A New Approach to the Translating Telephone

Robert Frederking **Christopher Hogan** **Alexander Rudnicky**

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15217 USA

Abstract

The Translating Telephone has been a major goal of speech translation for many years. Previous approaches have attempted to work from limited-domain, fully-automatic translation towards broad-coverage, fully-automatic translation. We are approaching the problem from a different direction: starting with a broad-coverage but not fully-automatic system, and working towards full automation. We believe that working in this direction will provide us with better feedback, by observing users and collecting language data under realistic conditions, and thus may allow more rapid progress towards the same ultimate goal. Our initial approach relies on the wide-spread availability of Internet connections and web browsers to provide a user interface. We describe our initial work, which is an extension of the Diplomat wearable speech translator.

1 Introduction

The Translating Telephone has been the goal of several major speech-to-speech translation projects, such as those at ATR (Moritomi et al. 1993) in Japan, the JANUS group at Carnegie Mellon University (Waibel et al. 1991, Woszczyna et al. 1994), and others in the C-STAR consortium. While there have been numerous significant accomplishments in these projects over a number of years, the ultimate goal of a useful Translating Telephone still seems remote. We in the Diplomat project (Frederking et al. 1997, Frederking et al. 1999) have therefore decided to attack the problem from a new direction, to determine whether more rapid progress can be made. Instead of gradually extending a limited-domain, fully-automatic system, we will begin with a broad-coverage, human-aided speech-to-speech translation system, and attempt to move towards full automation.

The rest of Section 1 further describes our strategic view of machine translation research, and our previous work in this particular strategic direction. Section 2 presents the current state of the Diplomat system, which provides the foundation for this new thrust. Section 3 then describes the surprisingly minor extensions required to transform Diplomat into a Translating Telephone. We conclude in Section 4.

1.1 Trade-offs in current MT

In order to understand the choice between these two strategic approaches, one must understand the trade-offs involved in the current state of the art in machine translation (MT). As we see it, the ultimate goal of MT is a system that is:

1. Fully-automatic (no human intervention required)
2. General purpose (not limited-domain)
3. High-quality (the output is essentially as good as human-produced)

It is currently possible to achieve any two of these three objectives in a single system, using different state-of-the-art MT techniques; but achieving all three at once is still well beyond current capabilities.

Seen from the most strategic level, one can therefore envision working from an initial system that satisfies any two of these goals, and striving to meet the third. The question that needs to be answered is which characteristics to begin with, and which one to work on extending. Since speech translation systems must support conversations in order to be tested effectively, a reasonable level of quality is necessary (otherwise the human participants cannot keep the conversation going; it immediately breaks down). This leaves us with a two-way choice for the initial system: limited-domain or human-intervention.

The main benefit we expect to receive from choosing human-aided, unrestricted MT as our starting point is the ability to have users engage in useful, successful dialogs on a variety of topics right from the start. We expect that this will provide us with large amounts of data for training and analysis, collected under realistic conditions, and will clarify which areas have the highest priority for further work: those areas that require the greatest amount of human intervention will be the ones with highest priority for improvement. Even better, if we can reach the point of providing ordinary users with a service they find useful, we should be able to collect arbitrarily large amounts of data.

1.2 Earlier related work

We have employed this strategy of starting with a human-aided system and working towards full automation in two other projects: the Pangloss project (Frederking et al. 1994), and the Communicator project (Constantinides et al. 1998).

The Pangloss MT system was specifically designed as a human-in-the-loop system. It made use of a Translator's Work Station (TWS) as its development platform and user interface. A major focus of work was producing a graphical user interface (GUI) that would allow the human translator to interact with the results of the MT sysicni as efficiently as possible (Frederking et al. 1993). The most important piece of this GUI was a popup menu of alternative translations; any time the user came across an oddly translated segment, they could use the popup editor to see and select alternative translations for that segment[1]. The current Diplomat project (described in Section 2 below) and its extension to a Translating Telephone (Section 3 below) both make use of the same style of popup menu interface, extended and reimplemented in technology appropriate for each application (C++ or Java, respectively).

