JANUS-II — TRANSLATION OF SPONTANEOUS CONVERSATIONAL SPEECH

A.Waibel M.Finke D.Gates M.Gavaldà T.Kemp A.Lavie L. Levin M.Maier L.Mayfield A.McNair I.Rogina K.Shima T.Sloboda M.Woszczyna T.Zeppenfeld P.Zhan

INTERACTIVE SYSTEMS LABORATORIES
at Carnegie Mellon University,USA
and University of Karlsruhe, Germany

ABSTRACT

JANUS-II is a research system to design and test components of speech-to-speech translation systems as well as a research prototype for such a system. We will focus on two aspects of the system: 1) new features of the speech recognition component JANUS-SR, 2) the end-to-end performance of JANUS-II, including a comparison of two machine translation strategies used for JANUS-MT (PHOENIX and GLR*).

1. INTRODUCTION

Currently JANUS-II components for English, German, Korean, Japanese, and Spanish speech input and translation are under development; though not all language pairs can always be kept at the same performance level, multilinguality is required to ensure generality in the recognition and translation approaches. A number of smaller and larger scale research projects contribute to the JANUS-II system [1], including language identification [2], robust speech recognition [3], recognition speed [4], noise modeling [5], new word modeling [6], portability to new languages [7], language modeling [8], user interfaces and repair strategies [9], interfaces between speech recognition and translation [10], machine translation issues [11], discourse modeling and software enginering. Covering all of them would go beyond the scope of a conference paper. We can therefore only focus on some selected aspects of the system. For general descriptions of other parts of the recognizer and the GLR* and PHOENIX parsers please refer to the list of references at the end of the paper.

2. THE SCHEDULING TASK DATABASE

JANUS is currently built around and evaluated on an appointment scheduling Task. We are collecting a large database of human-to-human dialogs centered around that scenario for English, German, Korean, Japanese and Latin-American Spanish. The collection sites are Carnegie Mellon University, the University of Pittsburgh, and Multicom (USA), Karlsruhe University (Germany), ETRI (Korea), UEC and ATR (Japan); additional German data collected at the Universities of Bonn, Kiel, Hamburg and München is available through the *VerbMobil* project sponsored by the BMBF. In each recording session, two subjects are each given a calendar and asked to schedule a meeting with the dialog partner. For most recordings the recording setup allows only one person to speak at a time by way of a push-to-talk switch and close-speaking microphones. Some of the Spanish data is recorded without push-to-talk system and includes crosstalk.

		utterances	hours
English	ESST	12430	40.3
German	GSST	12292	30.5
Japanese	JSST	6600	16.0
Korean	KSST	4395	10.2
Spanish	SSST	5730	10.7

Table 1. Spontaneous Scheduling Data currently used for developing JANUS-II

These dialogs were computed in a human-to-human setup rather than a human-machine-human situation. The average number of phonemes per word is only 2.9 for ESST (compared to 4.2 for both ATIS and WSJ); this makes the task more difficult for speech translation since speech recognition is harder on shorter words and the number of ambiguities increases with the number of words in the sentence. Some dialogs were recorded using the current JANUS setup. They have shorter utterances, longer words, less ambiguity and clearer articulation.

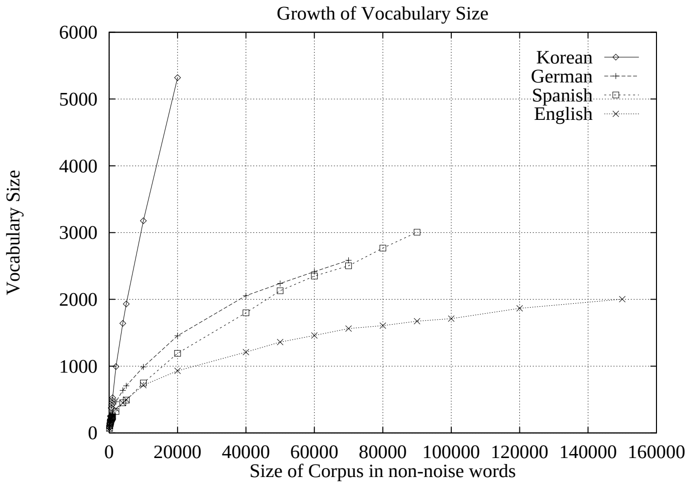

Figure 1. Development of Vocabulary Size

The steep vocabulary growth for Korean is due to the definition of a word unit: for Korean and Japanese the natural unit is similar to a phrase and it cannot be split to smaller units in a straightforward way. To make Korean and Japanese accessible to speech recognition and transla-

