly improved the system's ability to understand spontaneous speech that is often agrammatical. More recently, a new formalism of TINA called "layered bigrams" was introduced so that the system could accommodate robust parsing in conjunction with full integration between speech recognition and language understanding [7].

Meaning Representation The parse tree generated by TINA is converted to a hierarchical *semantic frame* which is intended to capture the meaning of the input utterance in a language-independent form. To produce the semantic frame, each active node in the parse tree is mapped to a corresponding *semantic* class, which in turn is associated with a specific *syntactic* role (such as clause, topic, predicate, quantifier, etc.) [1]. An example semantic frame for the sentence, "Where is the library near Central Square," is shown in Figure 2.

The semantic frame serves many roles in our spoken language systems: it is used as the basis for accessing information from application databases, to maintain a discourse history, and also for natural

language generation. Since the frames are intended to capture the relevant semantic information of the input query, they can also be used to paraphrase the input. This latter capability has proved to be quite useful for multilingual development.

Response Generation Natural language generation in our conversational systems is achieved using GENESIS, which has been ported to five domains and seven languages [8].[2] GENESIS serves the dual role of paraphrasing meaning representations and generating responses to the user. The input to GENESIS is a semantic frame created by TINA, which may be augmented to include information returned from a database query. The three data structures of GENESIS – a lexicon, a set of message templates, and a set of rewrite rules – are language-dependent and external to the GENESIS system itself, thus facilitating its porting to a new language.

Discourse and Dialogue The discourse component is responsible for pronoun resolution, implicit predicate inheritance, filling obligatory case roles, and dealing with fragments. It has multimodal capabilities, in that mouse-clicked items become reference entities. It operates on semantic frames and requires no knowledge of the input or output language. A declarative table specifies the inheritance needs of a particular domain. Details can be found in [9].

The dialogue component is based on the notion of an electronic form (*E*-form) that keeps a record of a transaction in progress. The *E*-form is particularly relevant for complex transactions, such as purchasing an airline ticket or an automobile (see [10] for details). It is consulted frequently in the course of a dialogue, and the system often initiates specific requests to fill particular slots in the form.

2.2. Multilingual Issues

Approach Our approach to developing multilingual conversational systems is predicated on the assumption that it is possible to extract a *common*, language-independent semantic representation from the input, similar to the *interlingua* approach to machine translation [11]. Whether such an approach can be effective for unconstrained machine translation remains to be seen. However, we suspect that the probability of success is high for spoken language systems operating in restricted domains, since the input queries will be goal-oriented and therefore more constrained. In addition, the semantic frame may not need to capture all the nuances associated with human-human communication, since one of the participants in the conversation is a computer. Thus far, we have applied this formalism successfully across several languages and domains.

To develop a multilingual capability for our spoken language systems, we have adopted the strategy of requiring that each component in the system be as language transparent as possible. Referring back to Figure 1, the system manager, discourse component, and the database are all structured so as to be independent of the input or output language. In fact, the input and output languages are completely independent from each other so that a user could speak in one language and have the system respond in another. In addition, since contextual information is stored in a language indepen-

---

[2]We have not actively developed a text-to-speech system and have instead relied on systems developed by others for speech generation.

E:	Where is the library near Central Square
F:	Où se trouve la bibliotheque qui est près de Central Square
I:	Dove sta la biblioteca vicino a Central Square
J:	Sentoraru sukuea no chikaku no toshokan wa doko desu ka

Figure 3: Sentences in English (E), French (F), Italian (I) and Japanese (J) that produce the semantic frame shown in Figure 2.

dent form, linguistic references to objects in focus can be generated based on the output language of the current query. This means that a user can carry on a dialogue in mixed languages, with the system producing the appropriate responses to each query.

Where language-dependent information is required, we have attempted to isolate it in the form of external models, tables, or rules, as illustrated in Figure 1, for the speech recognition, language understanding, and generation components. For example, all four sentences shown in Figure 3 can be parsed by TINA, albeit with different grammar rules. They all result in the same semantic frame shown in Figure 2.[3] For speech recognition, we trained the basic SUMMIT system for the languages of interest, using data recorded from native speakers for each language. For text-to-speech synthesis we acquire an appropriate text-to-speech system for each language.

If we are to attain a multilingual capability within a single system framework, the task of porting to a new language should involve only adapting existing tables or models, without requiring any modification of the individual components. By incrementally porting the system to new languages we hope to slowly generalize the architecture of each component to achieve this result.