The Carnegie Mellon Communicator project (Constantinides et al. 1998) is developing human-to-machine speech dialog interfaces, in the domain of making business travel arrangements over the telephone. One of many interesting aspects of this system is the Communicator/Supervisor model. When the Communicator detects that a conversation is breaking down, for example because the same question is being asked repeatedly, it will tell the user "I will get my supervisor." The supervisor is a human with a display of the conversation up to this point, audio listening capability, and domain knowledge. The supervisor gets the conversation back on track, and then conversation with the Communicator resumes. This is analogous to currently deployed automatic directory assistance services in the U.S., which use a human fallback when they have difficulty recognizing and obtaining a desired listing. The automatic directory system is of value to the telephone company, despite not being fully automatic, because it greatly reduces the number of human beings required to provide a given level of service.

2 Pre-existing Diplomat speech translation work

The technical basis for this Translating Telephone effort is the Diplomat (Frederking et al. 1997, Frederking et al. 1999) rapid-development speech translation project. This project has as its goal the production of a speech translator that can be rapidly moved to new languages and new domains, and deployed on a wearable platform for use by. e.g., peace keeping forces. Rapid development is necessary due to the speed and unpredictability with which new crises can develop, relative to the length of time typical MT and speech recognition systems require to achieve a useful level of performance.

The MT techniques employed to achieve rapid deployment and broad coverage require human-intervention, but this is readily available, since there is always a trained user present, and we assume that both speakers are trying to cooperate in communicating. This system therefore provides the type of general-purpose, human-assisted MT that we need for the Translating Telephone. We will briefly describe the most pertinent aspects of the current system here; more details are available elsewhere (Frederking et al. 1997, Frederking et al. 1999).

2.1 Multi-Engine Machine Translation

Diplomat uses the Multi-Engine Machine Translation (MEMT) architecture (Frederking and Nirenburg 1994). As shown in Figure 1 below, MEMT feeds an input text to several MT engines in parallel, with each engine employing a different MT technology. Morphological analysis, part-of-speech tagging, and possibly other text enhancements can be shared by the engines. Each engine attempts to translate the entire input text, segmenting each sentence in whatever manner is most appropriate for its technology, and putting the resulting translated output segments into a shared chart data structure (Kay 1967, Winograd 1983) after giving each segment a score indicating the engine's internal assessment of the quality of the output segment. These output (*target language*) segments are indexed in the chart based on the positions of the corresponding input (*source language*) segments. Thus the chart contains multiple, possibly overlapping, alternative translations. Since the scores produced by the engines are estimates of variable accuracy, we use statistical language modeling techniques adapted from speech recognition research to select the best overall set of outputs (Brown and Frederking 1995). These selection techniques attempt to produce the best overall result, taking the probability of transitions between segments into account as well as modifying the quality scores of individual segments. A recent evaluation (Hogan and Frederking 1998) has demonstrated that the MEMT architecture can indeed produce better translations than any single component MT engine.

[1] This design was based on the observation that a reasonably good translation is almost always available as one of the alternative translations.

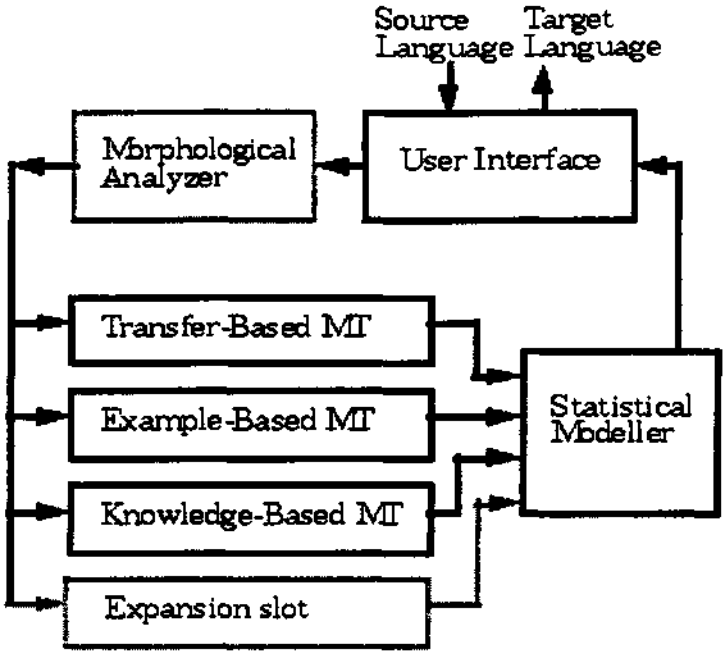

Figure 1: Multi-Engine MT Architecture

For the languages developed so far, the primary engines that we have produced have been EBMT and lexical-transfer MT:

- **EBMT** (Brown 1996) uses a sentence-aligned corpus to produce translations. When such a corpus is available, fairly good quality MT for a new domain is available essentially immediately. EBMT is basically a more sophisticated version of Translation Memory, in that sub-sentential chunks of words are matched, allowing much greater coverage. Sentences that match in full are translated exactly, but sub-sentential chunks are matched with a variety of heuristics, which are reflected in the quality scores assigned to the corresponding outputs.
- **Lexical-transfer MT** employs a very simple, very old technology: bilingual dictionaries and phrasal glossaries are used to translate pieces of source text. While this is a low-quality technique, the simplicity of the technique allows us to quickly and semi-automatically develop large databases using native speakers with no special training, allowing an initial rapid-deployment of an MT system even when parallel corpora are unavailable. Quality scores can be statically assigned on a per-glossary basis, using our overall confidence in a particular glossary.

The EBMT and lexical-transfer MT engines used in Diplomat are described in more detail elsewhere (Brown 1996, Nirenburg et al 1995).

While our original reason for employing the MEMT architecture was to reduce the time and cost required to develop new languages, there also appears to be an interesting match between the properties of spoken input and the properties of a rapid-development MEMT system. Compared to text translation, the input to speech translation is of much lower quality, due both to the word-error-rate of state-of-the-art real-time speech recognition and to the disfluencies present in spontaneous speech. That is. spontaneous speakers often do not utter the complete, grammatical sentences that linguistic analysis typically expects. Thus although Knowledge-Based MT (KBMT) systems produce better output translation quality than the EBMT and lexical-transfer MT engines employed in rapid-deployment MEMT, the degraded quality of spoken input means that the quality difference between KBMT and our MEMT is less important. Given a string of words containing several random word substitutions in addition to structural anomalies, it appears to us that a general-purpose MEMT system can do about as well as a (much more costly) KBMT system. This claim of course would require serious testing before it could be asserted as fact.

For the purposes of this paper, there are two other important aspects of the MEMT architecture:

- The initially deployed versions are quite error-prone, although generally a correct translation is among the available choices. This necessitates a strong user interaction capability and significant field testing to determine whether the initial versions of the system are in fact usable for the intended application.

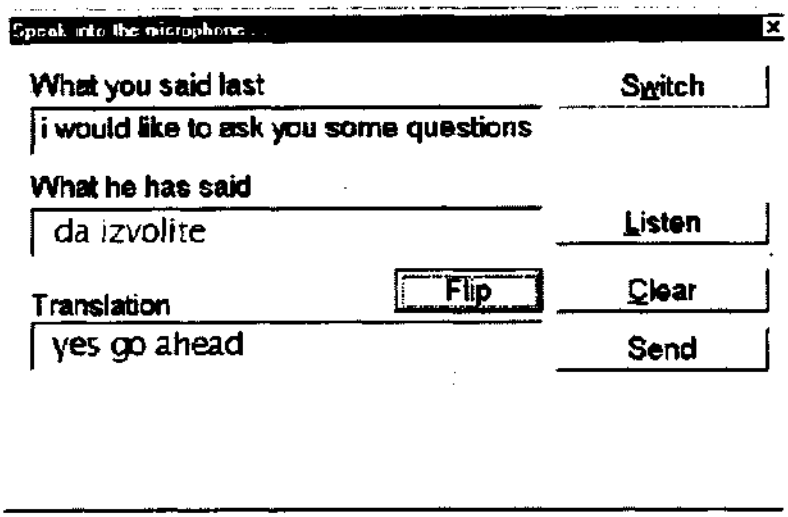

Figure 2: Diplomat User Interface

- The unchosen alternative translations are still available in the chart structure after scoring by the target language model. This allows later user interaction to improve the automatic selection, if the user wishes.

