[*Progress in Machine Translation*, ed. Sergei Nirenburg. Amsterdam: IOS Press, 1993.]

Overview of Speech Translation at ATR

• AKIRA KUREMATSU

1 Introduction

A speech translation system will transform a spoken dialogue from the speaker's language to the listener's automatically and simultaneously. It will undoubtedly be used to overcome language barriers and facilitate communication among the peoples of the world. Creation of such a system will first require developing the various constituent technologies: speech recognition, machine translation, and speech synthesis. These individual subsystems will then be integrated to form a speech translation system, moreover an automatic telephone interpretation system. In Japan, the general view toward the research and development for an automatic telephone interpretation system was reported in 1986. It reported the necessity of long-term research and development in component technologies and software architecture. ATR Interpreting Telephony Research Laboratories was established in 1986 to initiate basic research for automatic telephone interpretation. Currently, ATR Interpreting Telephony Research Laboratories are engaged in research in the areas of continuous large-vocabulary Japanese speech recognition, integrated processing of speech and language, machine translation of spoken language from Japanese to English, and high-quality speech synthesis. Experimental spoken language translation system *(SL-TRANS)* was developed at ATR.

2 Overview of Experimental Speech Translation System

An experimental spoken language translation system *(SL-TRANS)* has been implemented to investigate the major problems inherent in the integration of speech and language processing at ATR (Morimoto et al. 1990; Kurematsu et al. 1991). The system includes a speaker adaptation, a speech recognition system *(HMM-LR),* a controller of phrase candidates, a machine translation system *NADINE,* and a speech synthesizer. The system configuration is shown in Fig. 1. Experiments have been conducted for goal-oriented dialogues on the domain of "conference registration." The number of vocabulary in the initial experiment is about

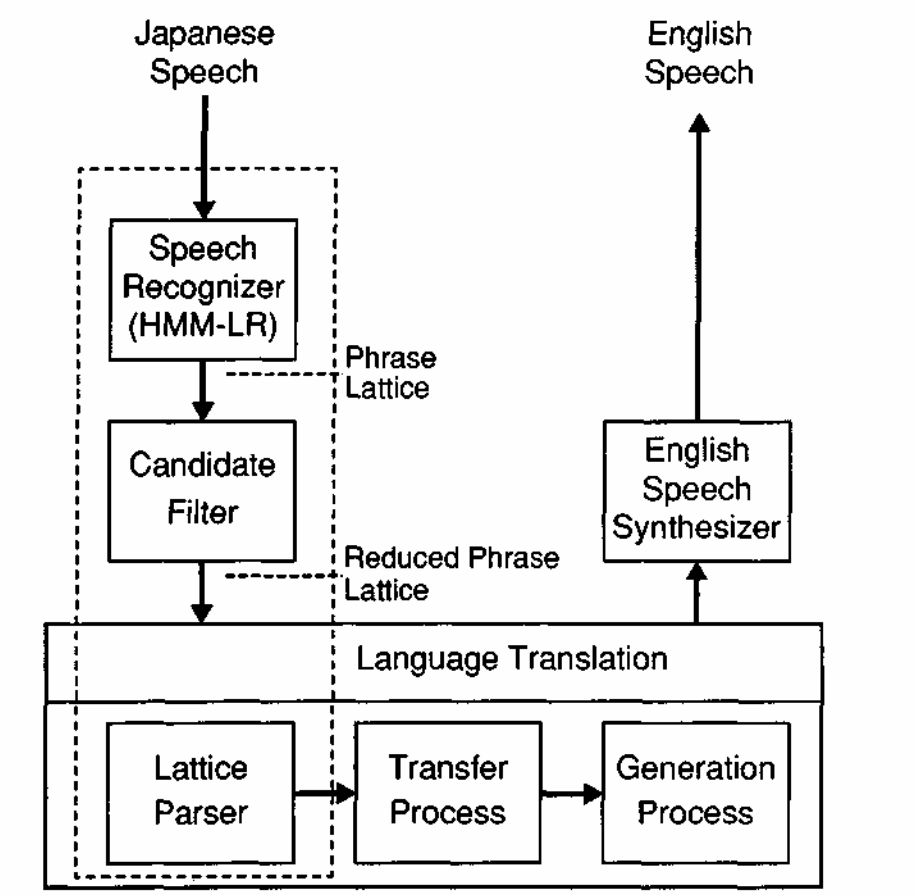

Figure 1: Configuration of Experimental Spoken Language Translation Systems (SL-TRANS)

400. It is going to be expanded to about 1500 which covers the basic dialogue expressions in Japanese.

The experimental translation system *(NADINE)* consists of three basic modules corresponding to processes in the intention translation method, i.e. analysis, transfer and generation modules. The analysis module consists of a phrase structure analysis module and a zero pronoun resolution module. The grammar consists of 20 generic phrase structure rules and about 400 lexical entries. The transfer process consists of the feature structure rewriting system and rules.

