# Combining Human Inputters and Language Services to provide Multi-language support system for International Symposiums

**Takao Nakaguchi**
Graduate School of Informatics
Kyoto University
nakaguchi@i.kyoto-u.ac.jp

**Masayuki Otani**
Graduate School of Informatics
Kyoto University
m-otani@i.kyoto-u.ac.jp

**Toshiyuki Takasaki**
NPO Pangaea,
Graduate School of Informatics
Kyoto University
toshi@pangaean.org

**Toru Ishida**
Graduate School of Informatics
Kyoto University
ishida@i.kyoto-u.ac.jp

## Abstract

In this research, we introduce and implement a method that combines human inputters and machine translators. When the languages of the participants vary widely, the cost of simultaneous translation becomes very high. However, the results of simply applying machine translation to speech text do not have the quality that is needed for real use. Thus, we propose a method that people who understand the language of the speaker cooperate with a machine translation service in support of multilingualization by the co-creation of value. We implement a system with this method and apply it to actual presentations. While the quality of direct machine translations is 1.84 (fluency) and 2.89 (adequacy), the system has corresponding values of 3.76 and 3.85.

## 1 Introduction

Multi-language support is used to reduce the language barriers in international symposia whose participants come from various countries and who speak different languages. The de facto multi-language support tool is simultaneous translation by human translators. Simultaneous translation is a very demanding task, especially between languages with different structures like Japanese and English, and costs are high because it takes long time to train the translators. In simultaneous translation, translators listen to the speech, translate the text, and then speak the translation result while closely following the speaker. Several studies have attempted to replace human translators with relatively low cost systems. Automatic speech recognition (ASR) performs the listening task, machine translation the translation task, and speech synthesis performs the speaking task; speech translation technologies like VoiceTra can perform all tasks. Because machine translation receives text as its input, several captioning schemes, which are normally to allow the deaf and hard of hearing to join in the dialogue, can be candidates for performing the listening task. Figure 1 shows the relationships among these technologies.

Translators convert the speech of the speaker directly into a language some of the audience can understand. Trying to provide complete translation coverage for all speaker/audience combinations is impractically expensive and it may impossible to find translators for some minor languages. One solution is using machines to replace or partly replace the translators. The first challenge is the creating inputs that suit the machine translation (MT) service. MT is widely available at reasonable cost and MT results can be given to the audience as text (text to speech (TTS) systems are also possible). Unfortunately, translation quality is very sensitive to the input material. Given that speeches at public meetings tend to be rather extemporaneous and not so fluent, we need to pre-edit the source text to suit the capabilities of the machine translation service if we are to get good quality. Thus, we propose the method that could combines the listening task with MT. We implement the method in a system that is put into practice in two real fields: presentations at an international convention and presentations at a laboratory. This paper introduces related works, describes our method and our implementation of a multi-language support system, explains how it was put into practice, evaluates and discusses the results.

*Proceedings of WLSI/OIAF4HLT*,
pages 28–35, Osaka, Japan, December 12 2016.

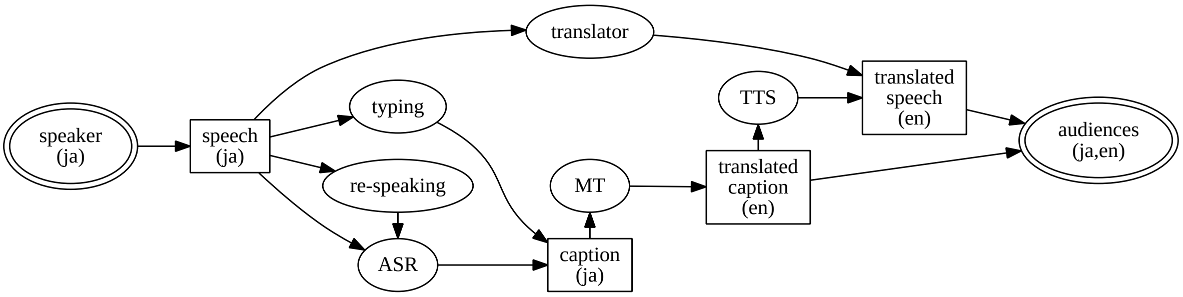


Figure 1: Methods to realize multi-lingualization of speech.