Published in "Proceedings of ICASSP-96"

tion word-like morphological units and decompositions are currently explored.

3. THE RECOGNITION ENGINE

The recognizer used in the current JANUS-II prototype system is a CDHMM based recognizer. The exact configuration varies from task to task. For the last *VerbMobil* evaluation on German scheduling dialogs we used the following preprocessing: on a 7 frame melscale window an LDA was computed to reduce the amount of parameters to 32. These were organized as two 16 coefficient streams. For acoustic modeling we used CDHMM's with 70 codebook vectors per codebook and top-all evaluation. Important new features of the recognizer are the flexible decoder and an utterance level speaker adaptation technique.

3.1. The Decoder

The decoder has been substantially expanded to fulfill the growing needs of both, international large vocabulary speech recognition evaluations and real time performance for the JANUS-II prototype system. All search passes involving acoustic scoring are now forward oriented, avoiding time delays and model inversion.

T-pass: A first tree structured pass without tree copies selects probable words for each starting point. This pass uses only approximated bigrams and trigrams. To compute cross word triphones, each state in the first diphone of each tree can model a different allophone, depending on its back pointer. To model the right context across word boundaries, leaves of the tree are allocated for each word begin phoneme when the last phoneme of a word is reached. As the tree is built on an allophone vocabulary, it is not very dense and only improves the overall speed for vocabularies over 1000 words.

F-pass: The second pass uses a flat, linear structured vocabulary allowing full bigrams and a better trigram computation. Trigrams are still approximated to minimize overhead. As the F-pass only works on a subspace of the T-pass it is about 10 times faster. Due to the more accurate language modeling, the word error on GSST of the F-pass is reduced by about 10% to 12% relative compared to the T-pass.

L-pass: By pruning the back pointer table of the second path a word-lattice is computed using full trigram information. From this lattice, we extract the best hypothesis. As the L-pass does not access the scoring module, it is typically 100 times faster than even the F-pass, but it requires the full utterance for pruning. Using trigrams instead of bigrams in all passes of the search yields a 4% error reduction for GSST. Half of this is already achieved by using approximate trigrams in the F-pass, the other half requires running the L-pass.

3.2. Speaker Adaptation

For the hypothesis H_1 of an initial recognition the viterbi path $S = (s_{i_1}, s_{i_2}, \dots s_{i_T})$ is computed. We are now looking for a transformation $\mu \rightarrow A\mu + b$ for all codebook vectors that increases the probability of observing the acoustic of the current sample given H_1. After the transformation this probability is given by

$$\prod_{t=1}^{T} p_{(A,b)}(x_t|s_{i_t}) =$$
$$\prod_{t=1}^{T} \frac{1}{\sqrt{(2\pi)^d|\Sigma_{i_t}|}} e^{(x_t-(A\mu_{i_t}+b))^{\mathrm{T}}\Sigma_{i_t}^{-1}(x_t-(A\mu_{i_t}+b))}$$

therefore we need to find

$$(\bar{A},\bar{b}) = \mathrm{argmax}_{(A,b)} \prod_{t=1}^{T} p_{(A,b)}(x_t|s_{i_t})$$

and then replace each codebook vector μ by $\bar{A}\mu + \bar{b}$. With the modified codebooks a new hypothesis is computed.

For adaptation on longer sequences, groups of codebooks are clustered together and each cluster is adapted individually.

In the June 1995 *VerbMobil* evaluation this adaptation on the utterance level yielded a relative error reduction of 2-3%. On SWB, where adaptation on whole dialogs is possible relative error reductions of 3-5% can be achieved.

