JANUS 93: TOWARDS SPONTANEOUS SPEECH TRANSLATION

M.Woszczyna *N.Aoki-Waibel* *F.D.Buø* *N.Coccaro* *K.Horiguchi* *T.Kemp* *A.Lavie*
A.McNair *T.Polzin* *I.Rogina* *C.P.Rose* *T.Schultz* *B.Suhm* *M.Tomita* *A.Waibel*

Carnegie Mellon University — USA
University of Karlsruhe — Germany

ABSTRACT

We present first results from our efforts toward translation of spontaneously spoken speech. Improvements include increasing coverage, robustness, generality and speed of JANUS, the speech-to-speech translation system of Carnegie Mellon and Karlsruhe University. Recognition and Machine Translation Engine have been upgraded to deal with requirements introduced by spontaneous human to human dialogs. To allow for development and evaluation of our system on adequate data, a large database with spontaneous scheduling dialogs is being gathered for English, German and Spanish.

1. OVERVIEW

JANUS [1, 2] has been among early systems to attempt the translation of spoken dialogs. It had initially been built based on a speech database of 12 read dialogs of the conference registration task, encompassing a vocabulary of around 500 words. It was designed as a speaker-independent system which translates spoken utterances from English and also from German into one of German, English or Japanese.

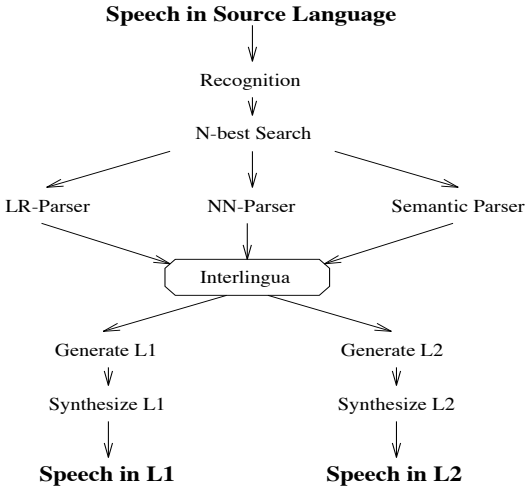

Figure 1. Overview of the System

In cooperation with partner efforts at ATR [3] and Siemens, feasibility and potential of multilingual speech translation on limited task has been demonstrated by a public demonstration in spring 1993. Independently, other speech translation systems [4, 5] have been presented, showing the growing interest in the field.

To begin extending our system to spontaneous human-to-human dialogs, however, improvements and changes along several dimensions of our earlier system are necessary to increase speed, robustness and coverage of the system in the face of ill-formed and ungrammatical spontaneous input. In the following, we will report on the state of these most recent efforts.

2. THE SCHEDULING TASK DATABASE

We have started collecting a large database of human to human dialogs centered around the scenario of appointment scheduling. In each recording session, two subjects are each given a calendar (one of 13 scenarios) and asked to schedule a meeting with the dialog partner. The recording setup allows only one person to speak at a time by way of a push-to-talk switch. Data is being collected in similar fashion in German, English and Spanish using the same setup at Karlsruhe University and Carnegie Mellon.

Dialogs	recorded	transcribed
German	450	150
English	200	140
English (phone-quality)	70	70
Spanish	75	30

Table 1. The Spontaneous Scheduling Task Database

On average, the resulting dialogs cover about 10-12 utterances, each up to 50 seconds long. The test set perplexities using smoothed bigrams on an initial set of dialogs were found to be around 45 for English and 88 for German; the vocabulary size (Fig. 2) for German is about 40% larger than for English, mostly due to the larger number of inflections and compound words.

After recording, the dialogs are transcribed following the same conventions at all sites. The transcription format covers a set of spontaneous effects like mispronunciations, restarts, human and nonhuman noises as well as pauses.

A first training and evaluation set for English and German consists of about 90 dialogs for training and 20 dialogs set apart as development test set. 20-40 new dialogs are reserved as independent evaluation set.

