An experimental dialogue system: Waxholm

Blomberg, M. and Carlson, R. and Elenius, K.
O. E. and Granström, B. and Gustafson, J.
and Hunnicutt, S. and Lindell, R. and Neovius,
L. and Nord, L.

KTH Computer Science and Communication

AN EXPERIMENTAL DIALOGUE SYSTEM: WAXHOLM*

*Mats Blomberg, Rolf Carlson, Kjell Elenius, Björn Granström, Joakim Gustafson, Sheri Hunnicutt, Roger Lindell, Lennart Neovius and Lennart Nord***

ABSTRACT

Recently we have begun to build the basic tools for a generic speech-dialogue system, WAXHOLM. The main modules, their function and internal communication have been specified. The different components are connected through a computer network. A preliminary version of the system has been tested, using simplified versions of the modules. We will give a general overview of the system and describe some of the components in more detail. Application specific data are collected with the help of Wizard-of-Oz techniques. The dialogue system is used during the data collection and the wizard only replaces the speech-recognition module.

1. THE DEMONSTRATOR APPLICATION

We are currently building a generic system in which speech synthesis and speech recognition can be studied in a man-machine dialogue framework. In addition, the system should facilitate the collection of speech and text data that are required for the development of the system. The demonstrator application, that we call WAXHOLM, gives information on boat traffic in the Stockholm archipelago (see Fig. 1). A fleet of some twenty boats from the Waxholm company connect about two hundred ports. Different days of the week have different time tables.

Besides the speech recognition and synthesis components, the system contains modules that handle graphic information such as pictures, maps, charts, and time-tables. This information can be presented to the user at his/her request. The application has great similarities to the ATIS domain within the ARPA community and other similar tasks in Europe, for example SUNDIAL. The possibility to expand the task in many directions is an advantage for our future research on interactive dialogue systems. An initial version of the system based on text input has been running since September 1992.

The dialogue system is implemented as a number of independent and specialized modules that run as servers on our computer system. A notation has been defined to control the information flow between them. The structure makes it possible to run the system in parallel on different machines and facilitates the implementation and testing of alternate models within the same framework. The communication software is based on UNIX de facto standards, which will facilitate the reuse and portability of the components.

