# Cooperative Spoken Language Understanding for Robust Speech Translation

**Mark Seligman**
GETA-CLIPS
Université Joseph Fourier
Grenoble, France
and
Spoken Translation, Inc.
mark.seligman@
spokentranslation.com

**Mike Dillinger**
Spoken Translation, Inc.
1100 West View Dr.
Berkeley, CA 94705
mike.dillinger@
spokentranslation.com

**Chengqing Zong**
National Laboratory of
Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
P. O. Box 2728
Beijing 100080, China
cqzong@nlpr.ia.ac.cn

## Abstract

This paper argues that the time is now right to field practical and robust Spoken Language Understanding (SLU) systems. It argues that, at the present state of the art, robustness can best be achieved through user cooperation and compromise with the system. If this insight guides design, several sorts of reliable SLU systems can be deployed over the next few years. Further, SLU systems can be arranged on a scale, in terms of the degree of cooperation or compromise they require from users. In general, the broader the intended linguistic or topical coverage of a system, the more user cooperation or compromise it will presently require. The authors focus on Spoken Language Translation as a specific example application.

## 1 Introduction

We take the field of spoken language understanding (SLU) to study the combination of speech recognition and language analysis, where analysis can be broadly understood to include syntactic parsing, the derivation of semantic representations, and information extraction.

For concreteness, this paper will focus on Spoken Language Translation (SLT) as one specific application of SLU technology. In this application, the analysis output supplied by SLU is subjected to further processing in order to derive a translation. (In many systems, the derived translation is then pronounced using text-to-speech software, so that speech-to-speech translation is achieved.) Thus the success of spoken language translation depends directly on the quality of this output.

At present, most systems combining SR and language analysis still remain in the laboratory. This paper, however, will urge that the time is now right to field practical SLT and other spoken language understanding systems. The key to sufficient robustness for practical use, we believe, is the recognition that, at the present state of the art, users must cooperate and compromise with the programs. We further suggest that SLU systems can be usefully ordered, or arranged on a scale, in terms of the degree of cooperation and/or compromise they require from users. We point out a direct relation between the degree of cooperation a system requires and its intended linguistic coverage or topical coverage. (By linguistic coverage, we mean the system's tolerance for freely varied sentence structures and vocabulary, as opposed to memorized or fixed phrases. By topical coverage, we mean the system's ability to translate language outside of a single narrow domain, such as the making of hotel reservations.) We will argue that, in general, the broader the intended linguistic or topical coverage of the system, the more user cooperation or compromise it will presently require.

We begin in Section 2 by presenting our case for extensive user cooperation in robust SLU at the present state of the art. Section 3 very briefly surveys various sorts of such cooperation. Section 4 describes three classes of spoken language understanding applications – specifically, spoken language translation systems – whose varying degrees of linguistic and topical coverage derive from their varying degrees of cooperation. Finally, Section 5 presents our conclusions.

## 2 The Need for User Interaction

This section presents our argument that extensive user cooperation is presently necessary for robust SLU. As a reference point for further discussion, we can consider idealized spoken language understanding systems requiring *no* user cooperation or compromise at all – and immediately reject them as over-ambitious for short-term deployment. For the translation task, these might be communicators straight out of Star Trek or the Hitchhiker's Guide to the Universe, providing effortless and transparent translation. Equipped with such communicators, you could talk just as you habitually do when addressing fellow native speakers of your own language: you could freely shift topics; use your full range of vocabulary, idioms, and structures; use extemporaneous language full of fragments, false starts, and hesitations; mumble; converse in noisy environments; and completely ignore the translation program. Meanwhile, your foreign-language-speaking partner would hear a simultaneously and perfectly translated version of your utterances.

Unfortunately, we believe that this vision is unlikely to be achieved soon. In the first place, high-quality spoken language understanding, especially at the level required for translation, appears to require a great deal of linguistic and extra-linguistic knowledge, efficiently integrated from moment to moment; however, it is unclear for current systems when this amount of knowledge can be added and this degree of integration achieved.

We can go on to note that even human interpreters who do have the requisite knowledge and integration capabilities rarely perform as transparently as the Star Trek gadget: they often understand and translate interactively. That is, when they fail to immediately understand an utterance because ill-formed expressions, unknown vocabulary, or other problems have been encountered, they often ask the speaker for repetition or further explanation (Oviatt and Cohen, 1991). Further, human interpreters, like SLU programs, are subject to processing overload: they often skip or mistranslate the input in order to keep up with the speaker (Dillinger, 1994). Thus it is unrealistic to expect an SLT system, for example, to outperform human interpreters by producing correct translation results without ever clarifying the speaker's intention, resolving certain ambiguities, etc. We believe that the same point can be made for SLU systems in general.

