JANUS: TOWARDS MULTILINGUAL SPOKEN LANGUAGE TRANSLATION

B. Suhm[1], *P. Geutner*[2], *T. Kemp*[2], *A. Lavie*[1], *L. Mayfield*[1], *A. E. McNair*[1], *I. Rogina*[2], *T. Schultz*[2], *T. Sloboda*[2], *W. Ward*[1], *M. Woszczyna*[1], *A. Waibel* [1,2]

Interactive Systems Laboratories
[1] Carnegie Mellon University (USA)
[2] Karlsruhe University (Germany)

ABSTRACT

In our effort to build spoken language translation systems we have extended our JANUS system to process spontaneous human–human dialogs in a new domain, two people trying to schedule a meeting. Trained on an initial database JANUS-2 is able to translate English and German spoken input in either English, German, Spanish, Japanese or Korean output. To tackle the difficulty of spontaneous human–human dialogs we improved the JANUS-2 recognizer along its three knowledge sources acoustic models, dictionary and language models. We developed a robust translation system which performs semantic rather than syntactic analysis and thus is particulary suited to processing spontaneous speech. We describe repair methods to recover from recognition errors.

1. Introduction

JANUS [1, 2] has been among the first systems attempting to provide spoken language translation. While the previous JANUS-1 system processed syntactically wellformed read speech over a 500 word vocabulary, JANUS-2 operates on spontaneous human–human dialogs in a scheduling domain with vocabularies exceeding 2000 words. Currently, English and German spoken input can be translated in either English, German, Spanish, Japanese or Korean output. Work is in progress to add Spanish and Korean as input languages.

This paper reports on the current status of the system and ongoing efforts to extend and improve the recognition component. Then, we describe our new approach to robust translation of spoken language. We briefly describe and compare the alternative approach to parsing and translation we pursue, based on a generalized robust LR parser and an ILT. Finally we report on efforts to detect erroneous system output and provide interactive methods to recover from such errors.

2. Current Status of JANUS

2.1. Data Collection

Data collection to establish a large database of spontaneous human–human negotiation dialogs in English and German has started about 18 months ago. In the meantime, several sites in Europe, the US and Asia have adopted the Scheduling task under several research projects and funding sources. Since the same calendars and data collection protocols are used the data elicited shares the same domain and procedural constraints.

English Scheduling		
	dialogs	words
recorded	1984	505 K
transcribed	1826	460 K
German Scheduling		
	dialogs	words
recorded	734	158 K
transcribed	534	115 K
Spanish Scheduling		
	dialogs	words
recorded	340	79 K
transcribed	256	70 K
ATIS3		
transcribed	n./a.	250 K

Table 1: Comparison of Databases (as of December 1994)

Table 1 summarizes the current status of data collection. Since Scheduling utterances typically consist of more than one sentence, there is already more data available for English Scheduling than ATIS [1]. More data collection will establish databases in size at least comparable to ATIS for all languages.

In Spanish, we have explored two different data collection scenarios: To allow only one person to speak at a time the *push-to-talk* scenario requires the speaker to push a button while talking to the system. The *cross-talk* scenario allows speakers to speak simultaneously without push button. The speech of each dialog partner is recorded on separate channels.

2.2. System Overview

The main system modules are speech recognition, parsing, discourse processing, and generation. Each module is lan-

[1] The about 18000 utterances in English Scheduling correspond to some 30000 sentences.

guage-independent in the sense that it consists of a general processor that applies independently specified knowledge about different languages.

The recognition module decodes the speech in the source language into a list of sentence candidates, represented either as a word lattice or Nbest list. At the core of the machine translation components is a language independent representation of the meaning, which is extracted from the recognizer output by the parsing module. As last step, the final language independent representation is sent to the generator to be translated in any of the target languages. Figure 1 shows the system architecture.

After parsing, a discourse processor can be used to put the current utterance in the context of previous utterances, opening possibilities to integrate the speech and natural language processing compenents of the system to resolve parsing ambiguities and dynamically adapt the vocabulary and language model of the recognizer based on the current discourse state.