## 2 Related works

Several research projects have tackled speech captioning in real time. ASR is a technology with a long history, but because ASR accuracy is low, we need to train a recognition model, set-up a low-noise environment or use another person to re-speak the text (Miyoshi et al., 2008) to achieve adequate ASR accuracy.

Captioning is a manual way of creating text from speech. The captionist hears the speech and types it into a computer. To achieve adequate speed, the captionist must be well-trained and use a special keyboard, like a steganography keyboard, so costs get high. To reduce this, *SCRIBE* (Lasecki et al., 2012) proposed to use crowd sourcing for captioning. They divide the speech voice data into small parts, send them to many captionists and merge their outputs. This eases the speed requirements. *IPTalk* (Kurita et al., 2013) proposed another method in which several captionist cooperate to input a speech. All captionists hear the same voice data and decide who inputs which part of speech by monitoring what the others are typing by using a type monitor module on the input screen. *SCRIBE* captionists need not be aware of whole speech but can concentrate on that the part of the speech they must input, while each *IPTalk* captionist is aware of the whole speech but input boundaries are decided by predicting the inputs of others.

The goal of this work is to provide near real-time speech translations with high quality and wide coverage at relatively low-cost. To achieve this, we focus on overcoming the problems raised by the target application, MT of symposium speeches. To achieve adequate translation quality, the input text must be recast (pre-edited) to suit the MT service used and then transcribed to allow MT processing. To achieve acceptable quality and speed we modify the cooperative captioning approach.

## 3 Problems of multi-lingual support for international symposium

A key problem is that the sentences uttered in symposia tend to be longer and more ambiguous than those in printed texts. The value of pre-editing as a preliminary step to MT has been confirmed for many languages including English(Jachmann, Grabowski and Kudo., 2014), French(Bouillon et al., 2014) and Japanese(Miyata et al., 2015). However, the studies published to date apply pre-editing to written texts, not to spoken text. Accordingly, we evaluated the quality of translation with or without pre-editing of speech material to confirm the effectiveness of pre-editing. Table 1 shows the results. We used *IPTalk*, a cooperative input method, to do caption the speech and *JServer* in the Language Grid(Ishida, 2011) (Murakami et al., 2012) to translate Japanese into English and Chinese. We evaluated fluency and adequacy based on the method written in the evaluation guideline(Linguistic Data Consortium, 2002).

As the translation inputs, we created sentences by concatenating *IPTalk* outputs and creating sentences by setting periods. As evaluation metrics we used fluency, adequacy and concept preservation ratio; maximum score (best) is 5 and minimum (worse) is 1. We also calculated word accuracy (WA) and concept preservation rate to determine how well the surface meaning and intent of the speech text were

Table 1: Translation quality from Japanese speech dictation by cooperative input with or without preedit.

| sentence creation method | number of sentence | WA | concept saving rate | target language | average fluency | average adequacy |
|---|---|---|---|---|---|---|
| cooperated captioning | 22 | 79.98 | 4.8 | Chinese | 1.77 | 2.68 |
| | | | | English | 1.91 | 3.09 |
| pre-editing afterward | 40 | 42.63 | 4 | Chinese | 3.13 | 3.70 |
| | | | | English | 3.18 | 3.25 |

retained. We did morphological analysis by applying *Mecab* to the *IPTalk* outputs (real-time input was given) and the resulting transcript. WA was calculated by the following formula.

$$WA = 1 - \frac{S + D + I}{N}$$

N denotes the number of words in final transcript, S denotes the number of replaced words, D denotes the number of deleted words and I denotes the number of inserted words relative to real-time input. In captioning, the inputters basically enter the words exactly as spoken, but may delete redundant words, shorten sentences or insert words to allow easier comprehension when read. In terms of modifications there were 56 replacements, 126 deletions, and 9 insertions. Concept preservation rate indicates how well the original concepts were expressed by the final transcript (evaluated by Japanese native speakers). The results in Table 1 show that simply applying MT to *IPTalk* output does not yield good translation quality and we can improve the quality by pre-editing the MT inputs.