These reservations are not meant as discouragement. On the contrary, our main point is precisely that SLU technology in general, and SLT technology in particular, can after all be put into practical use soon. We do believe, however, that realistic expectations are crucial for the field at present, since perfectionism or exaggerated aims can delay the development of usable but less dramatic systems.

Of course, at a very high level, the nature of the SLU problem is clear enough: the component technologies (speech recognition, parsing, word sense disambiguation, etc.) are still quite imperfect when considered individually, and are quite challenging to effectively integrate. When the particular application is spoken language translation, additional translation processes are also involved. While we believe that most of these individual technologies have now passed the point of usability, at least for some applications, their combination may nevertheless fall below the usefulness threshold, since their error rates generally combine, and sometimes compound.

Effective component integration seems likely to elude us for some time to come. Ideally, we would like to understand (and e.g. translate) speech through some seamless combination of the technologies. Presently, however, most working systems maintain a division between speech recognition and language analysis. (Working experimental SLT systems generally maintain at least three separate components: speech recognition, machine translation – incorporating its own analysis routines – and text-to-speech generation.)

This separation between components implies that speech input must be transformed into text before the machine translation's analysis component can begin to handle it: the system cannot analyze or translate incrementally, while speech recognition is progressing. Thus any information which might be gained by deeply analyzing input utterances will for now remain unavailable to current speech recognition components. This lack of information can only increase the error rate during recognition. In addition, when these recognition errors are passed to the machine translation component, they can only increase its burden, even if robust parsing techniques are applied.

The separation between components is clearly artificial, and can be seen as a temporary tactic to be employed only until we learn enough to implement more integrated approaches. (For discussion of some issues facing such tight integration, see Seligman (2000).) However, this learning process is likely to continue indefinitely. While close integration of speech recognition and translation remains an important research goal for the middle and long term, it would be unwise to postpone the building of practical systems until it is achieved. Programs which aim to produce practical systems in the near term must treat the separation between speech recognition and language analysis as a fact of life.

When we add to these challenges the current lack of robust multiple knowledge sources which may eventually improve the automatic operation of the language analysis and translation components themselves, the present need for user cooperation and compromise becomes clear. These accommodations are presently the best ways to compensate for the inevitable shortcomings of state-of-the-art systems. As systems mature, the need for them should gradually lessen. For now, however, SLU systems which aim for fully transparent or effortless operation in the Star Trek mode will remain out of reach.

## 3 Cooperative Spoken Language Processing

We can assume that, for the foreseeable future, some degree of user cooperation and compromise will be required for robust and practical spoken language understanding systems, e.g. for spoken language translation.

What sorts of cooperation and compromise might users be expected to contribute? Depending on the system design, they might for example be asked to:

- speak loudly and clearly
- use a standard dialect
- speak in quiet environments
- accept restrictions on the use of audio input equipment, networks, etc.
- correct speech recognition errors, by voice or by typing
- train speaker-dependent acoustic models
- provide only syntactically well-formed input
- use vocabulary and sentence structures from a limited array
- provide extra information in the input to aid analysis, e.g. word separations or brackets
- resolve lexical or structural ambiguities (by taking the initiative, or by responding to prompts from the system)
- provide missing information, e.g. referents for zero pronouns
- tolerate the results of rough or incomplete understanding, e.g. as they affect translations
- spell or type out words that prove hard to recognize
- use richer or more complex interfaces, e.g. including multi-modal GUIs as opposed to voice-only

In fact, several interactive approaches to spoken language translation have been proposed along these lines (Boitet, 1996; Blanchon, 1996; Waibel, 1996; Seligman, 1997a, 1997b, 2000; Ren and Li, 2000).

## 4 Three Classes of cooperative SLT Systems

Having very briefly surveyed some possibilities for user cooperation, we now present three classes of SLU systems – more specifically, SLT systems – which we believe can be constructed and put into practical use in the next three or four years. The three classes require increasing amounts of user cooperation or compromise, and in return offer increasing degrees of linguistic coverage (freedom of expression) and topical coverage (freedom to switch domains). We certainly do not suggest that our discussion will cover the full range of possible system configurations, but we do believe that these classes represent likely developments.