3 Speech Recognition

In the speech translation, large vocabulary speech recognition of continuous utterance is necessary. The size of vocabulary will be required at least more than a few thousands. For the recognition of large-vocabulary continuous speech, the problem of attaining high recognition rates must be overcome in order to lessen the burden of language processing. The

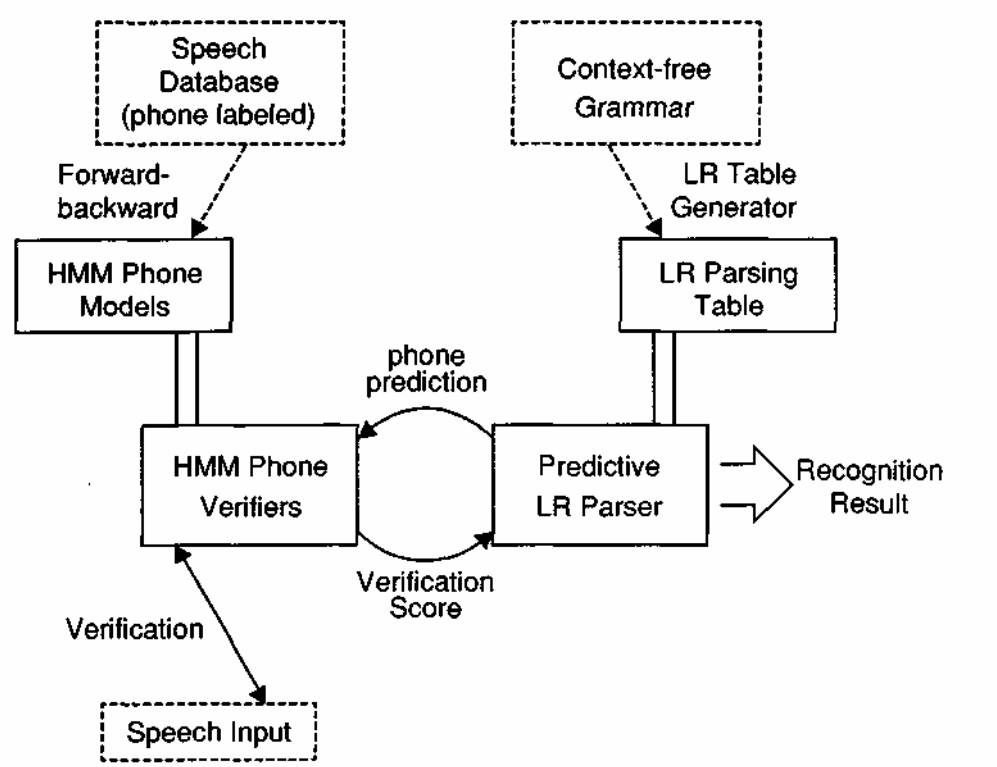

Figure 2: Schematic Diagram of HMM-LR Speech Recognizer

recognition of phrases or sentences are conducted though the recognition of phoneme.

3.1 HMM-LR Speech Recognition System

To treat speech of large vocabulary, Japanese phone models formalized in HMM have been introduced. Any word can be made by concatenating appropriate phone models. As a speech recognition mechanism, a grammar-based approach is adopted, that is, a parser is used to predict next possible phones from pre-compiled syntactical knowledge, then verifies their existence in the input data by comparison with HMM phone models. This process is continued until the input data ends.

The HMM phoneme models are integrated with the generalized LR predictive parser (Kita and Kawabata 1989), as shown in Fig. 2. The generalized LR parser, called LR parser for short, has been introduced to predict next word/phonemes. The HMM phoneme model is integrated with the generalized LR parser to efficiently recognize a Japanese phrase input. The LR parser was originally developed as a compiler, and then extended to handle arbitrary context-free grammar (Tomita 1986). The LR parser is considered the language source model in the speech recognizer. An LR parser, guided by an LR table automatically created from context-free grammar rules, proceeds from left to right without backtracking.