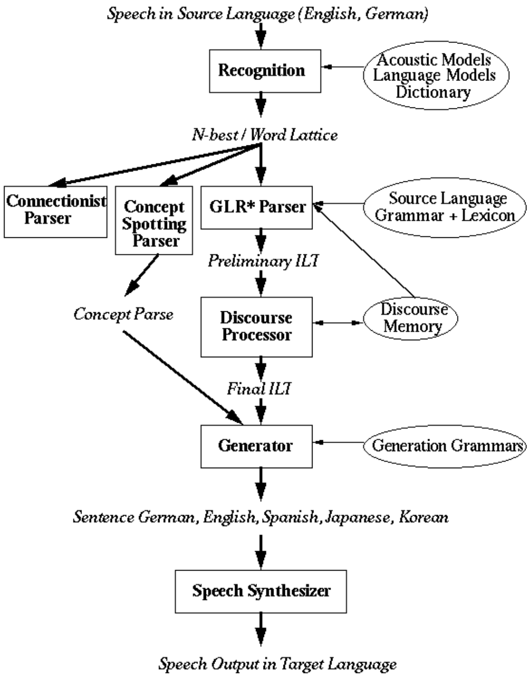

Figure 1: System Diagram

We explore several approaches for the main processes. For example, we are experimenting with TDNN, MS-TDNN [3], MLP, LVQ [4], and HMM's [5, 12] for acoustic modeling; n-grams, word clustering, and automatic phrase detection for language modeling [6]; statistically trained skipping parsing [7, 8], neural net parsing [9] and concept spotting parsing [10] for extracting the meaning; and statistical models as well as plan inferencing for identification of the discourse state [11]. This multi-strategy approach should lead to improved performance with appropriate weighting of the output from each strategy.

2.3. Recognition Performance Analysis

The baseline JANUS-2 recognizer can be described as follows:

- *Preprocessing*: LDA on melscale fourier spectrum and additional acoustic features (power, silence)
- *Acoustic modeling*: LVQ-2 or phonetically tied SCHMM, no cross–word triphones, explicit noise models
- *Decoder*: Viterbi search as first pass, followed by a word-dependent Nbest search, standard word bigram language model, word lattice output

Current recognition results on the English, German and Spanish Spontaneous Scheduling Task (ESST, GSST, SSST) can be seen in table 2.

	ESST	GSST	SSST
Word Accuracy	66%	72%	61%

Table 2: JANUS-2 baseline recognition performance

The low absolute recognition accuracies are due to the challenging nature of human–human spontaneous speech. In the official evaluation of the German VERBMOBIL project on the GSST task, the JANUS-2 decoder outperformed all other participating systems. In addition, recent evaluations on the Switchboard task confirm that human–human dialogs are much more difficult to recognize than human–machine spontaneous speech (like ATIS). Participating systems achieved word accuracies between 30% and 50%.

Analysis shows that human-human dialogs (like Scheduling or Switchboard) are more difficult to recognize than human-machine dialogs (e.g. ATIS). Perplexities lie between 35 and 90 for ESST, SSST and GSST, and somwhat over 100 for Switchboard. Additionally, human-human dialogs are significantly more disfluent [8]. Large variations in speaking rates and strong coarticulation between words contribute significantly to the difficulty of recognizing human-human spontaneous speech.

3. Improving the Recognition Component

We describe efforts to improve the recognition component along its major knowledge sources acoustic models [12], dictionary [13] and language models [14].

3.1. Data–Driven Codebook Adaptation

We developed methods aimed at automatic optimization of the number of parameters for the semi-continuous phonetically tied HMM used in JANUS-2. Usually, a fixed number of codebook vectors is assigned to each of the phonemes. However, as the available training data differs between phonemes and the size of the feature space phonemes cover varies greatly, constant codebook size leads to suboptimal allocation of resources.

We have therefore suggested [12] to adapt the codebook size of each phoneme according to the amount and the distribution of the training data, similar to [15]. During training, the size of the codebook is incrementally increased. Some quality criterion determines when to stop the process of increasing the codebook. We compared a *variance* criterion based on the average distance between data points and their nearest codebook vector with a *prediction* criterion which tries to capture how well the modeling of the recognizer can predict unseen data.