### 4.1 Class One: Voice-Driven Phrasebook Translation

*linguistic coverage: narrow*
*topical coverage: narrow*
*cooperation required: low*

It is possible to collect useful phrases – for ordering at restaurants, for checking into hotels, for responding to questions at customs inspections, and so on – and to store them in a suitable database along with prepared translations. A speech translation system can then directly search this phrase translation database.

A system in this class (see for instance Zong *et al.* (2000)) would allow little or no freedom of expressive choice. Instead, users would choose among the stored phrases. To gain slightly more freedom, they might also select templates, as in many existing (and profitable) travel phrasebooks in the style popularized by Berlitz and others. (A template contains at least one variable element with a specified range of values, e.g. *I'd like a bottle of [beer, wine, soda], please*. Or *I'd like a bottle of [BEVERAGE], please*.) As increasing numbers of variable elements are introduced, this translation memory-based approach grades into an example-based approach (compare Nagao, 1984; Sato, 1991; Sumita and Iida, 1992; Brown, 1996; Iida, 1996).

While a phrase-oriented system would probably offer a range of standard travel-oriented situational topics, the system would be useless outside of those situations (though extensions to new situations, or additions to existing ones, would always be possible). From the user's viewpoint, such a system would have several important advantages over a classical printed phrase book:

- there would be no need to carry a book: instead, the system could be accessed by standard telephone or the using a hand-held device;
- selection of phrase or template would be made by voice rather than by leafing through the pages or using an index; and
- the translation output would be pronounced by a native, rather than simply being printed and pointed out to target-language readers (or bravely spoken by the user with the aid of imperfect pronunciation tips, as in "PAR-lay voo ahn-GLAY?").

**Interface design for voice-driven phrase translation** Voice-driven phrase translation systems would be unlikely to perform well if users were simply invited to pronounce any phrase they'd like to have translated. Instead, some guidance would normally be provided for the user.

For instance, in a telephone environment, voice-driven dialogs (designed with VoiceXML or similar authoring systems) might guide users toward the phrases or templates a system could handle, gradually converging on the desired phrase the manner of a decision tree. A preliminary voice prompt might present a range of topics to be selected (by voice). Once the main topic had been chosen, subtopics might be successively presented using voice prompts. Finally, three or four phrases or templates might be pronounced, from which a final selection could be made vocally. The prepared translations and their pronunciations would then be produced.

Alternatively, a voice-driven phrase translation system equipped with a graphic user interface, e.g. in a handheld format, could present a menu of narrow areas within which very typical utterances might be blindly tried. The system could then return several of the fixed phrases in its database which most closely match the input. If one of these approximated the user's original intention, it could be selected from the short list. This alternative approach is in fact been developed for the open market by two companies at present: NEC (Okumura, 2003) and VoxTec (a division of Marine Acoustics, Inc.).

**Technology for voice-driven phrase translation**

The required technology for voice-driven phrase translation is all in place and commercially available.

- For speech recognition, the speaker-independent, grammar-based technology now widely used for interactive voice response (IVR) systems would be appropriate. Fixed phrases could appear in the current grammar as flat strings, while templates could include sub-grammars representing the variables.
- For translation, a range of memory-based approaches could be employed: at minimum, simple databases associating phrases with their prepared translations; at maximum, sophisticated example-based or generalizing approaches. At a middle level of sophistication, template-based approaches, in which simple finite-state grammars for translation were synchronized with simple speech recognition grammars, would seem especially promising.

Thus the project becomes mainly an engineering exercise, with emphasis on interface design. Accordingly, we judge the degree of risk in preparing phrase-translation systems during the next few years to be quite low.

### 4.2 Class Two: Robust Speech Translation Within Very Narrow Domains

*linguistic coverage: broad*
*topical coverage: narrow*
*cooperation required: medium*

A system in this class allows users to choose expressive words and structures quite freely, but in compensation handles only a sharply restricted range of topics, e.g. hotel or event reservation. Users can be made aware of those constraints, and need to cooperate by remaining tightly within a pre-selected topic. If inputs are misunderstood (as indicated through a written or synthesized voice response), users can cooperate by repeating or rephrasing. Their reward would be that, instead of having to work down through voice-driven decision trees or otherwise choose among fixed phrases as in a phrase translation system, they could express their needs relatively naturally, even to the extent of using hesitation syllables or fragments. For instance, any of the following utterances might be used to begin a reservation dialog:

*Uh, could I reserve a double room for next Tuesday, please?*
*I need to, um, I need a double room please. That's for next Tuesday.*
*Hello, I'm calling about reserving a room. I'd be arriving next week on Tuesday.*