At each recognition stage, probabilities for the sequences of phonemes are calculated and only those candidates with high probability are kept (i.e., the beam search is performed). This probability array is attached to each node of the partial parsing tree. When the highest

probability in the array is below a certain threshold, the parsing tree is pruned and then re-pruned by a beam searching algorithm with a certain beam width at each phoneme recognition stage. At the end of the input data, several candidates which are grammatically accepted and kept higher probabilities are output as final candidates.

HMM-LR is effective for a large vocabulary with high perplexity and is processed quite efficiently. For ease of recognition, we are adopting a method which recognizes continuous speech. The syntax of the phrases includes a general syntax structure of Japanese phrases, whose phoneme perplexity is about five. Assuming that the average word phoneme length is three, their perplexity is more than one hundred. The grammar is designed to cover linguistic expressions common in Japanese. The number of different words is 1,035. For the speaker-dependent case, the phrase recognition rate is about 89%.

Two-level *HMM-LR* using two-level LR parsing has been developed to recognize continuous speech sentence (Kita et al. 1991). This method makes it possible to use both intra- and inter-phrase grammatical constraints during speech recognition.

3.2 Speaker Adaptation

As an effective approach to the problem of speaker independence, speaker adaptation has been taken. As a means of spectral pattern learning for speaker adaptation, one promising approach is codebook mapping. Discrete spectrum space representation by vector quantization makes it possible to realize sophisticated speaker adaptation by codebook mapping.

This algorithm realizes general speaker adaptation which does not depend on speech recognition systems. Twenty five words will be enough to adapt to speaker characteristics. A speaker adapted phrase recognition was experimented by use of HMM phone model. The average recognition rates were 81.6% for the top candidate and 98.0% for the top 5 candidates for 1035 words with phone perplexity of 5.9 (Hanazawa et al. 1990).

4 Integration of Speech and Language Processing

Noisy and ambiguous output of the speech recognition has to be processed in the language processor. Based on the idea of the language source model, it carries out top-down prediction to the speech recognizer. The bottom-up results contain multiple candidates which must be narrowed down by the use of linguistic constraints and various knowledge information. Fig. 3 shows a block diagram of the proposed method for speech translation from Japanese to English.

In language processing systems, a function is necessary which can use syntactic and semantic knowledge to select the most appropriate candidate. The Japanese co-occurrence dependency relationship among phrases of HMM-LR candidates is used (Kakigahara and Morimoto 1989). In the speech recognition stage, syntactical knowledge for phrases (hereafter, a phrase means Japanese Bunsetsu[1]) is used. About 60 kinds of semantic relationships are defined and attached to each phrase in the text database. From this database, a possible phrase relationship and its frequency for two phrases are extracted. Using this information, only probable phrase candidates are selected from the HMM-LR output and the candidates with no phrase relationship on other candidates are discarded. At the final stage, sentence

[1] A Bunsetsu is a grammatical and phonological unit in Japanese. It consists of an independent-word such as noun, verb or adverb followed by a sequence of zero or more dependent-words such as auxiliary verbs, postpositional particles or sentence final particles.