3.3. Recognition Results

Table 2 gives the results of the official 1995 *VerbMobil* evaluation on German scheduling dialogs.

	Word Error
Uni Bielefeld	52.6 %
Daimler AG	38.5 %
Uni Hamburg / HTK	34.4 %
Uni Karlsruhe	31.7 %

Table 2. Results of the VerbMobil'95 evaluation (contrast test s2). Results for Karlsruhe were obtained using JANUS-SR.

The GSST, ESST and SSST results for German, English and Spanish scheduling in table 3 were obtained on internal evaluations on unseen data from our own databases.

The results presented for the Switchboard task are the results on a development set selected by NIST for SWB evaluation. The acoustic quality of the recording for SWB (telephone speech including crosstalk) is much lower than for GSST/ESST (close speaking microphone, no crosstalk) although both are human-to-human spoken dialogs and found to be considerably harder than read speech tasks. To allow JANUS-II to be used over telephone lines, improving the performance on telephone speech will be an important research issue for JANUS-SR. As we have only recently started to build Japanese and Korean recognition components there are no results for these systems yet.

	Word Error
GSST	28.0 %
ESST	27.8 %
SSST	27.7 %
Switchboard	48.0 %

Table 3. Performance of JANUS-SR on internal evaluations

4. PHOENIX AND GLR*

For a description of the parsing strategies developed in the JANUS project refer to [10, 11, 12]. In this section we will compare the performance on transcribed and spoken dialogs using two of these translation approaches: the GLR* skipping parser and the PHOENIX concept-based parser.

The evaluations comparing GLR* with PHOENIX on both the Spanish and English test sets indicate that the portion of acceptable translations produced with each of the parsers is very similar. On the Spanish transcribed test set, GLR* is slightly better (see table 4), while on English transcribed data, PHOENIX is slightly better. On speech recognized data, PHOENIX performed slightly better than GLR* in both Spanish and English. These slight performance differences should not, however, be regarded as statistically significant. The translation quality evaluations are necessarily very subjective. Although the grading is cross validated, differences in judgement between the graders have repeatedly amounted to 5% or more.

The two parsers have clear strengths and weaknesses. GLR* tries to match input utterances to an interlingua specification, so although words can be skipped with a penalty, the parser is less robust over disfluent input. Input that is parsed, though, is generated in the target language using syntactic constraints; this means that translations through GLR* are more likely to be complete grammatical sentences than those translated through PHOENIX, which parses and generates only at the speech act level.

GLR* tends to break down when parsing long utterances that are highly disfluent, or that significantly deviate from the grammar. In many such cases, GLR* succeeds in parsing only a small fragment of the entire utterance, and important input segments end up being skipped. PHOENIX is significantly better in analyzing such utterances. Because PHOENIX is a chart parser that is capable of skipping over input segments that do not correspond to any top level semantic concept, it can recover from out of domain segments in the input, and "restart" itself on the in-domain segment that follows. However, pre-breaking input to GLR* based on occurrences of human noise and parsing the shorter sub-utterances separately significantly reduced this problem. Pre-breaking benefits PHOENIX only slightly, mainly in better resolution of attachment ambiguities in time expressions. At the current time, PHOENIX uses only very simple disambiguation heuristics, whereas a parse quality mechanism helps to decide between possible parses in GLR*.

Computational requirements of GLR*, which is implemented in lisp, are far greater than those of PHOENIX, implemented in C. PHOENIX is also much faster, averaging less than one second per parse (including generation) compared to GLR*'s 25 seconds.

Because the two parsing architectures perform better on different types of utterances, they may be combined in a way that takes advantage of the strengths of each.

