FROM INTERFACE TO CONTENT:
TRANSLINGUAL ACCESS AND DELIVERY OF ON-LINE INFORMATION[1]

Victor Zue, Stephanie Seneff, James Glass, Lee Hetherington, Edward Hurley, Helen Meng,
Christine Pao, Joseph Polifroni, Rafael Schloming, Philipp Schmid

Spoken Language Systems Group, Laboratory for Computer Science
Massachusetts Institute of Technology, Cambridge, MA 02139 USA
http://www.sls.lcs.mit.edu

ABSTRACT

This paper describes our initial implementation of a system to provide world-wide weather information over the telephone. The information is gathered from several different sites on the Web, preprocessed, and cached locally into a relational database to make access both fast and selective. Our natural language tools, originally developed for processing user queries, are used here for understanding content, and for subsequently translating it into languages other than English. The system is operational, and we have been collecting data from real users via a toll-free number. We report here on an initial evaluation both of the full system in English and of the quality of the responses in German.

1. INTRODUCTION

Over the past several years, our group has been involved in the development of conversational interfaces for accessing on-line information using human language technologies. Increasingly, we see the need to utilize the same technologies for understanding the content in order to manipulate and deliver the information to the user, since so much of the available content (e.g., in newspapers, radio broadcasts, or Web pages) is linguistic in nature. To this end, we are developing a system called JUPITER, which allows a user to access and receive on-line weather information over the phone and in multiple languages. JUPITER utilizes the client-server architecture of GALAXY [1], but it specializes in weather-specific information world-wide obtained from a variety of Web sites. It can give a weather report for a particular day or several days, and answer specific questions about weather phenomena such as temperature, wind speed, precipitation, pressure, humidity, sunrise, etc. JUPITER serves as a testbed for several important areas that have surfaced on our research agenda, including displayless interaction, virtual browsing, information on demand, and translingual content management. The system currently has weather information for approximately 500 cities, mostly within the United States, but also selected major cities world-wide. The information is available in English, but JUPITER is also acquiring some "translingual" capabilities, by parsing and translating weather information into several other languages.

1.1. Language as Interface

From the standpoint of interface, our research is increasingly moving toward a telephone-based platform. A displayless environment, in which only the telephone is available for information access and delivery, is important be-

cause the telephone is so much more pervasive when compared to PCs equipped with Internet access. By using the telephone as a means of accessing the information, we can empower a much larger population. In the scenario that we envision, a user could conduct "virtual browsing" in the information space without ever having to point or click. Displayless information access poses new challenges to conversational interfaces. If the information can only be conveyed verbally, the system must rely on the dialogue component to focus the interaction into manageable subdialogues, the language generation component to express the information succinctly, and the text-to-speech component to generate natural and intelligible speech.

1.2. Language as Content

From the standpoint of content, the linguistic information that exists on-line often contains much more than the user would like. Weather forecasts are obtained from multiple sites on the World Wide Web, whose information content complement one another. If the system fully understands the information it is presenting to the user, it will be able to select subsets that are relevant to the question, avoiding boring the user with long monologues on irrelevant distractions. It will also be able to recognize when two sources are providing overlapping information, and can select one or the other based on quality considerations. Finally, by representing all linguistic data in a language-transparent semantic frame format, the notion of translating the weather reports into other languages becomes far more feasible.

2. CREATING THE DATABASE

JUPITER currently obtains all of its information from several complementary weather sources available on the Web. One of the sources provides up to 4-day weather and temperature predictions worldwide, whereas another is concentrated within the United States, but far more detailed in its descriptions. A third one is an excellent resource for quantitative information such as temperature, humidity, barometric pressure, and sunrise and sunset times.

The National Weather Service provides weather forecasts for 279 cities within the United States. These reports are manually written, and there seems to be no strict format, so they are problematic from the standpoint of regularization. But because they give a rich description of the weather, including predictions of amounts of precipitation, advisories for hurricanes, floods, etc., we feel it is worth the effort to process them.

JUPITER is updated several times a day, by polling the Web for any changes in predictions. An automatic procedure parses the data into semantic frames [2], and a second process sorts them into categories based on the meaning. As illustrated in Figure 1, each weather report is first converted to an indexed list of semantic frames, one for each sentence. The indices are then entered into

1. Wednesday: [date]

2. Becoming very windy and turning colder with a 60 percent chance of snow. [weather] [snow] [wind]

3. Near blizzard conditions and dangerous wind chills developing. [weather] [snow] [temperature]

4. High around 20 with temperature falling into the single digits by late afternoon. [temperature]

5. Northwest winds increasing to 25 to 45 mph. [wind]

Figure 1: Extract from an on-line weather report maintained by the National Weather Service, enumerated and annotated for categories.