2.2 User-interface design

As indicated in the introduction, our approach in the Diplomat project to coping with error-prone speech translation is to allow user correction wherever feasible. While we would like as much user interaction as possible, it is also important not to overwhelm the user with either information or decisions. This requires a careful balance, which we are trying to achieve through early user testing. We have carried out initial testing using local naive subjects (e.g., drama majors and construction workers), and intend to carry out early tests with actual end users as soon as specific ones can be identified.

The primary potential use for Diplomat identified so far is to allow English-speaking soldiers on peace-keeping missions to interview local residents. While one could conceivably train the interviewer to use a restricted vocabulary, the interviewee's responses are much more difficult to control or predict. Our current system has been designed to run on a either a laptop or a wearable computer, with each speaker taking turns using a graphical user interface (GUI) on a single display screen (see Figure 2).

Feedback from initial demonstrations made it clear that, while we could expect the interviewer to have roughly eight hours of training, we needed to design the system to work with a totally naive interviewee, who had never used a computer before. We responded to this requirement by developing an asymmetric interface, where any necessary complex operations were moved to the interviewer's side. The interviewee's GUI is now extremely simple, and a touchscreen has been added, so that the interviewee is not required to type or use the pointer. In addition, the interviewer's GUI controls the state of the interviewee's GUI. The speech recognition system continuously listens, so the participants do not need to physically indicate their intention of speaking.

A typical exchange consists of recognizing the interviewer's spoken utterance, translating it to the target language, backtranslating it to English[2], then displaying and synthesizing the (possibly corrected) translation. The interviewee's response is recognized, translated to English, and backtranslated. The (possibly corrected) backtranslation is then shown to the interviewee for confirmation. The interviewer receives a graphic indication of whether the backtranslation was accepted or not. (The actual communication process is quite flexible, but this is a normal scenario.)

In order to achieve such communication, the users currently can interact with Diplomat in the following ways:

[2] We realize that backtranslation is also an error-prone process, but it at least provides some evidence as to whether the translation was correct to someone who does not speak the target language at all.

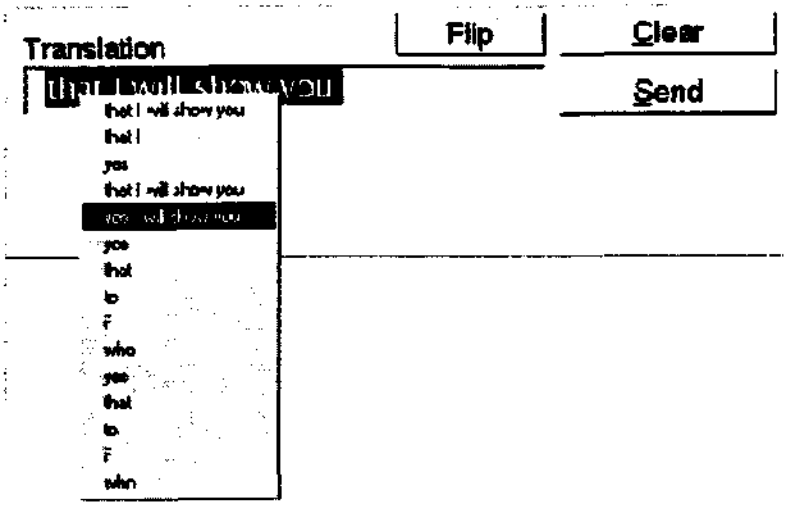

Figure 3: Diplomat Chart-editing Interface

Speech displayed as text: After any speech recognition step, the best overall hypothesis is displayed as text on the screen. The user can highlight an incorrect portion using the touchscreen, and respeak or type it.

Confirmation requests: After any speech recognition or machine translation step, the user is offered an accept/reject button to indicate whether this is "what they said". For MT, backtranslations provide the user with an ability to judge whether they were interpreted correctly. For the **interviewee,** these confirmation requests are presented using very simple displays, which have been designed to be easily customizable for the various languages required.