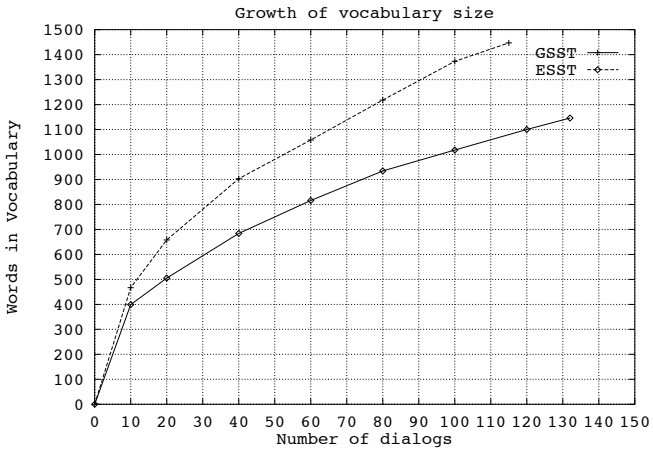

Figure 2. Development of Vocabulary Size

3. RECOGNITION ENGINE

3.1. Acoustic Modeling

For acoustic modeling, several alternative algorithms are being explored including TDNN, MS-TDNN, MLP and LVQ [6, 7]. In the main JANUS system, an LVQ algorithm with context-dependent phonemes is currently used for speaker independent recognition. For each phoneme, there is a context independent set of prototypical vectors. The output scores for each phoneme segment are computed from the euclidean distance using context dependent segment weights.

Recent changes include the introduction of noise models as well as the improvement of the training algorithms; the 1993 results in table 2 were obtained using triphone clustering, corrective training and feature weights. Further improvements using cross word triphones are possible but not evaluated here.

	1991	1993
Conference Registration	9.1 %	3.7 %
Resource Management	24 %	7.5 %

Table 2. Comparison of error-rates

3.2. Search

The search module of the recognizer builds a sorted list of sentence hypotheses. Speed and memory requirements have been improved considerably (table 3).

	1991	1993
Time for N-best	3-5 min	3-10 sec
Memory usage	50Mbytes	7Mbytes
Maximum Vocabulary Size	500	10000
Maximum Perplexity	5-10	120

Table 3. Recent Improvements of the Search Module. Timings and memory usage are given for the English CR Task.

This was achieved by using the word dependent N-best algorithm [8] as a backward pass in the forward backward pruning algorithm.

Recent experiments with the Wall Street Journal Task show, that the recognizer can handle vocabularies of up to 10000 words at a perplexity of 110.

3.3. Language Models

The most successful language model so far used for the scheduling task is a model that interpolates a cluster based model with smoothed bigrams. The cluster based model is using automatically build classes. The optimization criterion is to minimize the leaving-one-out perplexity on the training set [9]. The results shown in table 4 were obtained using 150 word classes.

	ESST	GSST
Smoothed Bigrams	46.8	88.2
Cluster	42.3	71.8
Interpolated	39.2	62.1

Table 4. Perplexity Reduction due to automatic clustering

3.4. Between Word Noise Models

Noises, filled pauses, and restarts, are much more frequent in spontaneous speech compared to read speech[1], particularly as our domain involves human-to-human negotiation:

Noise	**English**	**German**
human		
breathing	2166	1017
lip smack	992	144
laugh	36	41
hm	21	9
mmm	—	14
ähm	—	75
eh	—	136
um	546	—
uh	514	—
glottal	92	—
ah	49	—
oh	47	—
non-human		
key click	1028	477
paper rustle	148	13
finger hitting headset	50	13
copier noise	20	—
pen tap	35	—
silence		
1 to 2 seconds	547	385
longer than 2 seconds	18	23

Table 5. List of frequent noises

We have chosen 10 noise classes that are modeled by specialized phoneme models. Initially, we used one such model for each of the seven most frequent single noises, one for all remaining human noises, one for all remaining non-human

[1]The total number of utterances used for the ESST noise statistic is about twice the number used for the GSST statistic.

noises and a special model to deal with stutter and short false starts. Clustering these models into 6 generalized noise models results in further improvement. The total relative error reduction due to the introduction of between word noise models was 17% on the English Spontaneous Scheduling task.