## 4 Online Multi-lingual Discussion Tool (OMDT)

Though existing studies considered only static text and not real-time speech texts, the simple modification of simply shortening the text is known to be effective for improving translation quality. From this viewpoint, we design a system that helps inputters to transcribe and pre-edit speech texts for creating MT inputs.

The key components of our system are *Input Screen*, *Collaboration Server*, and *Display Screen*. *Input Screen* is one screen of *OMDT* and inputters use this screen to input speech text. This screen runs on a web browser and we can open this screen as necessary to support inputters and languages. The screen also has the ASR Client function to recognize speech by using ASR services on the Cloud. *OMDT* currently supports IBM, Google and Julius ASR engines. The result of ASR and typed entries are shared between *Input Screen*s and *Display Screen*s via the *Collaboration Server*. The *Collaboration Server* is responsible for message transfer among screens and invocation of the composite translation service. When the server receives typed text from the *Input Screen*, the server invokes DictTrans service to translate it. Though our system can access any translation service on *Language Grid*, we usually use DictTrans as it offers a bilingual dictionary. DictTrans combines translation service, bilingual dictionary service and morphological analysis service and can replace/modify special words in the dictionary to enhance translation quality. The server sends the translation result to the user screens. *Display Screen* shows the typed text and the translation result. That screen can show the results of several languages together and also we can open several screens to show many languages. In addition, *Collaboration Server* supports *IPTalk* as the input interface, so inputters can use it instead of the *Input Screen*.

Figure 2 shows one instantiation of our system. This construction is for translating Japanese speech to English at a symposium that has Japanese speakers with Japanese and English audiences. The ASR result is sent to the *Input Screens* for input processing as shown in Figure 3. *Input Screen* has several display components: Log area, Back translation, Type monitor and Input area in addition to ASR results. Log area shows the input history for the speech, Back translation shows the result of back translation of the text of Input area, Type monitor shows the typing state of other inputters and Input area shows the text currently being pre-edited.