INPUT

Speech

RECORD

RECOGNITION

WIZARD OF OZ — *Human interaction*

Words
Parts of speech
Semantic information
Pronunciation

LEXICON ⟷ GRAMMAR & SEMANTICS — *Context sensitive rules*
transition networks

DATA BASE ⟷ DATABASE SEARCH ⟷ DIALOG MANAGER — *Application specific*
modules

Ship schedules
Port locations
Lodging
Eating

GRAPHICS   SYNTHESIS   Sound

Charts Maps Tables   Speech   Recordings

OUTPUT
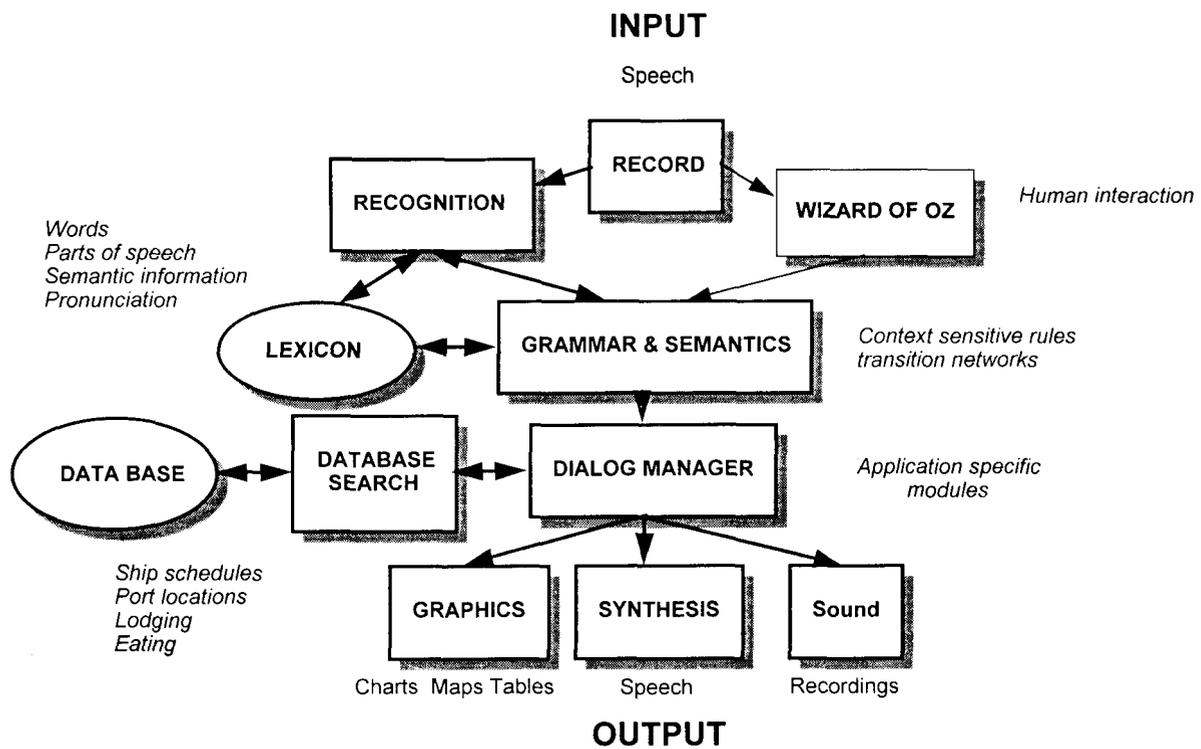
Figure 1. Block diagram of the demonstrator application Waxholm.

2. THE DATABASE

The database contains time tables, and also information about port locations, hotels, camping places, and restaurants. It is accessed by the standardized query language (SQL). The time table, which is the primary part of the database, comprises some inherent difficulties in our application.One is that a boat can go in "loops", i.e. it uses the same port more than once for departure or arrival. This has been solved by giving unique tour identification numbers to different "loops". Another problem is that the port Waxholm may be used as a "transit port" for many destinations, and transit tours are not included in the database to avoid redundancy. Transits are instead handled by recursively searching for tours from the departure port to Waxholm, and (backwards) from the destination port to Waxholm that are less than 20 minutes apart (Gustafson, 1993).

3. SPEECH RECOGNITION

The speech recognition component, which so far has not been integrated in the system, will handle continuous speech with a vocabulary of about 1000 words. The work on recognition has been carried out along two main lines: artificial neural networks and a speech production oriented approach. Neural nets are general classification tools and it is quite feasible to combine the two approaches.

3.1. Speech production approach

Our system uses a speech synthesis technique to generate spectral prototypes of words in a given vocabulary, see Blomberg (1991). A speaker-independent recognition

system has been built according to the speech-production approach, using a formant based speech-production module including a voice-source model. Whole word models are used to describe intra-word phonemes, while triphones (three-phoneme clusters) are used to model the phonemes at word boundaries. An important part of the system is a method of dynamic voic- source adaptation. The recognition errors have been significantly reduced by this method.

3.2. Artificial neural networks

We have tested different types of artificial neural networks for performing acoustic-phonetic mapping for speech signals, compare Elenius & Takács (1990); Elenius & Blomberg (1992), and Elenius & Tråvén (1993). The tested strategies include self-organizing nets and nets using the error-back propagation (BP) technique. The use of simple recurrent BP-networks has been shown to substantially improve performance. The self-organizing nets learn faster than the BP-networks, but they are not as easily transformed to recurrent structures.