Implementation To port a spoken language system to another language, the following steps must be taken. First, the system must be able to generate the appropriate responses in the target language from semantic frames, which are derived from a set of English training sentences. Second, the language generation capability will enable the collection of sentences in the target language for system development, training, and evaluation. This is done with a bilingual typist serving as the wizard, who translates the spoken utterance from the target language into an English sentence that the system is able to understand. The resulting semantic frame could then be used to generate the responses in the target language. Third, the collected sentences are used to develop a grammar for the target language, and to extend the capabilities of the natural language component. Fourth, lexical items (with associated pronunciations), acoustic models, and language models must be derived from the training sentences in order to bring up the recognizer in the target language. Finally, the performance of the system must be evaluated using previously unseen data. The capabilities of the system will be improved and refined as more training data are acquired.

3. INFORMATION ACCESS

The first multilingual conversational system that we developed was the VOYAGER system. VOYAGER can engage in verbal dialogues with users about a geographical region within Cambridge, Mas-

sachusetts, in the USA. It can provide users with information about distances, travel times, or directions between objects (e.g., restaurants, hotels, post offices, subway stops) located within this area, as well as information such as addresses or telephone numbers of the objects themselves. While VOYAGER is constrained both in its capabilities and domain of knowledge,[4] it nevertheless contains all the essential components of a conversational system, including discourse maintenance and language generation. As of 1994, VOYAGER operated in a trilingual mode, where the user can select among the three choices, English, Japanese, or Italian, for the communication language [1]. A user can also freely mix the three languages in a single conversation, and the system will incorporate context appropriately, regardless of the language of the context-setting query(s).

More recently, we have concentrated our multilingual development effort within the framework of GALAXY, which enables universal information access using spoken dialogue [3]. GALAXY differs from its predecessors in several important ways. First, it accesses *real* databases residing on the information highway. We believe this strategy of developing human language technologies within real applications will force us to confront some of the critical technical issues that may otherwise elude our attention, such as dialogue modelling, new word detection/learning, and portability across domains and languages. Second, GALAXY utilizes a distributed, client/server architecture that shares compute servers (for human language technologies) and domain servers among many users, and relies on lightweight clients for input/output. An example of a compute server would be a speech recognition server that converts the speech signal into hypothesized word strings.[5] Each domain server has some knowledge about a specific domain (e.g., air travel, weather, classified Ads for automobiles, or local area restaurants), and is capable of accessing specific databases. The client provides the interface to the user; it captures audio or typed input from the user, and presents the servers' responses using graphics, text, and synthetic speech. Third, it is our intention to minimize the computational needs of GALAXY's client program, thus providing information access to the widest user population in the most affordable way. At present, one can launch a GALAXY client program anywhere on the Internet and receive aural and visual information by simply using a telephone to talk to the servers running at MIT. Ultimately, we envision that the interface can simply be a telephone and a cable television, thus enabling mobile and affordable information access. At present, GALAXY is connected to many on-line databases, and it is relatively straightforward to extend it to include other domains [12, 13]. Users can query GALAXY in natural English (e.g., "what is the weather forecast for Miami tomorrow," "are there any hotels in Boston with a pool and a jacuzzi," "show me the flights from Boston to San Francisco," "do you have any convertibles with manual transmission") and receive verbal and visual responses.

It is our intent to make GALAXY a multilingual system for accessing on-line information. The current languages of interest are Japanese, Spanish and Mandarin Chinese. The first step of this process – the development of language generation capabilities – is well under-

---

[3] The only exception is that the quantifier is absent in the Japanese version.

[4] It only has a vocabulary of 500-700 words, depending on the language, and it knows about a few dozens objects.

[5] At present, a user can access the SUMMIT recognizer at MIT or the SPHINX-II recognizer from CMU.

E:	SHOW ME THE DIRECT FLIGHTS FROM BOSTON TO LONDON
J:	Boston kara London eno tyokko: bin o hyo:ji shite kudasai.
E:	HOW LONG WOULD IT TAKE TO WALK TO THIS BANK
M:	zou(3) dao(4) zhei(4) ge(0) ying(2) hang(2) yao(4) duo(1) jiu(3)?
E:	ARE THERE ANY MEXICAN RESTAURANTS IN BOSTON
S:	Hay algunos restaurantes en Boston que sirven comida Mexicana?

Figure 4: Examples of English (E) sentences and their corresponding paraphrases in Japanese (J), Mandarin (M), and Spanish (S) in the GALAXY domain. The numerics in Mandarin indicate lexical tones, and the colons in Japanese indicate long vowels.

way. Figure 4 show some example English sentences and their corresponding paraphrases in the target language.