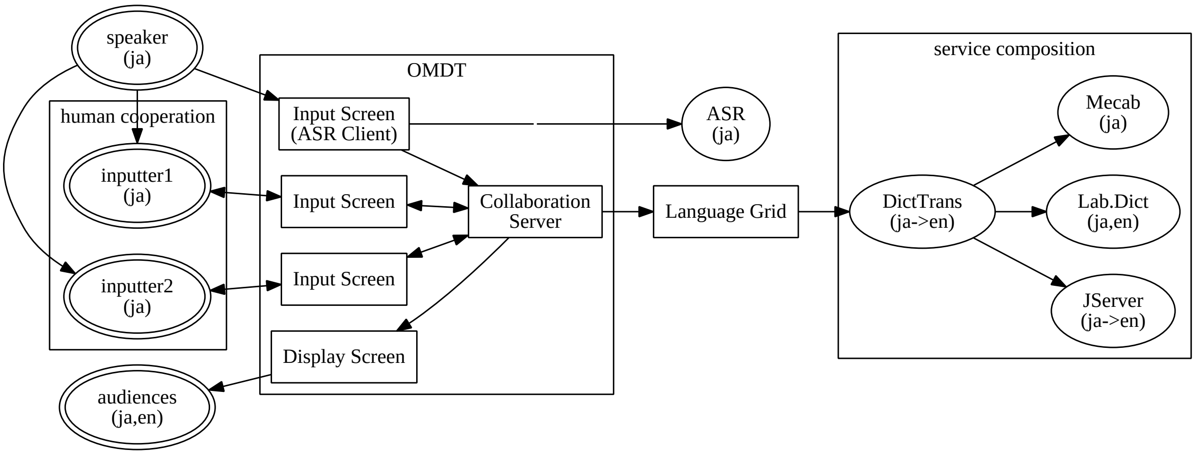
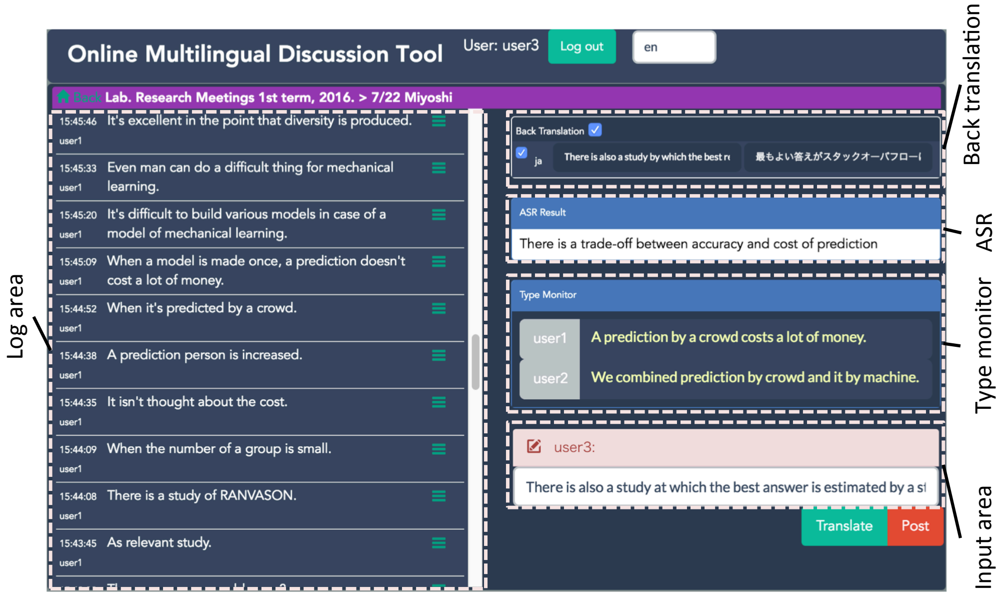


Figure 2: Example construction of whole system

Translated text is finally shown to the audience on the *Display Screen* as shown in Figure 4. We can increase the number of *Input Screens* and *Display Screens* to cover more languages. We can also increase the number of translation pairs the system can translate by adding translation settings (how to combine language services for specific translation pairs) by using *Language Grid*.



Lab. Research Meetings 1st term, 2016. > 7/22 Miyoshi

15:45:46 It's excellent in the point that diversity is produced.
user1

15:45:33 Even man can do a difficult thing for mechanical learning.
user1

15:45:20 It's difficult to build various models in case of a model of mechanical learning.
user1

15:45:09 When a model is made once, a prediction doesn't cost a lot of money.
user1

15:44:52 When it's predicted by a crowd.
user1

15:44:38 A prediction person is increased.
user1

15:44:35 It isn't thought about the cost.
user1

15:44:09 When the number of a group is small.
user1

15:44:08 There is a study of RANVASON.
user1

15:43:45 As relevant study.
user1

Back Translation ☑

☑ ja There is also a study by which the best r 最もよい答えがスタックオーバフロー

ASR Result

There is a trade-off between accuracy and cost of prediction

Type Monitor

user1 A prediction by a crowd costs a lot of money.

user2 We combined prediction by crowd and it by machine.

user3:

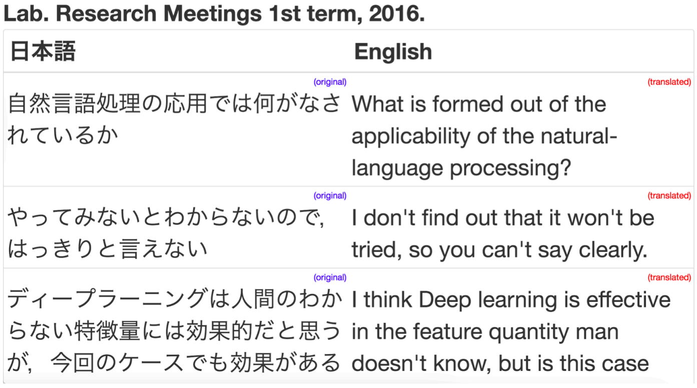
There is also a study at which the best answer is estimated by a st

Figure 3: Input screen

## 5 Take the system into real use

### 5.1 Practice 1: International Symposium

We applied our system in an international symposium. Two inputters were used (Inputter A and B) and their mother language was Japanese. To achieve adequate input speed, both were experienced in data input work and to maximize translation quality, we trained them in pre-editing and the use of *IPTalk* beforehand. For training, we had them perform, independently, three trial in 90 minutes. In each trial they listened to and pre-edited speech (in Japanese) and then checked the translation result. Next, the inputters



Lab. Research Meetings 1st term, 2016.

| 日本語 | English |
|---|---|
| (original)<br>自然言語処理の応用では何がなされているか | (translated)<br>What is formed out of the applicability of the natural-language processing? |
| (original)<br>やってみないとわからないので，はっきりと言えない | (translated)<br>I don't find out that it won't be tried, so you can't say clearly. |
| (original)<br>ディープラーニングは人間のわからない特徴量には効果的だと思うが，今回のケースでも効果がある | (translated)<br>I think Deep learning is effective in the feature quantity man doesn't know, but is this case |

Figure 4: Display screen

Figure 5: The look of conference hall

worked together in performing five trials in five hours. In the training sessions, inputter A entered 495 sentences while B entered 481. We provided multi-language support for Y's Men International 26th Asia Area Convention. This symposium drew 962 participants and was held over three days in Kyoto, Japan. Figure 5 shows a picture of the conference hall. We used our system to provide real-time multi-language translation for an 80 minute Japanese speech. The hall has three screens. The center screen shows the live video or slides as determined by the presenter, left screen displayed Japanese text and English translation and right screen displayed Japanese text and Chinese translation. 738 sentences were input, translated and displayed during the speech.

The system consisted of one server machine (MacBook), two machines to display the translations (Windows) and two machines to run *IPTalk* (used in the training sessions); all machines were notebook PCs.

### 5.2 Practice2: Presentation at laboratory

We also applied our system to a 30 minute presentation in our laboratory. The audience consisted of 20 people. There were two inputters and the two speakers used English. The system translated the speech into Vietnamese, Chinese, Dutch, Japanese and Korean. Two screens were used; the left one displayed the first three languages and the right one for the last two languages.

The system for this support consisted of one server (MacBook), two machines for inputters (MacBook) and two machines for displaying translation (Windows). We used one web screen of our system as the em Input Screen. Both inputters had no experience in captioning but we trained them for two hours (input

Table 2: Status of input of multilingual support

| Inputter | Exp.1(ja) | | Exp.2(en) | |
|---|---|---|---|---|
| Question | A | B | C | D |
| Number of input sentences | 290 | 448 | 94 | 76 |
| Ave. number of characters | 12.8 | 10.8 | | |
| Ave. numbers of words | | | 6.68 | 7.7 |
| Miss-inputs(%) | 4(1.38) | 1(0.22) | 4(4.26) | 2(2.63) |
| Redundant inputs(%) | 7(2.41) | 11(2.46) | 0(0) | 1(1.32) |
| Succession inputs(%) | 32(11.03) | 191(42.63) | 36(38.3) | 19(25) |