O *Haleakala* summit winds becoming southerly 40 to 60 mph with occasional higher gusts.
R Summit winds becoming southerly 40 to 60 mph with occasional higher gusts.
O Heavy snow warning *extended* through this morning for mountains *above 4500 feet.*
R Heavy snow warning *remains in effect* through this morning for *higher* mountains.

Figure 2: Example rephrasings to eliminate parse failure. O = original, R = rephrased.

the relational database under appropriate topicalized categories. To retrieve the answer to a particular user request, the system first retrieves the indices of the relevant sentences in the weather report via an SQL query, then orders them sequentially, and finally paraphrases each of the corresponding semantic frames in turn, to compose a verbal response. Delays are minimal, since the system has preprocessed all current information into semantic frames in a local cache.

After less than three months, the system had achieved a very high coverage of incoming weather reports. Any sentences that fail to parse can be rephrased by the system developer to recover an equivalent meaning. The grammar can later be expanded to encompass a broader base. Example rephrasings are given in Figure 2. By requiring a full parse and hand-editing sentences that fail to parse, we can guarantee a high reliability in the understanding and regeneration. We feel it is important to protect the system from providing the user with incorrect or garbled information, which would be much more likely with robust parsing strategies. The current grammar contains nearly 1000 unique words, with about 650 categories, half of which are pre-terminal.

An example of selective database retrieval is given in Figure 3. The user first asked whether it would snow today in North Dakota. The SQL query returned all the cities in North Dakota for which some mention of snow was made. A follow-up query asks for the particular conditions in Fargo, and the system replies with all parts of today's weather report that mention snow, which turns out to include a sentence mainly about the wind speed.

3. TRANSLINGUAL CONTENT

Because the information that JUPITER maintains is represented in a semantic frame format, it is straightforward to translate this information to another language besides English. We have begun an effort to paraphrase the weather responses into German, Mandarin Chinese, and Spanish. For each of these languages, a native speaker who is fluent in English is preparing the corresponding GENESIS generation rules [3]. We have not dealt extensively in the past with German, a particularly difficult language due to

U: WILL IT SNOW TODAY IN NORTH DAKOTA
Q: select distinct city from geo, event where event.apt_code = geo.apt_code and event.snowspk is not null
J: The forecast indicates snow for the following places in North Dakota: Bismarck, Dickinson, Devil's Lake, Fargo, Grand Forks, Williston, James-town, and Minot. Please select one.
U: HOW ABOUT FARGO
Q: select distinct dayspk, snowspk from event, geo where event.apt_code = geo.apt_code and event.day = 0 and event.snowspk is not null
J: In Fargo in North Dakota today: becoming windy with light snow developing. Wind east increasing to 20 to 30 miles per hour producing drifting snow. Can I help you with something else?

Figure 3: Example of a short dialogue between a user and JUPITER, showing the database query and the filtered responses. U = User, Q = SQL query, J = JUPITER.

```
clause: weather_event
   topic: accumulation
      name: snowfall
      pred: amount
         topic value, name: 2
            pred: to_value
               topic: value, name: 4, units: inches
      pred: by_time
         pred: time_interval
            topic: time_of_day, name: morning
```

Input:	2 to 4 inches snowfall accumulation by morning
English:	snowfall 2 to 4 inches by morning
German:	Schneefall 2 bis 4 Inch bis am Morgen
Spanish:	nevada 2 a 4 pulgadas antes de la manana

Figure 4: Example sentence and semantic frame along with paraphrases in English, German and Spanish.

its extensive use of inflectional endings. We had to augment GENESIS with a more sophisticated ability to deal with case, which can be assigned in the vocabulary file by prepositions and verbs. In addition, we needed to be able to specify the correct inflectional endings for nouns and adjectives as a function of case, gender, and number.

There were a few instances when the same word in English had to be given a different translation depending on the context. For example, the word "light," translates differently into Mandarin for the two phrases, "light wind" ("qinq1") and "light rain" ("xiao3"). GENESIS handles this situation using a semantic grammar that can categorize the two cases into different adjective types.

Figure 4 gives an example of a semantic frame for the sentence, "2 to 4 inches snowfall accumulation by morning," along with the corresponding paraphrases in three languages. Note that the preposition "by" has been interpreted in the semantic frame as denoting a time expression, allowing the appropriate translation of this diversely realized preposition.

4. DIALOGUE INTERACTION

We have discovered several interesting issues with regard to proper dialogue modelling to accommodate users' requests, and we are becoming increasingly aware of the benefits in letting real users influence the design of the interaction. One of the critical aspects of any conversational interface is the need to inform the user of the scope

of the system's knowledge. JUPITER only knows approximately 500 cities, and users need to be directed to select relevant available data when their explicit request yields no information. Even for the cities it knows, JUPITER does not necessarily have the same knowledge for all cities.