4. THE MACHINE TRANSLATION (MT) ENGINE

The MT-component that we have previously used has now been replaced by a new module that can run several alternate processing strategies in parallel. In translating spoken language from one language to another, the analysis of spoken sentences which suffer from ill-formed input and recognition errors is most certainly the hardest part. Based on the list of N-best hypotheses delivered by the recognition engine, we can now attempt to select and analyze the most plausible sentence hypothesis in view of producing an accurate and meaningful translation.
Two goals are central in this attempt: *high fidelity* and *accurate translation* wherever possible, and *robustness* or *graceful degradation* in face of ill-formed or misrecognized input. At present, three parallel modules attempt to address these goals: 1) an LR-parser 2) a semantic pattern based approach and 3) a connectionist approach. The most useful analysis from these modules is mapped onto a common Interlingua, a language independent, but domain-specific representation of meaning.

4.1. Robust GLR Parser

The first step of the translation process is parsing with the Generalized LR Parser/Compiler. It can use syntactic or semantic based grammars. For application to the spontaneously spoken English scheduling task, we found semantic based grammars most useful.

The Generalized LR parsing algorithm is an extension of LR parsing with a "Graph-Structured Stack" [10], and it can handle arbitrary context-free grammars while most of the LR efficiency is preserved.

We use a recently developed robust version of the GLR parser to parse the input sentence. The most important feature of the robust parser is a capability to skip words of the input in cases where the complete input sentence is not grammatical. Using a beam search technique, the parser attempts to detect and parse the grammatical subset of the input sentence with the fewest skipped words. It thus returns a parse for any sentence, unless no part of the sentence can be considered grammatical. If the complete input sentence is itself grammatical, the parser behavior is identical to that of the standard GLR parser.

On a first experiment on the English Spontaneous Scheduling Task, this parser achieved 40% error free parses on unseen text, using a semantic based grammar.

4.2. The Interlingua

The current output of the parser is an interlingua representation, that could be refined by a discourse plan tracker.

Figure 3 is an example of interlingua representation (ILT) produced from the sentence "twenty (pause) actually July twenty sixth and twenty seventh looks good". In the example, the sentence is represented as speech-act *SUGGEST-TIME. Other typical speech-acts for this task are *STATE-CONSTRAINT, *AFFIRM and *REQUEST-RESPONSE;

The interlingua ensures that alternate parsing modules can be applied in a modular fashion and that different output languages can be added without redesign of the analysis stage. It also allows the separate evaluation of parser and generator, by matching against and generating from a set of reference interlingua representations.

```
(TWENTY ACTUALLY JULY TWENTY SIXTH AND
                      TWENTY SEVENTH LOOKS GOOD $)

((SPEECH-ACT *SUGGEST-TIME)
     (SENTENCE-TYPE *STATE)
     (FRAME *FREE)
     (WHEN
       ((FRAME *TIME-LIST) (CONNECTIVE AND)
                           (ITEMS
                             (*MULTIPLE*
                               ((FRAME *SIMPLE-TIME)
                                (DAY 26)
                                (MONTH 7))
                               ((FRAME *SIMPLE-TIME)
                                (DAY 27))))))
     (ADVERB ACTUALLY))
```

Figure 3. Example for Interlingua Output

4.3. Semantic Pattern Based Parsing

Our robust semantic parser combines frame based semantics with semantic phrase grammars [12]. We use a frame based parser similar to the DYPAR parser used by *Carbonell, et al.* to process ill-formed text [11], and the MINDS system previously developed at CMU. Semantic information is represented in a set of frames. Each frame contains a set of slots representing pieces of information. In order to fill the slots in the frames, we use semantic fragment grammars. The operation of the parser can be viewed as "phrase spotting". A beam of possible interpretations are pursued simultaneously. An interpretation is a frame with some of its slots filled. Each slot type is represented by a separate Recursive Transition Network, which specifies all ways of saying the meaning represented by the slot. The grammar is a semantic grammar, non-terminals are semantic concepts instead of parts of speech.