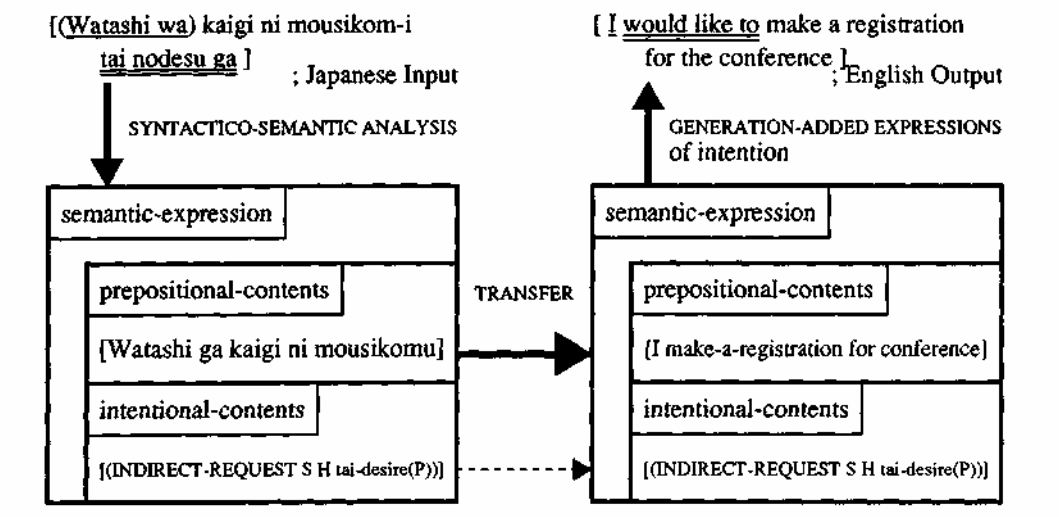

Figure 3: Block Diagram of Proposed Speech Translation System

analysis, the most plausible sentence is selected by checking strict syntactico-semantic and pragmatic appropriateness or by evaluating the preference of sentence structure. Preliminary experiments indicated the usefulness of this method: the number of candidates was reduced to less than one-third of the number of raw candidates.

5 Machine Translation for Spoken Dialogue

Major linguistic phenomena peculiar to Japanese spoken dialogues have been investigated from a linguistic viewpoint in order to construct a discourse-dialogue model. Spoken Japanese sentences contain certain inherent ambiguities, especially in the distribution of zero pronouns and the construction of predicate phrases. In order to disambiguate them, pragmatic constraints on the uses of expressions must be extracted, and the most plausible analysis candidate selected by using these constraints. The current approach to analyzing Japanese dialogues is based on a lexico-syntactic grammar framework in terms of typed feature structures and an analysis order controllable parser.

5.1 Intention Translation

The intention translation method for spoken dialogues based on the semantic transfer approach has been proposed (Kogure et al. 1989). It can be characterized by its two translating processes: one which extracts intentions in utterances, such as request, promise, greetings, etc., and the other which transfers propositional parts of utterances. The outline of the transfer process is shown in Fig. 4. Two different types of representations, propositional contents and the expressions of intention, are translated separately.

Hearing the speaker's utterance, the hearer receives communicative signs in addition to a propositional content. According to the speech act theory, these signs are classified as illocutionary forces governed by certain felicity conditions. Illocutionary forces can be useful to machine translation if propositional content is distinguished from the structure in

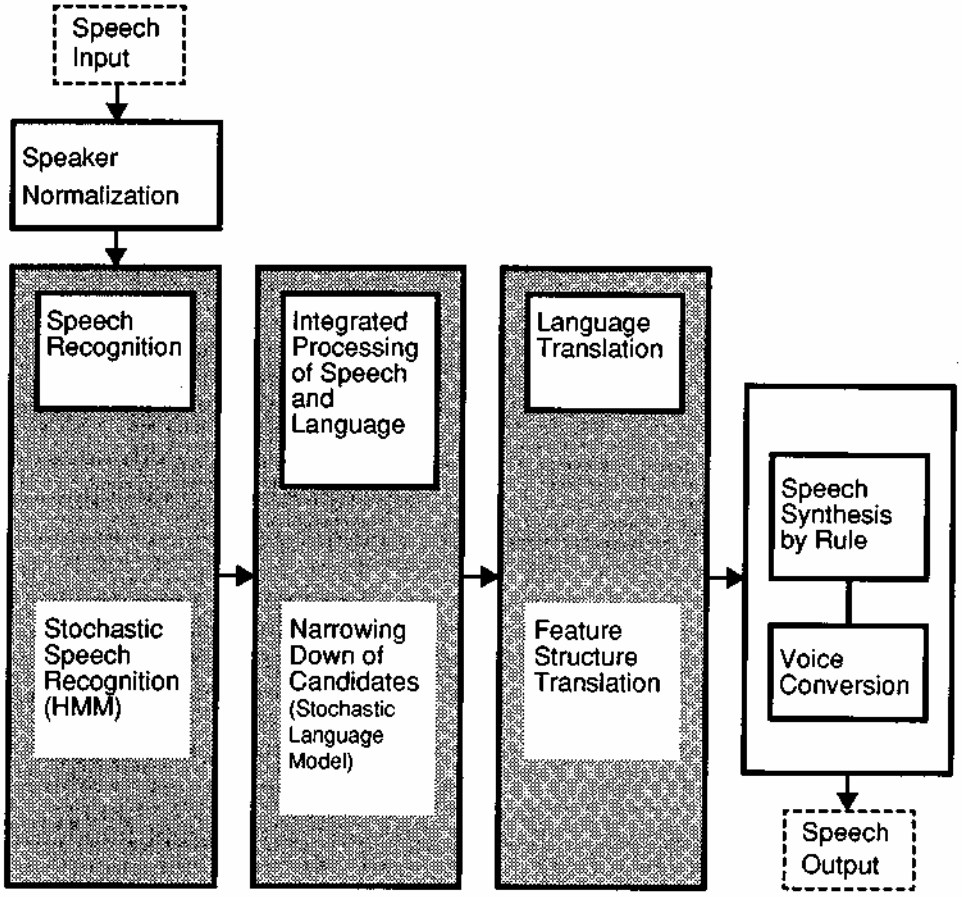

Figure 4: Outline of Dialogue Translation Method

the utterance analysis. There are many conventional forms for illocutionary forces which concern certain explicit expressions.