5. NEW FEATURES IN PHOENIX

We have recently implemented a series of extensions to the Phoenix parser, namely:

5.1. Usage of utterance boundaries

The markers <s> and </s> are used as special grammar symbols to convey information about utterance boundaries. The parser inserts the markers at the beginning and end, respectively, of each utterance to be parsed, and the grammar writers use them as terminal symbols of the grammar. Thus, it is possible to differentiate the occurrence of, say, the word *okay* when it constitutes the whole utterance, which is certainly meaningful, from a rather meaningless occurrence of the same word, as in *Why don't we meet... okay... the twenty third.*

5.2. Disambiguation heuristic

The parser now disambiguates parse trees by analyzing top-down each depth level and preferring the parse that has fewer sub-tokens, the underlying philosophy being that the more words a single, higher-level concept can cover, the better.

5.3. Rejection of out of domain phrases

Error analysis shows that a large proportion of bad translations arise from out of domain phrases that confuse the parser: since Phoenix skips unknown words, when confronted with a phrase that is out of domain the parser is still very likely to find a parse of some substring of the input phrase. For instances, the Spanish word "una" (which can both mean *one o'clock* and the indefinite article *a*), is very prone to confuse the parser if, say, followed by an unknown word. Therefore, in order to avoid the (wrong) translation of such out of domain phrases, it becomes necessary to have a mechanism that rejects such spurious parses.

Several approaches are under consideration but the one currently implemented simply looks for small islands of parsed words among non-parsed words. More concretely, let x be the number of contiguous parsed words and c the number of non-parsed words to the left and to the right of the parsed words. If there is at least one non-parsed word to the left and to the right of the parsed words, $x \leq X$ and $c \geq C$, then the parse originated by the parsed words will be filtered out. Different experimentation has shown that the best results are obtained with values of $C = 4$ and $X = 2$, in which case a 32.9% detection rate of out-of-domain parses is achieved, with no false alarms.

5.4. Open Classes

Finally, the most recent feature that has been added to the Phoenix formalism is the ability to define certain classes as open, i.e. able to parse new words. An example is for the concept of proper names: If defined as an open class, all new words that occur where a proper name could be parsed (e.g. after "Hello") will be accepted and parsed as proper nouns. They can even be automatically included into the appropriate grammar file and run-time lexicon.

6. SPEECH TRANSLATION RESULTS

The goal of the translation in JANUS is to preserve the content of an utterance. Therefore, the recognition (SR), translation (MT), and end-to-end quality needs to be assessed in terms of how well the meaning is preserved during the translation process. The speech recognition is considered to be a process with the spoken sequence as transcribed by a human for input, and as the output the word string recognized by JANUS-SR. The evaluation shows how much information is lost during the recognition process, and what is still available for the translation to work on.

Two grades were chosen for evaluation, `ok` and `bad`.
As an example, imagine the following sentence spoken into the system:

```
Spoken input:   Tuesday morning I have a meeting
```

If one or more important semantic concepts of an utterance are lost during recognition or translation, the whole recognition or translation is judged as `bad`; here are two examples of how the sentence could be corrupted, the first one for recognition the second for an English-English speech translation:

```
bad (SR):  you say morning I have a meeting
bad (MT):  Tuesday  morning works for me
```

If the meaning is preserved for all concepts in the utterance, it is judged as `ok` even if the sentence comes out somehow funny. For an `ok` recognition there is still a chance of getting a good translation. In our example, a recognition or translation scoring `ok` could look like this:

```
ok (SR):   Tuesday the morning I I have a meeting it
ok (MT):   Tuesday morning won't for me work
```

For the actual grading, utterances are broken down to sentences for better granularity. In table 4 the total percentage of acceptable recognitions or translations is given for all sentences. As the data is natural human-to-human speech, 10-30% of the sentences are not in the domain of appointment scheduling. If the parser does not reject or translate them correctly, they are counted as `bad` translations. The English and German data for this evaluation was recorded with a push-to-talk setup, yielding extremely long utterances with no crosstalk. The Spanish data was recorded without push-to-talk. The resulting utterances are shorter, but contain crosstalk and more spontaneous effects.