Such relatively broad linguistic coverage can be achieved through a variety of robust analysis techniques: the recognized text is treated as a source for pattern-driven information extraction, from which programs attempt to extract only crucial elements for translation, ignoring less-crucial segments. Thus all of the above utterances might be recognized as instances of some specialized speech act, e.g. ROOM-RESERVATION-REQUEST, by recognizing that *reserve* and *room* occur within them, along with such speech act markers as *could I*, *I need*, *I'm calling about*, etc. Given this particular speech act, the system can attempt to extract such specific information as ROOM-TYPE (here, a double room); ARRIVAL-DATE (here, Tuesday of the following week), etc. Narrowly restricted responses, e.g. from a reservation agent, can be treated in a comparable way.

**Advantages of Class Two systems**

Class two systems have several advantages from the system builder's viewpoint.

Such systems have been widely studied: most experimental speech translation systems have in fact been of this type (Corazza *et al*., 1999; Lavie *et al*., 1999; Sugaya *et al*., 1999; Zong *et al*., 1999; Wahlster, 2000). Thus a good deal of practical experience in building them has been gained, and could be quickly applied in the effort to field practical systems in the near term.

Another major advantage is that, since systems of this sort remain narrowly restricted with respect to topic, it has proven practical to incorporate within them technology, for both speech recognition and translation, which can be tightly tuned or optimized for particular applications.

With respect to speech recognition, as for Class One systems, it is possible to exploit standard speaker-independent, grammar-based technology. Because the range of input utterances is sufficiently limited, one can construct finite-state grammars covering significant segments of the relevant domain.

Concerning translation technology, the narrowness of the selected domain makes possible the use of interlingua-based translation approaches, those using pragmatic and semantic representations applicable for several languages. The most widely used interlingua within the current speech translation community is the Interchange Format, or IF, used by members of the C-STAR consortium. Other interlinguas, for example the UNL representation propounded by UNU (http://www.unl.ru/) could also be tried. However, whichever representation is chosen, the well-known advantage of this translation style is gained: multiple translation directions can be quickly built.

**Challenges of Class Two systems**

But of course, Class Two systems face several challenges as well.

To date, most systems stressing robust recognition with narrow topical coverage have not required users to correct speech recognition results before sending them to analysis programs. Perhaps it has been felt that users were already compromising enough in terms of topical coverage, and thus should not be required to expend this additional editing effort as well. In any case, if this approach continues to be followed, a certain amount of noise must be expected in the input to translation. At the same time, robust parsing remains an area at the forefront of research rather than a mature technology, and can itself be expected to provide noisy or imperfect results.

Thus the degree of risk in this approach must be considered at least medium: practical systems of this sort can probably be built soon given sufficient resources, but considerably more research and development time will be needed than for phrase translation systems. Further, since errors are inevitable, user frustration is likely to be somewhat greater. The hope is that this degree of frustration (along with the stringent restriction as to topic) will be tolerated in exchange for greater freedom of expression, i.e. linguistic coverage.

### 4.3 Class Three: Highly Interactive Speech Translation with Broad Linguistic and Topical Coverage

*linguistic coverage: broad*
*topical coverage: broad*
*cooperation required: extensive*

A speech translation system in this third class allows users to choose expressive structures quite freely, and allows free movement from topic to topic. However, to maintain sufficient output quality while permitting such freedom, the system requires that users monitor, and when necessary correct, the progress of both speech recognition and language analysis for translation. In effect, at the present state of the art, users of a Class Three system must pay for expressive and topical freedom by spending considerable time and effort to help the system.

Where speech recognition is concerned, since very broad-coverage speech recognition is required, dictation technology appears most promising. Acoustic models (individual voice profiles) must be created for all users (where each such registration takes several minutes); and underlying language models for running text must include at least tens of thousands of words, probably by using *n*-grams rather than grammars. Users must also expect to correct recognition results for each utterance (by using voice commands or by typing) before passing corrected text to the translation stage. While considerable linguistic freedom is allowed, spontaneous speech features are not: roughly any standard grammar and vocabulary can be used, but hesitation syllables or stutters may be misrecognized as words, repetitions and false starts will appear in the visible text output, etc.

Where language analysis for translation is concerned, users must provide standard, well-formed text for input: fragments or false starts will degrade quality. They must also be prepared to help the system to resolve ambiguities, e.g. by indicating the desired meaning of an ambiguous input word, or the desired attachment of an input prepositional phrase. For some translation directions, users will also need to supply information missing in the input, e.g. the referents of zero pronouns when translating from Japanese.