JUPITER has a separate geography table organized hierarchically, enabling users to ask questions such as "What cities do you know about in the Caribbean?" This table is also used to provide a means of summarizing a result that is too lengthy to present fully. For example, if the user asks where it will be snowing in the United States, there may be a long list of cities expecting snow. The system then climbs a geographical hierarchy until the list is reduced to a readable size. For example, JUPITER might list the states where it is snowing, or it might be required to reduce the set even further to broad regions such as "New England," and "Northwest." We try to restrict the size of an enumerated list to under 10 items, if possible.

During our data collection sessions, we noticed considerable frustration among users who were trying to find information about sunrise and sunset time, and frequently coming upon cities for which this information did not exist. We realized that the system needs to distinguish between the general set of cities it knows, and the particular knowledge associated with each of those cities. Based on these observations, we decided to augment the system with the capability of suggesting a list of alternative cities in the same geographic region for which the particular data requested *are* available.

In addition to these general considerations, several phenomena required special attention. For example, we had calls after midnight when users, asking for "tomorrow's" weather, really wanted "today's" weather, defined from midnight to midnight. We also had callers from Canada who wanted temperature information presented in degrees Celsius rather than Fahrenheit. We have augmented the system to take these issues into account. To encourage the user to continue the dialogue after each exchange, we implemented a simple mechanism to alternate among a set of continuation phrases that followed the delivery of information. Figure 5 shows an actual dialogue between a user and JUPITER, illustrating this behavior.

5.   DATA COLLECTION

We have been actively collecting data within the JUPITER domain since the beginning of 1997. As we've done in previous domains, we first developed a prototype JUPITER system and used it to collect spontaneous speech using a Wizard paradigm, with a human typist in the loop and subjects brought into the lab and given scenarios to solve. At the same time, we solicited read speech using both our Web-based data collection facility [4] and a phone number that subjects could call to read from pre-distributed lists.

Once these data had been collected, we were able to train a recognizer and move on to system-based data collection. We currently have a toll–free number that is available 24-hours/day for subjects to call to find out weather information. The utterances collected from this facility are also used as training data. The toll–free number has been a particularly powerful method for collecting data from a variety of subjects in a short period of time. We feel that these calls accurately reflect the way users want to interact with such systems.

Table 1 summarizes the current status of the data collection effort. The numbers for the spontaneous, i.e., system-in-the-loop, data collection reflect just one month's recording, with subjects solicited by word-of-

J: Hi! Welcome to JUPITER, the MIT Lab for Computer Science weather information system. This call is being recorded for system development. You may hang up or ask for help at any time. How can I help you?

U: CAN YOU TELL ME WHAT'S THE WEATHER LIKE IN BOSTON TODAY

J: In Boston this afternoon. Partly sunny early. Then becoming cloudy with 30 percent chance of showers toward evening. Tonight: rain. Possibly heavy at times. Near 100 percent chance of rain. Can I help you with something else?

U: YES WHAT ABOUT SAN FRANCISCO

J: In San Francisco today: patchy fog near the ocean morning. Otherwise mostly sunny. Tonight: areas of coastal fog and low clouds. Otherwise fair. Is there something else I can do for you?

U: YES HOW COLD IS IT IN DALLAS TEXAS TODAY

J: In Dallas today: high upper 40s. Tonight: lows near 32. What else would you like to know?

U: I'D LIKE TO KNOW WHETHER IT'S SNOWING IN DETROIT

J: I'm sorry, I didn't understand you. Please rephrase your query.

U: WHAT IS THE WEATHER IN DETROIT

J: In Detroit this afternoon: occasional light rain and drizzle. Tonight: rain showers likely. Becoming snow showers by late tonight. 70 percent chance of precipitation.

Figure 5: Example spoken dialogue between JUPITER and a user.

Type	Utterances collected	Number of calls	Utterances per call
Wizard	1059	27	39
Read	3600	103	35
Spontaneous	1052	259	4.1

Table 1: Summary of data collection efforts to-date.

mouth from among the families and friends of members of our group. We anticipate an even larger supply of such data once the phone number has been more widely publicized.

6.   PRELIMINARY EVALUATION

Since the initial stages of JUPITER system development, we have tried to assure that there is a formal method for evaluating progress in all of the different aspects of the total system. Below, we have divided the evaluation criteria along the dimensions of interface vs. content. Within each of those dimensions we have evaluated components separately as well as in aggregate. Evaluations of the interface are necessarily preliminary, as the system has only been operational for a few weeks at the time of this writing.

Evaluating Interface For evaluating JUPITER's performance during conversations with users, we selected a set of log files that corresponded to calls made to the system during approximately its third week in operation. Any calls made by system developers were eliminated from consideration. We evaluated this material in terms of its speech recognition performance, parse coverage, response generation (in both English and German) and overall speech understanding.