Recognition intelligibility	
English Recognition	65.9
German Recognition	74.0
Spanish Recognition	56.8
PHOENIX on transcriptions	
Spanish ⟶ English	81.4
English ⟶ German	88.3
German ⟶ English	75.5
Korean ⟶ Korean	80.6
GLR * on transcriptions	
Spanish ⟶ English	83.3
English ⟶ English	86.2
PHOENIX on recognizer output	
Spanish ⟶ English	73.3
English ⟶ German	60.5
German ⟶ English	66.4
Korean ⟶ Korean	50.0
GLR * on recognizer output	
Spanish ⟶ English	64.7
English ⟶ English	60.0

Table 4. Percent of sentences scoring `ok` for intelligibility of recognizer output or translation output

7. ACKNOWLEDGEMENTS

Research on the recognition engine was funded in part by grand 413-4001-01IV101S3 from the German Federal Ministry of Education, Science, Research and Technology (BMBF) as part of the VerbMobil project.
The JANUS project was supported in part by the Advanced Research Project Agency and the US Department of Defence. We also gratefully acknowledge support and cooperation with ATR Interpreting Telecommunication Laboratories and the University of Electro-Communications in Japan, ETRI in Korea, and the Siemens AG.
The views and conclusions contained in this document are those of the authors.

REFERENCES

[1] 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 *JANUS 94: Towards Spontaneous Speech Translation* ICASSP'94, Volume 1, pp345;

[2] T.Schultz, I.Rogina, A.Waibel *LVCSR-Based Language Identification* to appear in ICASSP'96;

[3] I.Rogina, A.Waibel *Learning state-dependent Stream Weights for multi-codebook HMM Speech Recognition Systems* ICASSP'94, Volume 1, pp217;

[4] M.Woszczyna, M.Finke *Minimizing search errors due to delayed bigrams in real-time speech recognition systems* to appear in ICASSP'96;

[5] T.Schultz, I.Rogina *Acoustic and Language Modeling of Human and Nonhuman Noises* ICASSP'95, Volume 1, pp293;

[6] T.Kemp, A.Jusek *Modelling Unknown Words in Spontaneous Speech* to appear in ICASSP'96;

[7] T.Sloboda *Dictionary Learning: Performance Through Consistency* ICASSP'95, Volume 1, pp453;

[8] Petra Geutner *Using Morphology Towards Better Large-Vocabulary Speech Recognition Systems* ICASSP'95 Volume 1, pp445;

[9] A.E.McNair, A.Waibel *Improving Recognizer Acceptance through Robust, Natural Speech Repair* ICSLP94;

[10] L.J.Mayfield, M.Gavaldà, Y-H.Seo, B.Suhm. W.Ward, A.Waibel *Parsing real input in JANUS a concept based approach to spoken language translation* TMI'95, pp196;

[11] L.J.Mayfield, M.Gavaldà, W.Ward, A.Waibel *Concept-based Speech Translation* ICASSP'95, Volume 1, pp197;

[12] F.D.Buø, T.Polzin, A.Waibel *Learning Complex Output Representations in Connectionist Parsing of Spoken Language* ICASSP'94, Volume 1, pp365;

[13] V.M.Jimenez, A.Castellanos, E. Vidal *Some Results with a Trainable Speech Translation and Understanding System* ICASSP'95, Volume 1, pp113;

[14] Young-Jik Lee, Young-Sum Kim, Jung-Chul Lee, Joon-Hyung Ryoo, Jae-Woo Yang *Korean-Japanese Speech Translation System for Hotel Reservation - Korean Front Desk Side* EUROSPEECH'95, Volume 2, pp1197

[15] Myoung-Wan Koo, Il-Hyun Sohn, Woo-Sung Kim, Du-Seong Chang *A Speech Translation System for Hotel Reservation and a Continuous Speech Recognition System for Speech Translation* EUROSPEECH'95, Volume2, pp1227

[16] M-S.Ägnas, H.Alshawi, I.Bretan, D.Carter, K.Ceder, M.Collins, R.Crouch, V.Digalakis, B.Ekholm, B.Gambäck, J.Kaja, J.Karlgren, B.Lyberg, P.Price, S.Pulman, M.Rayner, C.Samuelsson, T.Svensson *Spoken Language Translator: First Year Report* ARPA-SLT'94