4. SPEECH SYNTHESIS

For the speech-output component we have chosen the multi-lingual text-to-speech system developed in an earlier project (Carlson et al., 1991). The system will be modified in several ways for this application. The application vocabulary will be checked for correctness, especially considering the general problem of name pronunciation. Speaker-specific aspects are important for the acceptability of the synthetic speech. The WAXHOLM dialogue system will focus our efforts on modelling the speaking style and speaker characteristics of one reference speaker. Since the recognition and synthesis modules have the same need of semantic, syntactic and pragmatic information, the lexical information will, to a great extent, be shared. The linguistic module, STINA, will also be used for improved phrase parsing, compared to the simple function-word based methods that have been used so far in the synthesis project. However, in dialogue applications, such as the proposed WAXHOLM demonstrator, information on phrasing and prosodic structure can be supplied by the application control software itself, rather than by a general module meant for text-to-speech. In a man-machine dialogue situation we have a much better base for prosodic modelling compared to ordinary text-to-speech, since we, in such an environment, will have access to much more information than if we used an unknown text as input to the speech synthesizer.

5. NATURAL LANGUAGE COMPONENT

Our initial work on a natural language component is focused on a sublanguage grammar, a grammar limited to a particular subject domain: that of requesting information from a transportation database. This component provides syntactic and semantic knowledge to the recognizer.

Our aim is to develop a parser that is technically robust - a parser that is efficient and fast, that is statistically sound, and that fails gracefully. We are stressing an interactive development environment in order to have control over the system's progress as more components are added.

The fundamental concepts are inspired by TINA, a parser developed at MIT (Seneff, 1989). Our parser, STINA, i.e., Swedish TINA is knowledge-based and is designed as a probabilistic language model (Carlson and Hunnicutt, 1992). It contains a context-free grammar which is compiled into an augmented transition network (ATN). Probabilities are assigned to each arc after training. Features of STINA are a stack-decoding search strategy and a feature-passing mechanism to implement unification.

5.1. Lexicon

The lexicon entries are generated by processing each word in the Two-Level Morphology (TWOL) lexical analyser of Koskenniemi (1983) and Karlsson (1990). Each entry is then corrected by removing all unknown homographs. New grammatical and semantic features, which are used by our algorithm and special application, are then added.

5.2. Features

The basic grammatical features can be positive, negative or unspecified. Unspecified features match both positive and negative features.

Semantic features can be divided into two different classes. The basic features like BOAT and PORT give a simple description of the semantic property of a word. These features are hierarchically structured. Fig. 2 gives an example of a semantic feature tree. During the unification process in STINA all features which belong to the same branch are considered. Thus, a unification of the feature PLACE engage all semantic "non-shaded" features in Fig. 2.
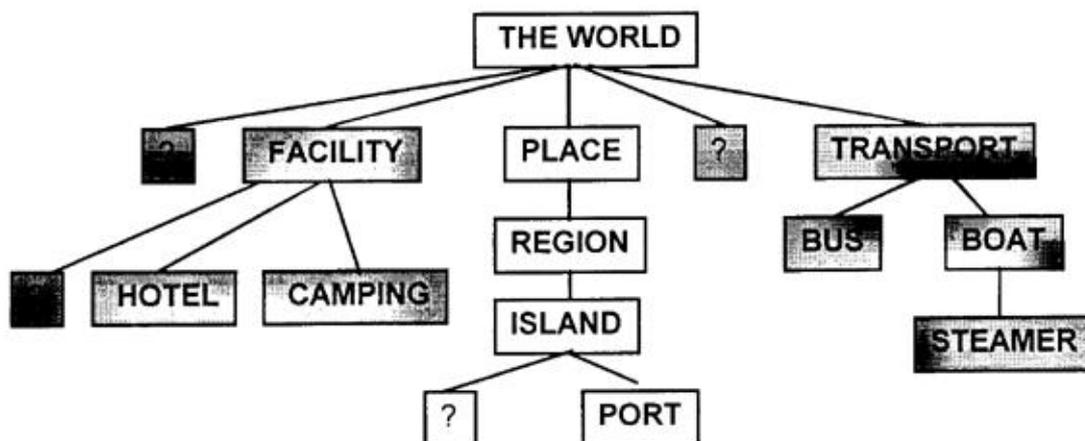
Figure 2. Example of a semantic tree feature structure.