Model	Codebook Size	Word Accuracy
baseline	4600	66.9%
variance	4201	69.9%
prediction	1677	67.8%

Table 3: Results for Codebook Adaptation (GSST)

Table 3 compares recognition accuracies and codebook sizes of the baseline models, with models automatically adapted using the variance and prediction criterion. As can be seen, codebook adaptation leads to significant error reduction if the same number of parameters is used.The number of parameters can be reduced by 40% with still better performance than the baseline system.

3.2. Dictionary Learning

Due to the enormous variability in spontaneous human–human dialogs creating adequate dictionaries with alternative pronunciations is crucial [16]. However, hand tuning and modifying dictionaries is time consuming and labor intensive. Pronunciations of a word should be chosen according to their frequency. Modifications of the dictionary should not lead to higher phonetic confusability after retraining. Therefore we have proposed [13] a data-driven approach to improve existing dictionaries and automatically add new words and pronunciation variants whenever needed.

The learning algorithm requires transcripts for the whole training set and a phoneme confusability matrix of the speech recognizer used. First, phonetic transcriptions for all appearances of each word are generated by help of a phoneme recognizer. Then, variants which are infrequent or which would lead to erroneous training of confusable phonemes are eleminated. Finally, the acoustic models are retrained allowing for the newly aquired pronunciations variants.

As can be seen in table 4, our algorithm for adapting and adding phonetic transcriptions to a dictionary improves the recognition accuracy of the decoder significantly and leads to performance that is comparable to the context dependent results (cf. table 2). The baseline decoder for these experiments uses 69 context independent phoneme models. Evaluation using context dependent models is in progress.

Dictionary	Word Accuracy
baseline	61.7%
adapted	65.6%

Table 4: Results Dictionary Learning (GSST)

3.3. Morpheme Based Language Models

Based on our scheduling databases we noticed that in morphologically rich languages such as German and Spanish, dictionaries grow much faster with increasing database size, compared to English (cf. figure 2). This is due to the large number of inflections and compound words. One way to limit this growth with increasing dictionary sizes is to use other base units than words.

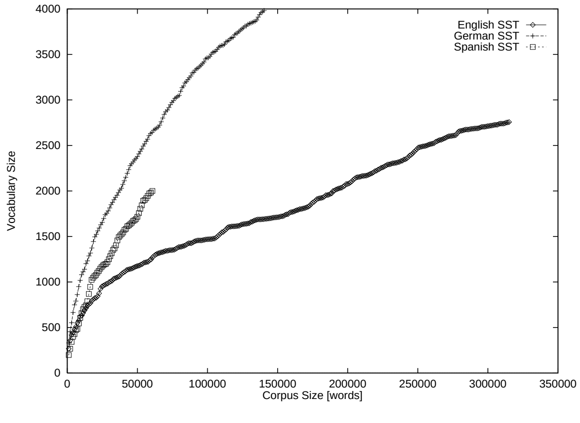

Figure 2: Vocabulary Growth

We compared three different decomposition methods:

- strictly *morpheme* based decomposition, e.g. weggehen (to go away) $\longrightarrow$ weg–geh–en, Spracherkennung (speech recognition) $\longrightarrow$ Sprach–er–kenn–ung
- decomposition in *root forms*, e.g. weggehen (to go away) $\longrightarrow$ weggeh@, Spracherkennung (speech recognition) $\longrightarrow$ Spracherkenn@

- combination of strictly morpheme based decomposition and root forms

Table 5 shows dictionary size, bigram perplexity and recognition accuracy using the respective decomposition method, based on 250 GSST dialogs. As can be seen, all decomposition methods significantly reduce vocabulary size and perplexity. The impact on recognition accuracy is still small. This may be due to the fact that the acoustic modeling suffers from smaller units and thus deteriorate the gain in the language model. In a real interface, however, this reduction in vocabulary growth leads to a reduction of new words. Further research will focus on finding more efficient and acoustically less confusable decompositions automatically, and also test the impact on translation.

	Dictionary	Perplexity	Accuracy
Baseline	3821	88	64.7%
Morphemes	2391	46	65.4%
Root Forms	3205	79	63.5%
Combined	2998	59	65.1%

Table 5: Comparison of Decomposition Methods (GSST)

4. Concept Based Speech Translation

We have developed a robust translation system based on the information structures inherent in the appointment scheduling task being performed, described in detail elsewhere [10]. The basic premise is that the structure of the information conveyed is largely independent of the language used to encode it. Our system tries to model the information structures in a task and the way these structures are realized in words in various languages. This system is an extension of the Phoenix Spoken Language System [18]. It uses the Phoenix parser to parse input into slots in semantic frames, and then uses these frames to generate output in the target language.