4.   LANGUAGE LEARNING

In 1993, our group started to investigate the feasibility of utilizing spoken language technology for foreign language learning and maintenance. The outgrowth is a system called LANGUAGE TUTOR, which provides a non-threatening, interactive environment to help people acquire and maintain language skills. Students can learn how to pronounce words and sentences by listening to the spoken versions provided by the system, speaking their own version and receiving feedback on their pronunciation skills, or comparing the two versions. Feedback is obtained by having the SUMMIT speech recognition system shadow the student as he/she pronounce the appropriate words. If the student encounters difficulty, the system will prompt her/him by speaking the appropriate words and phrases. LANGUAGE TUTOR is currently operating for English and Japanese.

While the LANGUAGE TUTOR can potentially be helpful in learning a new spoken language, the system is nevertheless limited in its ability to provide an active learning environment. A possible, novel approach for language learning may be to dovetail the LANGUAGE TUTOR with a multilingual conversational system such as VOYAGER or GALAXY. Each lesson, in addition to introducing new vocabulary and linguistic constructs, would contain a scenario specifically designed for the lesson (e.g., looking for a hotel with a business center, or checking the arrival time of a particular flight). Students are then asked to practice their passively acquired language skills in an active setting, in which they must learn to interact with the system in order to obtained the desired responses. This environment, we suspect, will enable students to practice interactions in a risk-free setting. It has the potential advantage of going beyond the mechanics of standard reading/speaking exercises, and stimulating real-world interactions in a language laboratory. Our approach to developing multilingual conversational system offers another advantage. Since the core of the conversational system is language transparent, students can speak in their native language and hear responses in the target language, or vice versa, thus providing a flexible alternative to practice speaking and listening in a structual manner. For example, if the student forgets the word "library" for Japanese, he/she can simply set the system in "cross-language" mode and ask a question in English such as, "Where is the nearest library?" She/he can presumably deduce the Japanese word "toshokan" from the system's re-

sponse. We have investigated the feasibility of this approach, and have obtained some encouraging, albeit anecdotal, results. People generally found the combination of LANGUAGE TUTOR and multilingual VOYAGER to be fun and useful. In fact, several members of our group enjoyed learning Japanese more by using the system than through a classroom setting. Much work remains to be done, however, if one is to pursue such an approach. In addition to the development of spoken language technologies, the participation of teachers of foreign languages would be indispensable.

5.   SUMMARY

This paper describes our approach to developing multilingual conversational systems, and provides a status report on our research in developing such systems. Our near-term multilingual research will be conducted within the framework of GALAXY, focusing on Japanese, Mandarin, and Spanish. The existing generation capabilities for these languages will soon enable us to collect data from native speakers, and to use the resulting data for the development of language understanding and speech recognition capabilities.

6.   ACKNOWLEDGEMENT

7.   REFERENCES

1. Glass, J. et al., "Multilingual Spoken Language Understanding in the MIT VOYAGER System," *Speech Communication*, vol. 17, 1–18, 1995.

2. Zue, V. et al., "PEGASUS: A Spoken Language Interface for On-Line Air Travel Planning," *Speech Communication*, vol.15, 331–340, 1994.

3. Goddeau, D. et al., "GALAXY: A Human Language Interface to Online Travel Information," *Proc. ICSLP-94*, 707–710, Yokohama, Japan, 1994.

4. Glass, J. et al., "A Probabilistic Framework for Feature-Based Speech Recognition," *These Proccedings*.

5. Seneff, S., "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, Vol. 18, No. 1, 61–86, 1992.

6. Seneff, S., "Robust Parsing for Spoken Language System," *Proc. ICASSP*, 189–192, San Francisco, CA, 1992.

7. Seneff, S. et al., "Language Modelling for Recognition and Understanding Using Layered Bigrams," *Proc. ICSLP*, 317–320, Banff, Alberta, Canada, 1992.

8. Glass, J. et al., "Multilingual Language Generation Across Multiple Domains," *Proc. ICSLP*, 983–986, Yokohama, Japan, 1994.

9. Seneff, S. et al., "Multimodal Discourse Modelling in a Multi-User Multi-Domain Environment," *These Proceedings*.

10. Goddeau, D., et al., "A Form-Based Dialogue Manager for Spoken Language Applicaions," *These proceedings*.

11. Hutchins, W.J., and Somers, H.L., *An Introduction to Machine Translation,* Academic Press, 1992.

12. Meng, H. et al., "WHEELS: A Conversational system in the Automobile Classifieds Domain," *These Proceedings*.

13. Seneff, S. and Polifroni, J., "A New Restaurant Guide Conversational System: Issues in Rapid Prototyping for Specialized Domains," *These Proceedings*.