Another type of semantic feature controls which nodes that can be used in the syntactic analysis. For example, the node DEPARTURE TIME can not be used in connection with verbs that imply an arrival time. This is a powerful method to control the analysis of responses to questions from the dialogue module. The question "Where

do you want to go?" conditions the parser to accept a simple port name as a possible response from the user.

6. DIALOGUE MANAGEMENT

Dialogue handling is also implemented in STINA making use of the lexical semantic features. Topic selection is done by a probabilistic approach that needs application-specific training. Thus, data collection is of great importance for the progress of the project. In Fig. 3 some of the major topics are listed. The decision on which path to follow in the dialogue is based on several factors, such as the dialogue history and the content of the specific utterance. The utterance is coded in the form of a "semantic frame" with slots corresponding to both the grammatical analysis and the specific application. Each semantic feature found in the syntactical and semantical analysis is considered in the form of a conditional probability to decide on the topic. The BOAT feature can be a strong indication for the TIME-TABLE topic but this can be contradicted by a HOTEL feature.

The dialogue will be naturally restricted by application-specific capabilities and the limited grammar. So far we also assume that the human subjects will be co-operative in pursuing the task. Recovery in case of human-machine "misunderstandings" will be aided by informative error messages generated upon the occurrence of lexical, parsing or retrieval errors. This technique has been shown to be useful in helping subjects to recover from an error through rephrasing of their last input (Hunnicutt et al., 1992) .

TIME-TABLE
    Goal: to get a time table presented with departure and arrival times specified
    between two specific locations.
    Example: När går båten? (When does the boat leave?)

GET-POSITION
    Goal: to get a chart or a map displayed with the place of interest shown.
    Example: Var ligger Vaxholm? (Where is Vaxholm?)

EXIST
    Goal: to display the availability of lodging and dining possibilities.
    Example: Var finns det vandrarhem? (Where are there hostels?)

REPEAT
    Goal: Repeat the synthesis
    Example: Jag förstår inte. (I do not understand)

Figure 3. Some of the main topics used in the dialogue.

7. CURRENT ACTIVITIES

7.1. Application-specific data collection

We are currently collecting speech and text data using the WAXHOLM system. Initially, a "Wizard of Oz" (a human simulating part of a system) is replacing the speech recognition module, Fig. 4. The user is placed in a sound-treated room in front of a terminal screen. The wizard sitting outside the room can observe the subject's screen on a separate display.

The user is initially requested to pronounce a number of sentences and digit sequences to practice talking to a computer. This material will be used for speaker adaptation experiments. After this the subject is presented with a task to be solved. The scenario is presented both as text and as synthetic speech. An advantage of this procedure is that the subject becomes familiar with the synthetic speech. During the data collection, utterance-size speech files are stored together with the transcribed text entered by the wizard.
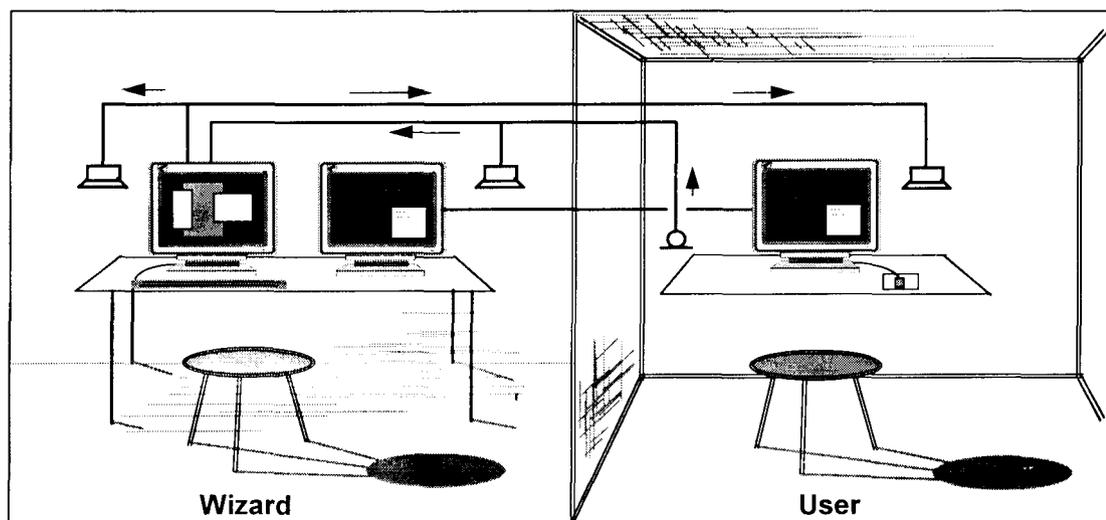
Figure 4. Hardware setup for data collection, with the help of a wizard.