Interactive chart editing: As mentioned above, the MEMT technology produces as output a chart structure, similar to the word hypothesis lattices in speech systems. After any MT step, the **interviewer** is able to edit the best overall hypothesis for either the forward or backward translation using a popup-menu-based editor (Figure 3), as in our earlier Pangloss text MT system. While we have not yet had an opportunity to test the chart-editing subsystem in Diplomat, the Pangloss version was shown to be an effective tool (Frederking et al. 1993). The editor allows the interviewer to easily view and select alternative translations for any segment of the translation. Editing the forward translation causes an automatic reworking of the backtranslation. Editing the backtranslation allows the interviewer to recognize correct forward translations despite errors in the backtranslation; if the backtranslation can be edited into correctness (using existing alternative translations), the forward translation was probably correct.

However useful it is, a good GUI will not always suffice; for example, a major challenge for handling Haitian Creole (one of Diplomat's languages) is that between 45% and 85% of the Haitian population is illiterate. Moreover, the Haitians who can read generally know French, and so are not as difficult to communicate with. So in order to handle the original Diplomat goals, we will have to develop an all-speech version of the interviewee-side interface. As we have done with previous interface designs, we plan to carry out user tests early in its development to ascertain whether our intuitions on the usability of this version are correct, and to iteratively improve the interface as necessary.

2.3 Speech recognition and synthesis

The speech understanding component used in Diplomat is the Sphinx II HMM-based speaker-independent continuous speech recognition system (Huang et al. 1992. Ravishankar 1996). with newly-developed techniques for rapidly developing acoustic, lexical, and language models for new languages (Eskenazi et al. 1998, Rudnicky 1995, Damiba and Rudnicky 1998).

The speech synthesis component is Phonebox. a newly-developed concatenative system (Lenzo et al 1998) based on variable-sized compositional units. Its use of subword concatenation is especially important, since this is the only currently available method for rapidly bringing up synthesis for a new language.

These components of Diplomat are less relevant to the main points of the current paper than the preceding components, and are described more fully elsewhere (Frederking et al. 1997, Frederking et al. 1999).

3 Extension of Diplomat to Translating Telephone

Given the technology base described in Section 2 above, we recently realized that a relatively small additional effort would allow us to attack the Translating Telephone problem. Diplomat was initially envisioned as a portable translator for field use; but from a software point of view, there is no inherent reason why the MT and speech processing would need to be located at the user's site; only the physical interfaces and suitable communication channels would need to be there. A remote computer, equipped with the necessary communication channels and interfaces, could

host the speech translation software for multiple conversations.

To support an interesting Translating Telephone application, the physical interfaces and communication channels need to be widely available. Telephones can obviously provide widely available speech interfaces (although with reduced recognition accuracy, due to the low fidelity of telephone equipment, compared to a close-talking microphone). But although we plan to eventually move Diplomat to an all-spoken interface, we currently depend on using a GUI for efficient user-correction of errors. A key realization was that a web browser running on a home computer could provide the GUI that we (currently) need for user interaction. The recent wide-spread availability of Internet connections and web browsers is thus an important aspect of our approach.

We will first describe the user's viewpoint, and then the implementation of our new extensions. In our current prototype, the person initiating the conversation (the *first party)* connects to a "Diplomat Translating Telephone" URL using a browser. After the first party enters language preference and contact information for the person he wants to call into an HTML form, the system contacts the *second party* either by email or telephone. If the second party wishes to talk, they use a browser to connect to a second URL that they are given when contacted, and use their phone to call into a number they were given (or, if they were contacted by telephone, they simply do not hang up). The second party is presented with a web page that essentially replicates the current Diplomat GUI (see Figure 4). The first party is then given a phone number to call, and their browser then displays a new page, which is similar to the second party's, but with reversed language direction . At this point, both parties have telephone and web connections to the server. The speech recognition, translation, and user interaction can then all proceed in the same general fashion as described in Section 2 above[4].