4.4. Connectionist Parsing

The connectionist parsing system PARSEC [13] is used as a fall-back module if the LR parser fails to analyze the input. One important aspect of the PARSEC system is that it learns to parse sentences from a corpus of training examples. This eliminates the very difficult work of writing robust grammars. Another aspect is that it has proven robust towards spontaneous utterances which frequently are "corrupted" with disfluencies, restarts, repairs or ungrammatical constructions. Third, integration with other information sources, e.g intonation, is easier.
More information about the recent developments in PARSEC can be found in [14]

4.5. The Generator

The generation of target language from an Interlingua representation involves two steps. First, with the same Transformation Kit used in the analysis phase, Interlingua representation is mapped into syntactic f-structure of the target language. The f-structure is then fed into sentence generation software called "GENKIT" to produce a sentence in the target language. A grammar for GENKIT is written in the same formalism as the Generalized LR Parser: phrase structure rules augmented with pseudo unification equations.

As a first experiment for generation on the spontaneous scheduling task, we tried Japanese generation on 264 new hand coded ILT's. More than 75% of the generated sentences were found to be good or acceptable (see table 6).

output quality	%
good	65.2
acceptable	11.0
bad	4.5
no output	19.3

Table 6. Quality of Japanese Generation

5. CONCLUSION

In this paper we have described a number of system improvements and extensions that have recently been introduced in JANUS, to accomodate extension of the speech translator to spontaneously spoken (not read) human-to-human dialogs.

A database of spontaneous negotiation dialogs is being collected in German, English and Spanish, and first results of system components on this data have been reported.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the collaboration with ATR Interpreting Telecommunications Laboratory, NEC, and Siemens, that made this research possible.

REFERENCES

[1] M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T .Polzin, I. Rogina, C.P. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel, and W. Ward, *Recent Advances in Janus, a Speech to Speech Translation System*, EUROSPEECH 1993.

[2] L. Osterholtz, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, and M. Woszczyna, *Testing Generality in JANUS: A Multi-Lingual Speech to Speech Translation System*, ICASSP 1992, volume 1, pp 209–212.

[3] T. Morimoto, T. Takezawa, F. Yato, S. Sagayama, T. Tashiro, M. Nagata, and A. Kurematsu, *ATR's Speech Translation System: ASURA*, EUROSPEECH 1993, pp 1295–1299.

[4] D.B. Roe, F.C.N Pereira, R.W. Sproat and M.D. Riley, *Efficient Grammar Processing for a Spoken Language Translation System*, ICASSP 1992, volume 1, pp 213–216.

[5] M. Rayner et al., *A Speech to Speech Translation System Built from Standard Components*, ARPA HLT Workshop Proceedings, March 1993, Session 6.

[6] J. Tebelskis and A. Waibel, *Performance through consistency: MS-TDNNs for large vocabulary continuous speech recognition*, Advances in Neural Information Processing Systems, Morgan Kaufmann.

[7] I. Rogina and A. Waibel, *Learning State-Dependent Stream Weights for Multi-Codebook HMM Speech Recognition Systems*, ICASSP 1994.

[8] S. Austin and R. Schwartz, *A Comparison of Several Approximate Algorithms for Finding N-best Hypotheses*, ICASSP 1991, volume 1, pp 701–704.

[9] R. Kneser and H. Ney, *Improved Cluster Techniques for Class-Based Statistical Language Modeling*, EUROSPEECH 93.

[10] M. Tomita (ed.), *Generalized LR Parsing*, Kluwer Academic Publishers, Boston MA, 1991.

[11] J.G. Carbonell and P.J. Hayes, *Recovery Strategies for Parsing Extragrammatical Language*, Carnegie-Mellon University Computer Science Technical Report 1984, (CMU-CS-84-107).

[12] W. Ward, *Understanding Spontaneous Speech*, DARPA Speech and Natural Language Workshop 1989, pp 137–141.

[13] A.J. Jain, A. Waibel and D. Touretzky, *PARSEC: A Structured Connectionist Parsing System for Spoken Language*, ICASSP 1992, volume 1, pp 205–208.

[14] F.D. Buø, T. Polzin and A. Waibel *Learning Complex Output Representations in Connectionist Parsing*, ICASSP 1994.