Japanese sentences in dialogues often have complex expression in a final predicate phrase which consists of one main predicate (e.g., verb, adjective, etc.) and combination of several auxiliary verbs and sentence final particles. These phrases are important for expressing intentions of the speaker. The exact meaning (semantics) of the speaker's intention is calculated compositionally from this part, and then its illocutionary force type (IFT) is determined.

The utterance analysis process extracts the values of output phrase structure analysis results as the semantic feature. Semantic representations consisting of surface illocutionary force relationships in terms of language-independent concepts and propositional content parts in terms of language-dependent concepts.

Most of the information used in these processes is described in terms of feature structures and is modified by the unification operation. Representation formalism using feature structures has various advantages over other representation formalisms. One of the most significant advantages is that they permit integrated descriptions of information from various

kinds of information sources such as syntax, semantics and pragmatics.

5.2 Unification-Based Utterance Analysis

Unification calculation permits integrated descriptions of information from various sources. In analyzing spoken utterances such as fragmental and various intentional phrases, constraints between syntax, semantics and pragmatics can be described in terms of feature structures.

A complement-head grammatical structure based on Head-Driven Phrase Structure Grammar *(HPSG)* is adopted to implement the unification-based approach (Gunji 1987). In this approach, a grammar has a small number of general phrase structure rules corresponding to sets of grammatical principles, and most of the grammatical information must be specified in descriptions of lexical items.

Abundant linguistic information is defined in lexical items as a set of feature structures. In grammar specification, the context free grammar (CFG) rule and the constraints that the elements in the CFG should satisfy are defined. These constraints are described as a set of path equations between feature structures.

In this schema, not only general grammatical constraints such as the HEAD feature principle or the SUBCAT feature principle, but also semantic and/or pragmatic constraints, can be specified. At the same time, the semantics calculation formula is specified declaratively as a path equation (Nagata and Kogure 1990).

Many ellipses can be resolved by using pragmatic felicity conditions such as honorific expressions (Yoshimoto 1988). The utterance analysis in the translation process consists of phrase structure analysis and zero pronoun resolution sub-processes. The zero pronoun resolution process takes the output structures and fills in the semantic information corresponding to the zero pronouns by matching pragmatic constraints with dialogue-participant information such as social relationships (Dohsaka 1990).

5.3 Utterance Transfer

The transfer process of the machine translation system accepts an analysis result represented by a feature-based semantic representation. The transfer process uses a feature structure rewriting system which rewrites input feature structures according to feature structure rewriting rules. In order to maintain high modularity and flexibility in this transfer process, a rewriting environment has been introduced into the rewriting system and constraints have been attached to each rewriting rule instead of describing rule dependencies explicitly. Two different types of representations, propositional contents and intention expressions, are translated separately by a feature rewriting process. The propositional contents are taken into account as language-dependent concepts. On the other hand, intention expressions consisting of an illocutionary force are considered to be language-independent representation and can be passed through the transfer process.

Propositional contents are represented by recursively defined relationships. Such a relationship consists of a relationship name and its case roles. The transfer process converts such relationships in terms of source language concepts (e.g. concepts in Japanese) into corresponding relationships in terms of target language concepts (e.g. concepts in English). The load of this transfer process is small compared to the load in the traditional syntactic transfer approach because this process does not require many transfer rules for treating syntactic information.

5.4 Utterance Generation

The generation process has two phases: the first is to make up a target description on a surface linguistic structure concerning a propositional content and intentional contents. An illocutionary force type is used to determine an utterance form and the propositional content part is expanded to a sequence of its predicate and case role constituents. The second is to spell out the description according to the rules of the target language writing style and morphology.