This prototype system has been implemented using "plug-ins" and Java applets through a standard web browser, such as Netscape, plus an audio connection that could be provided by an ordinary voice telephone. (The initial prototype uses a close-talking microphone, since we have not yet adapted our system to telephone-quality speech.) The translation subsystem had already been implemented using a client/server architecture, and the Sphinx II speech recognizer had already been adapted for use in an earlier plug-in application. The speech synthesis component is initially being used as-is in the new system, but we may need to adapt it to function as a plug-in, if we wish it to function well over low-bandwidth network connections such as dial-up lines. Thus the only major new developments required were a web server built in Java and a user interface built from HTML and Java applets. The server sets up the initial connections described above and coordinates the various other components. The user interface applet directly substitutes for the GUI in the wearable-translator that Diplomat has been working with.

In addition to serving as a Translating Telephone, this new version of the system allows us to translate speech between two people in a single location, without any special equipment or advance software installation. As long as someone knows the URL. they can start up the translator just by bringing up a web browser and downloading our plug-in, wherever they are. Thus speech translation will be immediately available to anyone, anywhere there is a web connection and a telephone. As part of this vision, we also intend to produce online tutorial web pages, which can be skipped by experienced users, to alleviate any need for the user to carry a manual.

4 Conclusion

By working from a new strategic direction, we believe that we may relatively quickly achieve a Translating Telephone with a useful level of capability in several language pairs. The system initially requires a significant amount of human-intervention, but we expect that to decrease significantly over time. We find the prospect of making Translating Telephones available to anyone who knows the URL very exciting. The recent wide-spread availability of web browsers is clearly an important aspect of our approach; a few years ago, the system as envisioned here would not have been usable in any reasonable number of households. We have begun work in this new direction, and expect to demonstrate initial capabilities at the MT Summit.

References

Brown, R. (1996). "Example-Based Machine Translation in the Pangloss System." In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96).*

Brown, R. and Frederking, R. (1995). Applying Statistical English Language Modeling to Symbolic Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95),* pp. 221-239.

Constantinides, P., Hansema, S., Tchou, C., and Rudnicky, A. (1998.) A Schema Based Approach to Dialog Control. In *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP-98),* Sydney, Australia.

[3] This is actually just one possible configuration. Another possibility is that the correction interface is separated out, and a trained translator provides the repairs; this follows the Communicator/Supervisor model described in Section 1.2. and would only be feasible on a large scale once the translation quality reached a level that required less intervention.

[4] We are currently experimenting with different configurations, which resemble either the current asymmetric Diplomat configuration or the symmetric interaction used in the original Diplomat prototype, to varying degrees.

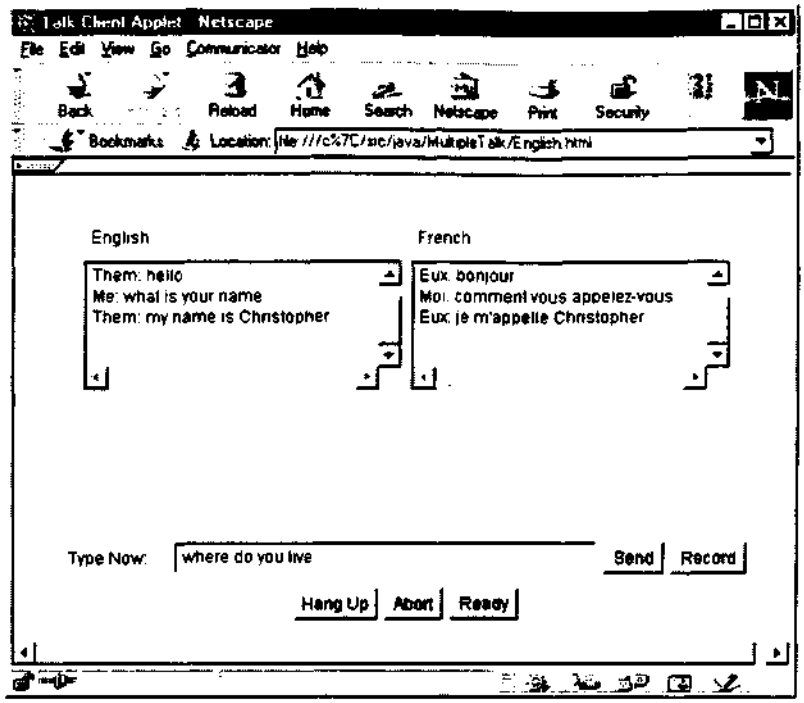

Figure 4: HTML-version of User Interface

Damiba, B. and Rudnicky, A. (1998). Language-Independent Lexical Acquisition. Unpublished manuscript. Available at: http://www.cs.cmu.edu/~air/papers/DamibaRudnickv98.pdf.