4.1. The Parser

Unlike individual words, semantic units used in a task domain are not language specific. Based on transcripts of scheduling dialogs, we have developed a set of fundamental semantic units in our parse which represent the different concepts a speaker would use. For instance, a typical *temporal* token could have *date* as subtoken, which could in turn consist of *month* and *day* subtokens. The *temporal* could be part of a statement of unavailability.

In contrast to previous speech translation systems, we presently don't perform syntactic analysis. Speaker utterances, as decoded by the recognizer, are parsed into semantic chunks which are concatenated without grammatical rules.

```
Original utterance:
  THAT SATURDAY I'M NOT SURE ABOUT BUT YOU SAID
    YOU MAY BE BACK IF YOU THINK YOU'LL BE BACK
    THE THIS SUNDAY THE TWENTY EIGHTH I COULD SEE
    YOU AFTER ELEVEN AM ON THAT IF YOU'RE BACK

Translated:
```

Saturday that's not so good for me Sunday the twenty eighth works for me after eleven a.m. (ENGLISH)

El sábado no me va demasiado bien pero el domingo veintiocho me va bien después de las once de la mañana. (SPANISH)

Samstag könnte ich nur zur Not aber Sonntag der Achtundzwanzigste geht bei mir ganz gut nach elf Uhr morgens. (GERMAN)

Figure 3: Translation Example

This approach is particularly well suited to parsing spontaneous speech, which is often ungrammatical and subject to recognition errors. This approach is more robust than requiring well-formed input and reliance on syntactic cues provided by short function words such as articles and prepositions.

4.2. The Generator

The generation component of the system is a simple left-to-right processing of the parsed text. The translation grammar consists of a set of target-language phrasings for each token, including lookup tables for variables like numbers and days of the week. When a lowest-level token is reached in tracing through the parse, a target-language representation is created by replacing tokens with templates for the parent token, according to the translation grammar. The result is a meaningful, although terse translation, which emphasizes communicating the main point of an utterance. An examples is illustrated in figure 3.

4.3. Results

We have implemented this system for bi-directional translation between English, German and Spanish in our scheduling task. Table 4 shows the performance of parser and subsequent generator on transcribed data. Evaluation of the system based on speech decoded by the JANUS-2 recognizer is still underway.

	Parsed from		Translated into
	token	utterance	utterance
English	95.6%	90.0%	90.2%
German	92.4	89.6	87.3
Spanish	88.8	58.3	82.2

Figure 4: End–to–End evaluation on transcribed data

One disadvantage of this approach is the telegraphic and repetitive nature of the translations. This could be overcome by providing multiple translation options for individual tokens in the target-language module, different levels of politeness, etc. However at present we feel that it is sufficient for intelligible communcation.

5. GLR* Parser

In addition to the concept based Phoenix parser we pursue GLR* as robust extension of the Generalized LR Parser. It attempts to find maximal subsets of the input that are parsable, skipping over unrecognizable parts of the input sentence [7]. By means of a semantic grammar GLR* parses input sentences into an interlingua text (ILT) as language independent representation of the meaning of the input sentence, described in more detail elsewhere (e.g. [8]).

Compared to Phoenix parses the ILT generated by GLR* offers greater level of detail and more specificity, e.g. different speaker attitudes and levels of politeness. Thus, translation based on ILTs is more natural, overcoming the telegraphic and terse nature of concept based translation.

A drawback of GLR* was that it expected input segmented into sentences for efficiency reasons. However, typical Scheduling utterances consist of 2-3 sentences. To integrate the parser with the speech decoder, we developed methods which extend the parsing capabilities from single sentences to multi-sentence utterances. We extended the grammar with a high-level rule that allows the input utterance to be analyzed as a concatenation of several sentences and developed two methods to constrain the number of sentence breaks that are considered by the parser. The first is a heuristic which prunes out all parses that are not minimal in the number of sentences. The second is a statistical method to disregard potential sentence breaking points that are statistically unlikely.