Feature structure-directed generation is useful for a bi-directional grammar in analysis and generation. An appropriate auxiliary phrase is added if necessary. Control of the generation process has been studied by selecting appropriate rules to apply. Typed feature structures are utilized to describe the control of the generation process in a declarative way. The disjunctive feature structure is introduced to solve the inefficiency in making multiple copies of the phrase structure when the generation process encounters multiple rule candidates.

6 Further Research

An extensive effort must be made to raise the level of technology of speech recognition, machine translation, and speech synthesis if automatic interpretation of telephone conversations is to be realized. Further research is now being directed to the following points.

The capability of recognizing large vocabulary continuous speech should be further enhanced. In order to obtain better phoneme recognition performance, a scheme to integrate current approaches of applying relevant knowledge about speech will be necessary to be formulated. In speaker-independent recognition, a method applicable to large-vocabulary continuous speech recognition must be explored by using a large scale speech database. Prosodic information such as pitch, stress, and duration, along with information on phrase boundaries, will be useful in order to increase the precision and speed of algorithms for phrase recognition.

In the integration of speech and language processing, a scheme to predict the level of words or utterances will have to be explored. The introduction of the statistical characteristics of grammar or statistical constraints on the input which takes the form of estimates of the probability of a particular sequence of words will be effective to reduce the perplexity. Further, utilization of higher level information such as dialogue structure will be investigated. Knowledge about language itself and domain-specific extralinguistic knowledge will be formulated as the common base for various aspects of spoken language interpretation.

In machine translation, the enrichment of grammar and lexical dictionaries to an advanced level will be carried out to cope with large vocabulary translation. The challenge will be directed to the general methodologies which can be expanded to large vocabularies and various task domains. Considering the requirement of real-time processing, a high speed computational scheme will be researched to shorten the considerable existing gap between theoretical computational linguistics and software implementation. To enhance deep understanding, translations based on context processing will be explored. Experimental speech translation system will be developed with the capability of larger vocabulary and wider linguistic expressions.

In speech synthesis, speech synthesis by rule will be enhanced to obtain more natural speech quality in conversational sentences. The linguistic information in language generation will be reflected to the control of rule in speech synthesis. Voice individualization over different languages will be developed.

Massive databases of speech and language corpora are essential and extensive efforts will be continued Because of the vast complexity of natural language, however, the goal should be reached by gradually improving levels and techniques. Also, it will be necessary to consider the expandability of the system in terms of domain size, different domain applications, multi-language application.

References

Dohsaka, K. 1990. Identifying the referents of zero-pronouns in Japanese based on pragmatic constraint interpretation. In *Proc. of ECAI 90*, 240-245.

Gunji, T. 1987. *Japanese phrase structure grammar*. D.Reidel.

Hanazawa, T., K.Kita, S.Nakamura, T.Kawabata and K.Shikano. 1990. ATR HMM-LR continuous speech recognition system. In *Proc. of ICASSP'90*

Kakigahara, K. and T.Morimoto. 1989. A method of bunsetsu candidate selection using Kakari-uke semantic relationships. In *Proc. Fall Meeting of Acoust. Soc. Japan* (in Japanese)

Kita, K. and T.Kawabata. 1989. HMM continuous speech recognition using predictive LR parsing. In *Proc. ICASSP'89*, 703-706.

Kita, K., T.Takezawa and T.Morimoto. 1991. Continuous speech recognition using two-level LR parsing. *Trans. of IEICE*, **74-E**(7): 1806-1810.

Kogure, K., H.Iida, T.Hasegawa and K.Ogure. 1989. NADINE: an experimental dialogue translation system from Japanese to English. In *Proc. InfoJapan'90*, volume 2, 57-64.

Kurematsu, A., H.Iida and T.Morimoto. 1991. Language processing in connection with speech translation at ATR Interpreting Telephony Research Laboratories. *Speech Communication*, **10**(1): 1-9

Morimoto, T., H.Iida, A.Kurematsu, K.Shikano and T.Aizawa. 1990. Spoken language translation: towards realizing an automatic telephone interpretation. In *Proceedings of the INFO JAPAN 90: International Conference of the Information Processing Society of Japan*, 553-559.

Nagata, M. and K.Kogure. 1990. HPSG-based lattice parser for spoken Japanese in a spoken language translation system. In *Proc. of ECAI 90*, 461-466.

Tomita, M. 1986. *Efficient parsing for natural language*. Kluwer Academic Publishers.

Yoshimoto, K. 1988. Identifying zero pronouns in Japanese dialogues. *In Proceedings of the 12th Int. Conf. on Computational Linguistics.*