Table 2 shows the understanding results, judged subjec-

tively by examining the log files of user interactions. Overall, the system was able to correctly answer over half of the queries. Users rarely spoke out of domain. An additional 48 queries were discarded due to various mishaps, e.g., the server was inoperational or the recording contained no evaluable speech. We also assessed the parse coverage of the orthographic transcription of the user queries. We found that 83% of the queries obtained a parse, which was always correct. We evaluated the recognition performance at the time of data collection and obtained a word error rate of 29.6% [5]. When sentences containing out-of-vocabulary items are removed, this improves to 21.9%. This same system had a word error rate of 5% when evaluated on read speech.

Evaluating Content Shortly after we began parsing weather reports, we decided to maintain a careful record of parse coverage over time, for the National Weather Service material, to see whether the system is reaching convergence. A plot of parse coverage, averaged over weekly intervals, is given in Figure 6. It was reassuring to see a steady increase in coverage, up to about 99%. Parse coverage is now hovering in the 99% range; there always seem to be a few sentences that fall outside of the grammar's domain, which, as mentioned previously, can be readily repaired.
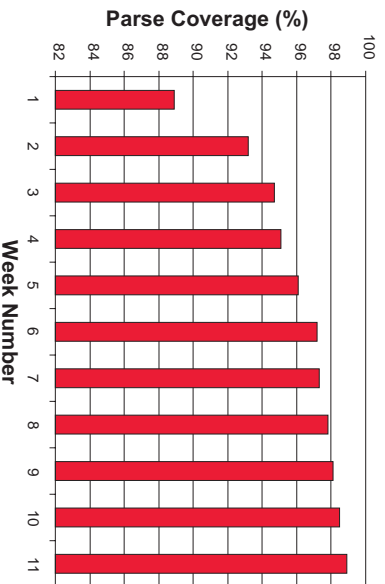

Figure 6: Parse coverage over one-week intervals for National Weather Service weather reports.

To evaluate the quality of German translations of system-generated weather reports, we reparsed all the English weather reports spoken within the log files used for understanding evaluation and then generated paraphrases in German from the resulting semantic frames. This also served as a sanity check on our English paraphrasing capabilities. There were a total of 885 sentences, of which 2 failed to reparse due to inappropriate English paraphrases. A summary of the German performance on the remaining sentences is given in Table 3.

7. SUMMARY & FUTURE WORK

In this paper we have described a system that provides access to dynamic on-line databases via the telephone. The system carries on a conversation with the user, and answers specific questions about weather world-wide. It remains current through frequent updates into a relational database, which also allows it to answer questions that would be difficult, or impractical, to answer through directly accessing the source Web sites, e.g., identifying tornado or hurricane warnings issued anywhere in the United States.

We are steadily collecting data from real users, and these will be incorporated into the training of our acoustic and language models. The system has a discourse and dialogue mechanism that is becoming increasingly sophisticated as we peruse log files.

Much work remains to be done. With audio-only responses, the quality of the synthesized speech becomes an important consideration. While we are currently using a commercial, off-the-shelf TTS system, we feel it should be possible to implement a high-quality concatenative speech synthesizer, and we have begun research in this area. We would like to augment the system with dynamic language modelling capabilities. For example, if a list of cities is presented to the user, then these cities should be accentuated in the recognizer's language model. We also need to develop a more sophisticated discourse-context-dependent help mechanism, that will transparently educate the user as to the system's complete capabilities. Of course, we will continue the development of the response capabilities for German, Mandarin, and Spanish, and we will develop a corresponding ability to handle *input* queries in these languages.

We plan to develop other on-line services as natural extensions to JUPITER. There are a number of similar domains for which the information is dynamic and the vocabulary is sufficiently limited to support practical conversational interfaces.

8. REFERENCES

[1] D. Goddeau, E. Brill, J. Glass, C. Pao, M. Phillips, J. Polifroni, S. Seneff, and V. Zue, "GALAXY: A Human-language Interface to On-line Travel Information," *Proceedings, ICSLP*, pp. 707–710, Yokohama, Japan, 1994.

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

[3] J. Glass, J. Polifroni and S. Seneff, "Multilingual Language Generation Across Multiple Domains," *Proceedings, International Conference on Spoken Language Processing*, pp. 983–986, Yokohama, Japan, Sept., 1994.

[4] E. Hurley, J. Polifroni, and J. Glass, "Telephone Data Collection Using the World Wide Web", in *Proc. Fourth International Conference on Spoken Language Processing* (Philadelphia), pp. 1898–1901, Oct. 1996.

[5] J. Glass, J. Chang, and M. McCandless, "A Probabilistic Framework for Feature-Based Speech Recognition", in *Proc. Fourth International Conference on Spoken Language Processing* (Philadelphia), pp. 2277–2280, Oct. 1996.