Eskenazi. M.. Hogan. C.. Allen, J., Frederking. R. (1998). Issues in Database Design: Recording and Processing Speech from New Populations. In *Proceedings of the First International Conference on Language Resources and Evaluation,* vol. 2. pp. 1289-1293. Granada. Spain.

Frederking. R., Grannes, D., Cousseau, P.. and Nirenburg. S. (1993). An MAT Tool and its Effectiveness. In *Proceedings of the DARPA Human Language Technology Workshop,* Princeton, NJ.

Frederking, R. and Nirenburg, S. (1994). Three Heads are Better than One. In *Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP-94)*. Stuttgart, Germany.

Frederking, R., Nirenburg, S., Farwell, D., Helmreich, S., Hovy, E., Knight, K., Beale, S., Domashnev, C., Attardo, D., Grannes, D., Brown, R. (1994). Integrating Translations from Multiple Sources within the Pangloss Mark HI Machine Translation System. In *Proceedings of the First AMTA Conference,* pp. 73-80, Columbia, MD.

Frederking, R., Rudnicky, A., and Hogan, C. (1997). Interactive Speech Translation in the Diplomat Project. In *Proceedings of the Workshop on Spoken Language Translation,* pp. 61-66, Madrid, Spain. ACL/ELSNET.

Frederking, R., Rudnicky, A., Hogan, C., and Lenzo K. (1999). Interactive Speech Translation in the Diplomat Project. *Machine Translation Journal: Special Issue on Spoken Language Translation.* To appear.

Hogan, C. and Frederking. R. (1998). An Evaluation of the Multi-Engine MT Architecture. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA-98).* Langhome, PA. Springer-Verlag.

Huang, X., Alleva, F., Hon, H.W., Hwang, M.Y., and Rosenfeld, R. (1992). The Sphinx-II Speech Recognition System: An Overview. Technical Report CMU-CS-92-112, Carnegie Mellon University.

Kay, M. (1967). Experiments with a Powerful Parser. In *Proceedings of the Second International COLING.*

Lenzo, K., Hogan, C., and Allen, J. (1998). Rapid-Deployment Concatenative Speech Synthesis in the Diplomat System. In *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP-98),* Sydney, Australia.

Moritomi, T., Takezawa, T., Yato, F., Sagayama, S., Tashiro, T., Nagata, M., and Kurematsu, A. (1993). ATR's Speech Translation System: ASURA. In *Proceedings of the Third Conference on Speech Communication and Technology,* pp. 1295-1298, Berlin, Germany.

Nirenburg, S. (ed.). (1995). The Pangloss Mark III Machine Translation System. Technical Report issued

as CMU-CMT-95-145, Computing Research Laboratory (New Mexico State University), Center for Machine Translation (Carnegie Mellon University), Information Sciences Institute (University of Southern California).

Ravishankar, M. (1996). Efficient Algorithms for Speech Recognition. PhD thesis. Carnegie Mellon University.

Rudnicky, A. (1995). Language Modeling with Limited Domain Data. In *Proceedings of the ARPA Workshop on Spoken Language Technology,* pp. 66-69. San Mateo.

Waibel, A., Jain, A., McNair, A., Saito, H., Hauptmann, A., and Tebelskis, J. (1991). JANUS: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-91),* vol. 2, pp. 793-796, Toronto, Canada. IEEE.

Winograd, T. (1983). *Language as a Cognitive Process.* Volume 1: Syntax. Addison-Wesley.

Woszczyna, M., Aoki-Waibel, N., Buo, F.D., Coccaro, N., Horiguchi, K., Kemp, T., Lavie. A., McNair, A., Polzin, T., Rogina. I., Rose, C.P., Schultz, T., Suhm, B., Tomita, M., and Waibel, A. (1994). Towards Spontaneous Speech Translation. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-94),* pp. 345-349. Adelaide. Australia. IEEE.