Table 3: The questionnaire of inputter(1:worst-5:best)

| Inputter | Exp.1(ja) | | Exp.2(en) | |
|---|---|---|---|---|
| Question | A | B | C | D |
| Was the input screen easy to look at? | | | 4 | 4 |
| Was the Back translation useful? | | | 4 | 4 |
| Was the Type monitor useful? | | | 5 | 5 |
| Was the whole input function easy to use? | | | 4 | 3 |
| Was the IPTalk easy to use? | 4 | 4 | | |
| Could you cooperate other inputter well? | 5 | 4 | 4 | 5 |
| Could you input whole speech? | 4 | 4 | 3 | 3 |

separately for 30 minutes and input cooperatively for one and half hours) beforehand. 170 sentences were inputted, translated and displayed during the presentation.

## 6 Results and discussions

### 6.1 Questionnaire

Table 2 shows the details of inputs in the two experiments. Inputters input in Japanese in Experiment 1 and English in Experiment 2. Average number of characters is the average number of characters per Japanese sentence and average number of words is the average number of words per English sentence. Miss-inputs indicates the number of sentences that have obvious input faults such as ”shrae”(should be ”share”). Redundant inputs is the number of sentences which are similar to previous input (i.e. one inputter input ”from Germany,” and the other inputter input ”From Germany and Denmark.” in succession). Successive inputs is the number of sentences consecutively entered by the same inputter.

We conducted questionnaires after the experiments. Table 3 shows the results of inputters in Exp.1 and Exp.2. The questions examined the functionality or usability of input functions, the cooperation with other inputter, and subjective evaluation of inputters about cover rate; a five-point scale was used. For input functions, we used *IPTalk* in Exp.1 and the *Input Screen* of our system in Exp.2, so questions for Exp.2 include each part of the *Input Screen.* Table 4 shows the impressions of the audience. We asked five questions (five-point response) and totaled the number of people who choose the same point. From Exp. 2, 16 of 20 people in the audience responded. Table 5 shows the translation quality of 50 sentences extracted randomly from Exp.1.

### 6.2 Input method

We assume inputters of our system would enter sentences by turns by cooperating with each other while the translation service would accept the pre-edited speech to keep translation quality as high as possible. Thus we attempted to train the inputters well beforehand. As Table 2 shows, however, sequential input was common. We considered that turn taking would be naturally selected but in practice the inputters

Table 4: The questionnaire of audience(Exp.2)(1:worst-5:best)

| Question | Number of people of each score | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Was the translation screen easy to look at? | 1 | 3 | 2 | 7 | 3 |
| Was the translation displayed in a timely manner? | 0 | 6 | 7 | 3 | 0 |
| Was the content of the translation easy to understand? | 1 | 9 | 5 | 1 | 0 |
| Was the content of the translation helpful to understand the presentation? | 0 | 5 | 5 | 5 | 1 |
| Will you want to use this system in the future? | 0 | 2 | 6 | 5 | 3 |

Table 5: The quality of translation of Exp.1

| sentence creation method | number of sentence | target language | average fluency | average adequacy |
|---|---|---|---|---|
| cooperated pre-editing input | 50 | Chinese | 3.58 | 3.82 |
| | | English | 3.94 | 3.88 |

sometimes input sentences sequentially.

The inputters ’ response showed that the systems were viewed very positively by all four inputters (inputter D experienced some difficulty because f his customized keyboard). So we conclude that the input function posed no obvious difficulties to the inputters.

### 6.3 Translation quality

The shorter training time is preferred because we assume that some of audience become inputter and cooperate each other and translation service, and realize multi-language support. But we need certain length of training because inputters of cooperated pre-editing input must get accustomed to pre-editing and predicting which inputter inputs which part of speech. For Exp.1 we train them six and half hours, and for Exp.2 three hours. Table 5 shows the certain translation quality enhancement, that average fluency improved from 1.84 to 3.76 and average adequacy from 2.89 to 3.76, but the evaluation by audience of Exp.2 was low. Though we could not simply compare the results of each experiments because the language of input and output is different and the speech is also different, at least we can see the result that by training inputters six hours we could get certain translation quality and we could assume that we don’t need training and trained skills as simultaneous translators have. But we still need to improve input functions or way to training to reduce training time and advance input efficiency.