7.2. Integration of speech recognizer

Currently the most important work beside data collection is the integration of the speech recognizer into the system. The interaction between the parser and the recognizer still has to be specified. Our initial efforts will be using an N-best approach using a bigram language model, while a more tight coupling between the parser and the recognizer is a long-term goal for our project.

7.3. Language modelling

The collected corpus is being used for grammar development, for training of probabilities in the language model in STINA, and also for generation of an application-dependent bigram model to be used by the recognizer. The collected text is also being used to train word collocation probabilities. Our plan is to replace explicit formulations of semantic coupling by a collocation probability matrix.

Robust parsing and unknown word processing have so far not been included in the data collection system. However, such facilities will play an important part in following versions of the system.

The parser itself has been expanded to better handle robust parsing and unknown word problems. We are currently testing a simple application-independent grammar on unlimited text. This system will also be used as part of our general text-to-speech system, which is outside the scoop of this presentation.

REFERENCES

Blomberg, M. (1991): "Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references," *Speech Communication,* Vol. 10. pp 453-462.

Carlson, R., Granström, B., & Hunnicutt, S. (1991), "Multilingual text-to-speech development and applications," in *Advances in speech, hearing and language processing,* (A. W. Ainsworth, ed.), JAI Press, London, UK.

Carlson, R., & Hunnicutt, S. (1992): "STINA: A probabilistic parser for speech recognition," in *Technical Report No. 10,* FONETIK'92, Sixth Swedish Phonetics Conference, May 20-22, 1992, Dept. of Information Theory, Chalmers University of Technology, Gothenburg, Sweden.

Elenius, K. & Blomberg M., (1992): "Experiments with artificial neural networks for phoneme and word recognition," *Proceedings of ICSLP 92,* Vol. 2, pp. 1279-1282, Banff.

Elenius K. & Tråvén H. (1993): "Multi-layer perceptrons and probabilistic neural networks for phoneme recognition", in *Proceedings of EuroSpeech '93,* Vol. 2, pp. 1237-1240. Berlin, Germany.

Elenius K. and Takács, G. (1990): "Acoustic-phonetic recognition of continuous speech by artificial neural networks", *STL-QPSR 2-3/90,* Technical Report, Dept. of Speech Comm., KTH, Stockholm, Sweden.

Gustafson, J. (1992): *"Databashantering som del av ett talförståelsesystem,"* Thesis work, Dept. of Speech Comm., KTH, Stockholm, Sweden. (Only available in Swedish).

Hunnicutt, S., Hirschman, L., Polifroni, J., & Seneff, S. (1992): "Analysis of the effectiveness of system error messages in a human-machine travel planning task," *Proceedings ICSLP 92,* Vol. 1, University of Alberta, Canada.

Karlsson, F. (1990): "A Comprehensive Morphological Analyzer for Swedish", manuscript, University of Helsinki, Department of General Linguistics.

Koskenniemi, K. (1983): "Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production", Publications No. 11, University of Helsinki, Department of General Linguistics.

Seneff, S. "TINA (1989): "A Probabilistic Syntactic Parser for Speech Understanding Systems," *Proceedings ICASSP-89,* pp. 711-714.