There are two main technical requirements for the translation engine: (1) it must provide very broad coverage with sufficient output quality; and (2) it must allow interruption of the translation process, so that users can interactively provide lexical and structural constraints. In practice, these attributes are found mostly in mature commercial transfer-based or interlingua-based systems; but some statistical systems may also prove suitable.

Class Three spoken language translation systems have been demonstrated successfully, though with some significant limitations. The demonstrations of "quick and dirty" English<>French speech translation presented by Seligman (2000) in cooperation with CompuServe, Inc. fell in this category.

**Challenges of Class Three speech translation systems**

In the demos just mentioned, speech recognition and text-to-speech remained client-based. For maximum flexibility of use, they should instead be server-based, as were the translation facilities. Further, no interactive correction of translation (as opposed to dictation) was yet enabled.

Two of us (Seligman and Dillinger) are now pursuing commercial research and development addressing both of these limitations. Both efforts present significant challenges.

Where speech recognition is concerned, while commercial systems for server-based dictation have recently become available, current systems do not yet support immediate interactive feedback and correction. (Users must instead dictate blindly by telephone, and make corrections later using client-based programs.) In order to create remote dictation systems with real-time correction, engineers will need to eliminate disruptive lags due to transmission inefficiencies. They will also need to implement complex client-server interactions to maintain synchronization as dictation results are obtained by servers and displayed by clients.

With regard to interactive translation, while there has been substantial research in this direction (Melby, 1987), few interactive translation systems have come into practical use. (One pioneering exception can be seen in the online demo of a commercial English<>Spanish system at http://www.wordmagicsoft.com.) The present work of Seligman and Dillinger's group focuses upon interactive lexical disambiguation as especially significant and potentially tractable.

A further challenge for highly interactive speech translation systems relates to the burden imposed by the interaction itself. Will users accept the inconvenience of monitoring and correcting dictation and translation? For socializing, they may choose to pass up such oversight entirely and to tolerate the resulting errors. However, for serious communication, we anticipate that the degree of tolerance for interaction will rise with the importance of quality results. Research is needed to confirm this expectation and explore ergonomic issues. We plan to undertake such usability testing during 2004 in cooperation with the GETA-CLIPS organization of the Université Joseph Fourier in Grenoble, France.

In view of these challenges, the degree of risk for a Class Three speech translation system must be considered medium to high. Nevertheless, Seligman and Dillinger hope to present fully server-based prototype systems for English<>German and English<>Spanish by late 2004. We believe that a well-supported development effort could bring these and several additional highly interactive, broad-coverage speech translation directions into practical use shortly after.

## 5 Conclusion

The central theme of this paper has been that several sorts of practical and sufficiently robust Spoken Language Understanding (SLU) systems can be built over the next few years if system designers make adequate provision for user cooperation and compromise. We have suggested that SLU systems can be arranged on a scale, in terms of the degree of user cooperation and/or compromise they require. Generally, the broader the intended linguistic or topical coverage of the system, the more such cooperation or compromise will presently be needed.

After very briefly reviewing some ways in which users can cooperate with SLU systems, we have described by way of example three classes of "cooperative" Spoken Language Translation (SLT) systems that we believe can be put into practical use during the next several years.

Overall, we are arguing that the spoken language understanding field has reached a point at which basic and applied research can and should be distinguished. On the basic research side, it is clear that many unresolved problems remain with respect to system components – speech recognition, parsing, lexical and structural disambiguation, extraction of semantic structures – and their integration. Until considerable further progress is made along these lines, we doubt that the goal of fully automatic, high-quality, broad-coverage SLU can be reached. On the side of applied research and engineering, however, it should now be possible to develop a variety of useful SLU applications in specific domains. The field of Spoken Language Translation furnishes an excellent example.

While these applications, and the products created from them, will be far from perfect, they can nevertheless provide useful aids for human-machine and human-human communication.

SLU applications can be useful without meeting all needs. Thus, while SLU technology is not yet ripe, its immaturity no longer justifies keeping it within the lab. On the contrary: many new technologies can only mature through actual use.

And communication – after all, the original purpose of SLU – is inherently a cooperative process. Conversational partners have always cooperated to make themselves understood. In practical and robust SLU systems in general, and in SLT systems in particular, humans and computers should likewise cooperate, collaborate, and mutually accommodate.