For the English analysis grammar, time efficiency thus improved by about 30%. As an additional benefit, the parse quality improved because strange sentence breaks are rejected in favor of a more reasonable location.

6. Handling Unreliability

Although research has boosted performance of speech recognition and spoken language translation technology, recognition and translation errors will persist. To build a system for use in real applications we need repair methods to recover from errors in a graceful and unobstrusive way. We have developed a speech interface for repairing *recognition* errors by simply respeaking or spelling a misrecognized section of an utterance. While much speech "repair" work has focused on repairs within a single spoken utterance [19], we are concerned with the interactive repair of errorful recognizer hypotheses [20].

6.1. Identifying Errors

To be able to repair an error its location has to be determined first. We pursue two strategies to identify misrecognitions as subpieces of the initial recognizer hypothesis.

The *automatic subpiece location* technique requires the user to respeak only the errorful subsection of the (primary) utterance. This (secondary) utterance is decoded using a vocabulary and language model limited to substrings of the initial erroneous hypothesis. Thus, the decoding identifies the respoken section in the hypothesis. Preliminary testing showed that the method works poorly if the subpiece to be located is only one or two words long. However, this drawback is not severe since humans tend to respeak a few words around the error.

A second technique uses *confidence measures* to determine for each word in the recognizer hypothesis whether it was misrecognized. First, we applied a technique similar to Ward [21], which turns the score for each word obtained during decoding into a confidence measure by normalizing the score and using a Bayesian updating technique based on histograms of the normalized score for correct and misrecognized words. Since we found this not to work well on our English scheduling task, we are currently developing different methods to compute confidence measures based on decoder, language model and parser scores.

6.2. Robust Speech Repair

After locating and highlighting erroneous sections in the recognizer hypothesis misrecognitions are corrected.

The *spoken hypothesis correction* method uses Nbest lists for both the initial utterance and the respoken section. The Nbest for the highlighted section of the initial utterance is rescored using scores from decoding the secondary utterance. Depending on the quality of the Nbest lists, most misrecognitions can be corrected.

The *spelling hypothesis correction* method requires the user to spell the highlighted erroneous section. A spelling recognizer decodes the spelled sequence of letters. By means of a language model we restrict the sequence of letters to alternatives found among the Nbest from the located section.

To date, we have evaluated our methods over sentences from the Resource Management task. Table 6 shows the improvements in sentence accuracy, based on recordings from one speaker of the February and October 1989 test data. We selected a subset of erroneous utterances; therefore the accuracy of the baseline system is significantly lower than the 94% performance our system achieves on the whole test set. The results indicate that repeating or spelling a misrecognized subsection of an utterance can be an effective way to repair recognition utterances.

No Repair (baseline)	63.1%
Respeak	83.8%
Spell	88.5%
Respeak + Spell	89.9%

Table 6: Improvement of Sentence Accuracy by Repair

7. Conclusions

We have made significant advances towards building a multi lingual translation system for spontaneous human–human dialogs. Beyond speech recognition of spontaneous speech JANUS provides a framework to investigate important areas like robust parsing, machine translation of spoken language and developing methods to recover from recognition and parsing errors. To achieve acceptance in real applications, we have to embed the spoken language technology in a sensible and useful user interface that is carefully designed around human factors and common needs. To be flexible and robust, such interfaces should not only recognize speech but also recognize other communication modalities, provide freedom from headset and push-buttons, allow for graceful recovery from errors and miscommunications, know what they don't know, and model what the user does or doesn't know [23].

8. Acknowledgements

This research was funded in part by grants from the Advanced Research Project Agency and the German Ministry of Science and Technology (BMFT) under project Verbmobil. We also gratefully acknoledge support by ATR Interpreting Telecommunications Research Labs of Japan. The views and conclusions contained in this document are those of the authors.