## 7 Conclusion

In this paper, we tried to solve the following problems to realize low cost and high quality multi-language support.

- Translation quality improvement by the cooperation of human inputters and language services.
- Realizing a multi-language support system whose user interface supports inputter cooperation and language services.

The proposed method was shown to improve translation quality by keeping the original intent of the speech; this is done by combining real-time cooperative captioning and pre-editing for input to an MT service. We implemented the multi-language support system and put it into practice in actual international symposium and achieved fluency and adequacy scores of 3.58 and 3.82 for Japanese-Chinese translation, and 3.94 and 3.88 for Japanese-English translation. While the content of speech itself is different between Table 1 and Table 5, the speaker was same. It shows some positive effect of our method.

At the implementation of the system, we developed *Input Screen* that has log area, type monitor and back translation, and *Translation Screen*. By translating speech from Japanese to English and Chinese at Exp.1 and English to five other languages at Exp.2, we show our user interface has some usability by evaluation by users.

## Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (S) (24220002) of Japan Society for the Promotion of Science (JSPS). The authors wish to thank members of Y's men international Asia Area and president Okano for supporting this research.

## References

Pierrette Bouillon, Liliana Gaspar, Johanna Gerlach, Victoria Porro, Johann Roturier. 2014. *Pre-editing by Forum Users: a Case Study.* 3-10. Controlled Natural Language Simplifying Language Use.

Torsten Jachmann, Robert Grabowski, and Mayo Kudo. 2014. *Machine-Translating English Forum Posts to Japanese: On Pre-editing Rules as Part of Domain Adaptation.* 20th Annual Meeting. 808-811. Natural Language Processing.

Toru Ishida (Ed.). 2011. *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability.*, ISBN 978-3-642-21177-5. Springer.

Shigeaki Kurita, Sumihiro Kawano and Keiko Kondou. 2013. *The remote computer assisted speech-to-text interpreter system for reducing operational costs*, vol. 15, no. 8, SIG-ACI-10, 13-20. Human Interface Society.

Walter S. Lasecki, Christopher D. Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey P. Bigham. 2012. *Real-Time Captioning by Groups of Non-Experts* Proceedings of the 25th annual ACM symposium on User interface software and technology.

Linguistic Data Consortium. 2002. *Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations.*

Robert F. Lusch and Stephen L. Vargo. 2006. *The service dominant logic of marketing: Dialog, debate and directions.* Armonk, NY. M.E. Sharpe.

Shodai Matsuda, Xinhui Hu, and Yoshinori Shiga. 2013. *Multilingual speech-to-speech translation system: VoiceTra.*, Vol. 2. 14th International Conference on Mobile Data Management (MDM).

Rei Miyata, Anthony Hartley, Cécile Paris, Midori Tatsumi, and Kyo Kageura 2015. *Japanese Controlled Language Rules to Improve Machine Translatability of Municipal Documents.* 90-103. MT Summit XV.

Shigeki Miyoshi, Hayato Kuroki, Sumihiro Kawano, Mayumi Shirasawa, Yasushi Ishihara, Masayuki Kobayashi. 2008. *Support Technique for Real-Time Captionist to Use Speech Recognition Software*, International Conference on Computers for Handicapped Persons. Springer Berlin Heidelberg.

Y. Murakami, M. Tanaka, D. Lin and T. Ishida. 2012. *Service grid federation architecture for heterogeneous domains.* 539-546. IEEE International Conference on Services Computing,

Masahiro Tanaka, Yohei Murakami, Donghui Lin, and Toru Ishida. 2010. *Language Grid Toolbox: Open source multi-language community site.* 4th International Conference on Universal Communication Symposium (IUCS), IEEE.