## References

Blanchon, Hervé. 1996. A Customizable Interactive Disambiguation Methodology and Two Implementations to Disambiguate French and English Input. In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*. Col de Porte, France.

Boitet, Christian. 1996. Dialogue-based Machine Translation for Monolinguals and Future Self-explaining Documents. In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*. Col de Porte, France.

Brown, Ralf. 1996. Example-based Machine Translation in the Pangloss System. In *Proceedings of COLING-96* (pp. 169-174). Copenhagen, Denmark.

Corazza, Anna, Mauro Cettolo, *et al*. 1999. The ITC-IRST Speech Translation System. *Proceedings of the C-STAR II Workshop*. Schwetzingen, Germany.

Dillinger, Mike. 1994. “Comprehension during interpreting: what do interpreters know that bilinguals don’t?” In: Lambert, S. and Moser-Mercer, B. (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation* (pp. 155-190). Amsterdam: John Benjamins.

Iida, Hitoshi, Eiichiro Sumita, and Osamu Furuse. 1996. Spoken Language Translation Method Using Examples. In *Proceedings of COLING-96* (pp. 1074-1077). Copenhagen, Denmark.

Lavie, Alon, Lori Levin, *et al*. 1999. The JANUS-III Translation System: Speech-to- Speech Translation in Multiple Domains. In *Proceedings of COLING-96* (pp. 1074-1077). Copenhagen, Denmark.

Melby, Alan K. 1987. “On Human-Machine Interaction in Translation” In: Nirenburg, S. (Ed.), *Machine Translation* (pp. 145-154). New York: Cambridge University Press.

Nagao, Makoto. 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle In: Eithorn, A. and Banerji, R. (Eds.), *Artificial and Human Intelligence* (pp. 173-180). Amsterdam: North-Holland.

Okumura, Akitoshi. 2003. *Development of Speech Translation for Hand-held Devices*. Presented at Machine Translation Summit IX, New Orleans, USA.

Oviatt, Sharon and Philip Cohen. 1991. Discourse Structure and Performance Efficiency in Interactive and Non-Interactive Spoken Modalities. *Computer Speech and Language 5*(4): 297-326.

Ren, Fuji and Shigang Li. 2000. Dialogue Machine Translation Based upon Parallel Translation Engines and Face Image Processing. *Journal of Information, 3*(4): pp.521-531.

Sato, Satoshi. 1991. *Example-based Machine Translation*. Unpublished PhD dissertation. Kyoto University, Kyoto, Japan.

Seligman, Mark. 1997a. Interactive Real-time Translation via the Internet. In: *Working Notes, Natural Language Processing for the World Wide Web* (pp. 142-148). AAAI-97 Spring Symposium, Stanford University, Palo Alto, CA.

Seligman, Mark. 1997b. Six Issues in Speech Translation. *Proceedings of the ACL-97 Workshop on Spoken Language Translation* (pp. 83-89). Universidad Nacional de Educacin a Distancia, Madrid, Spain.

Seligman, Mark. 2000. Nine Issues in Speech Translation. *Machine Translation, 15*: 149-185.

Sugaya, Fumiaki, Toshiyuki Takezawa, A. Yokoo, Yoshinori Sagisaka and Sei-ichi Yamamoto. 1999. End-to-end Evaluation in ATR-MATRIX: Speech Translation System between English and Japanese. In *Proceedings of ESCA, Eurospeech 1999* (pp. 2431-2434).

Sumita, Eiichiro. and Hitoshi Iida. 1992. Example-based Transfer of Adnominal Particles into English, *IEICE Transactions on Information Systems, E75-D*(4): 585-594.

Wahlster, Wolfgang. 2000. “Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System”. In *Verbmobil: Foundations of Speech-to-Speech Translation* (pp. 3-21). New York: Springer Verlag.

Waibel, Alex. 1996. Interactive Translation of Conversational Speech. In: *Proceedings of ATR International Workshop on Speech Translation* (pp. 1-17).

Zong, Chenqing, Taiyi Huang and Bo Xu. 1999. “Technical Analysis on Automatic Spoken Language Translation Systems” (in Chinese). *Journal of Chinese Information Processing, 13*(2): 55-65.

Zong, Chenqing, Yumi Wakita, Bo Xu, Kenji Matsui and Zhenbiao Chen. 2000. Japanese-to-Chinese Spoken Language Translation Based on the Simple Expression. In *Proceedings of International Conference on Spoken Language Processing (ICSLP-2000)* (pp. 418-421). Beijing, China.