References

1. L. Osterholtz, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, M. Woszczyna: *Testing Generality in JANUS: A Multi-Lingual Speech to Speech Translation System,* Proc. ICASSP 92, vol. 1, pp. 209-212
2. M. Woszczyna, , N. Coccaro, A. Eisele, A. Lavie, A.-E. McNair, T. Polzin, I. Rogina, C. Pennstein-Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki–Waibel, A. Waibel, W. Ward: *Recent Advances in JANUS: A Speech Translation System,* DARPA Speech and Natural Language Workshop 1993, session 6 - MT
3. H. Hild and A. Waibel: *Connected Letter Recognition with a Multi–State Time Delay Neural Network,* Neural Information Processing Systems (NIPS-5), Morgan Kaufman
4. O. Schmidbauer and J. Tebelskis: *An LVQ based Reference Model for Speaker–Adaptive Speech Recognition,* Proc. ICASSP 92, Vol. 1, pp. 441-445
5. I. Rogina and A. Waibel: *Learning State–Dependant Stream Weights for Multi–Codebook HMM Speech Recognition Systems,* Proc. ICASSP 94
6. B. Suhm and A. Waibel: *Towards Better Language Models for Spontaneous Speech,* ICSLP 94, Vol. 2, pp. 831-834
7. A. Lavie and M. Tomita: *GLR* - An Efficient Noise-skipping Parsing Algorithm for Context-free Grammars,* Proceedings of Third International Workshop on Parsing Technologies, 1993, pp. 123–134
8. B. Suhm, L. Levin, N. Coccaro, J. Carbonell, K. Horiguchi, R. Isotani, A. Lavie, L. Mayfield, C. Pennstein-Rosé, C. Van Ess-Dykema and A. Waibel: *Speech–Language Integration in a Multi–Lingual Speech Translation System,* Workshop on Integration of Natual Language and Speech Processing, AAAI-94, Seattle
9. F.-D. Buø, T.-S. Polzin and A. Waibel: *Learning Complex Output Representations in Connectionist Parsing of Spontaneous Speech,* Proc. ICASSP 94, Vol. 1, pp. 365-368
10. L. Mayfield, M. Gavalda, W. Ward and A. Waibel: *Concept Based Speech Translation,* to appear in Proc. ICASSP 95
11. Carolyn Penstein Rosé, Alex Waibel: *Recovering From Parser Failures: A Hybrid Statistical/Symbolic Approach,* to appear in "The Balancing Act: Combining Symbolic and Statistical Approaches to Language" workshop at the 32nd Annual Meeting of the ACL, 1994
12. T. Kemp: *Data-Driven Codebook Adaptation in phonetically tied SCHMMS,* to appear in Proc. ICASSP 95
13. T. Sloboda: *Dictionary Learning: Performance through Consistency,* to appear in Proc. ICASSP 95
14. P. Geutner: *Using Morphology towards better Large Vocabulary Speech Recognition Systems,* to appear in Proc. ICASSP 95
15. U. Bodenhausen: *Automatic Structuring of Neural Networks for Spatio-Temporal Real-World Applications,* Ph.D thesis, University of Karlsruhe, June 1994
16. J.-L. Gauvin, L.-F. Lamel, G. Adda and M. Adda-Decker: *The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task,* Proc. ICASSP 94, vol. 1, pp. 557-560
17. R. Kneser and H. Ney: *Improved Clustering Techniques for Class-Based Statistical Language Models,* EUROSPEECH 93, Berlin, Vol. 2, pp. 973-976
18. W. Ward: *Understanding Spontaneous Speech: The Phoenix System,* IEEE International Conference on Acoustics, Speech and Signal Processing, 1991, Vol. 1, pp. 365–367
19. C. Nakatani and J. Hirschberg: *A Speech–First Model for Repair Identification in Spoken Language Systems,* in Proceedings of the ARPA Workshop on Human Language Technology, March 1993
20. A.-E. McNair and A. Waibel: *Improving Recognizer Acceptance through Robust, Natural Speech Repair* ICSLP 94, Vol. 3., pp. 1299-1303
21. S.-R. Young and W. Ward: *Learning New Words from Spontaneous Speech,* Proc. ICASSP 93, Vol. 2, pp. 590–591
22. N. Yankelovich, G.-A. Levow and M. Marx: *Designing SpeechActs: Issues in Speech User Interfaces,* to be presented at CHI 95, Denver
23. M.-T. Vo, R. Houghton, J. Yang, U. Bub, U. Meier and A. Waibel: *Multimodal Learning Interfaces,* to be presented at Spoken Language Technology Workshop 1995, Austin