# Building a Parallel Corpus for Monologues with Clause Alignment

**Hideki Kashioka, Takehiko Maruyama, Hideki Tanaka**

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City", Kyoto 619-0288, Japan
{hideki.kashioka,takehiko.maruyama,hideki.tanaka}@atr.co.jp

## Abstract

Many studies have been reported in the domain of speech-to-speech machine translation systems for travel conversation use. Therefore, a large number of travel domain corpora have become available in recent years. From a wider viewpoint, speech-to-speech systems are required for many purposes other than travel conversation. One of these is monologues (e.g., TV news, lectures, technical presentations). However, in monologues, sentences tend to be long and complicated, which often causes problems for parsing and translation. Therefore, we need a suitable translation unit, rather than the sentence. We propose the clause as a unit for translation. To develop a speech-to-speech machine translation system for monologues based on the clause as the translation unit, we need a monologue parallel corpus with clause alignment. In this paper, we describe how to build a Japanese-English monologue parallel corpus with clauses aligned, and discuss the features of this corpus.

## 1 Introduction

Recently, much research into speech-to-speech machine translation systems (SSMT systems) has been reported. In most of this research, attention is focused on dialogues (Takezawa, Sumita, Sugaya, Yamamoto and Yamamoto, 2002). This is because an SSMT system for handling travel arrangement dialogues is a desired application system. However, situations that require SSMT systems should handle not only dialogues (like travel arrangements) but also monologues (like lectures, broadcast news, and talks). Actually, many simultaneous interpretations can be found right in our own backyard. Therefore an SSMT system for monologues is an important issue. In addition, many corpora for monologues in Japanese have become available in recent years (Maekawa, Koiso, Furui and Isahara, 2000).

In considering an SSMT system for monologues, we look at the differences between dialogues and monologues. In a dialogue, each utterance unit is usually a short, simple sentence. A system for dialogues thus uses the sentence as the processing unit for translation. In a monologue, however, the speaker often produces long, complicated sentences, which are sometimes even ungrammatical (Takanashi, Maruyama, Uchimoto and Isahara, 2003). When a system handles monologues, the sentence is not a suitable processing unit, since long sentences often cause problems for parsing. The longer a sentence becomes, the more ambiguity is generated during parsing. Parsing errors naturally lead to errors in translation. Thus, it is necessary to define some shorter units for processing monologues in place of sentences. We propose "clauses" as basic units for processing monologues. In Japanese, clauses basically contain one verb phrase, thus are syntactically sufficient and semantically meaningful constituents. Therefore, they can be considered to be useful units for parsing and translation.

In this paper, we introduce a parallel corpus with "clause alignment" in monologues and discuss the features of the clause alignment of this parallel corpus. In section 2, we discuss how to detect the clause in Japanese. In section 3, we describe how to build a parallel corpus with clause alignment. In section 4, the features of the parallel corpus are shown in relation to clause alignment. Then we

discuss the advantages for using the clause as a translation unit in section 5.

## 2 Clause Boundary Annotation with Target Data in Japanese

In this section, we describe a method for annotating the clause boundary using the clause type in the source language.

### 2.1 Target data

In this paper, we use the corpus of a transcription of a TV commentary program called "asu-wo-yomu," which is broadcast by NHK (the Japan Broadcasting Corporation). Each program consists of a 10-minute presentation of a current event by a commentator. The parallel corpus that we built includes 250 programs. Each program contains about 60-70 sentences, and each sentence has about 30 words on average. Most of the sentences are complex or compound.

The average length of a sentence in a monologue is much longer than that in a dialogue. Therefore, translating a sentence that appears in a monologue is much more difficult than that for a dialogue. It is necessary to determine an effective, short unit for processing monologues.

A sentence consists of several constituents --- morphemes, phrases and clauses. A clause in Japanese can be defined as a meaningful constituent including one verb phrase. Thus we can consider it to be an appropriate processing unit for translation. Morphemes and phrases are too short for this purpose.

We can detect the positions of clause boundaries by parsing sentences. However, it is basically difficult to parse a long sentence accurately. We developed a program that detects clause boundaries and annotates the labels without parsing the sentences. Next, we describe how to detect the clause boundary automatically.

### 2.2 Method of clause segmentation

First, we will list a classification of clauses from the point of view of Japanese descriptive grammar. Then we will introduce our clause boundary annotation program, and examine its performance.

#### *2.2.1 Classification of clauses*

Generally, clauses can be classified into two types: a main clause and a subordinate clause. Main clauses are put at the end of a sentence. Subordinate clauses are put in the middle of a sentence and modify a main clause or other subordinate clauses. Japanese subordinate clauses can be roughly classified into four types according to their functions (Masuoka and Takubo, 1992).

The classification of subordinate clauses is shown below.

**[A. Supplement clause:]** Clauses working as arguments of verb phrases in a main clause with formal nouns (FM) and case marking particles.

健は　　**凛を　　見た ことを** 思い出した
Ken-TOP Lynn-ACC saw fm-ACC remembered
``Ken remembered that he saw Lynn."

**[B. Adverbial clause:]** Clauses modifying verb phrases in a main clause or a whole sentence.

健は　　**テレビを 見ながら** 夕食を　　とる
Ken-TOP TV-ACC watching supper-ACC have
``Ken has supper watching TV."

**[C. Adnominal clause:]** Clauses modifying the following nouns.

**健が 撮った** 写真
Ken-NOM took picture
``the picture that Ken took"

**[D. Compound clause:]** Clauses connected to a main clause on an equal footing.

**健は　　音楽が　　好きで** 凛は
Ken-TOP music-ACC likes Lynn-TOP
映画が　　好きだ
Movies-ACC likes
``Ken likes music, and Lynn likes movies."

As a whole, clauses can be classified into five types: a main clause, and four sub-types of subordinate clauses. Subordinate clauses can also be classified by their relational meanings between clauses.

Since Japanese is an SOV language, verb phrases or conjunctive particles are commonly placed at the end of clauses. Such a feature makes it possible to detect the boundaries between each clause rather precisely by considering part-of-speech (POS) tags, especially conjugated forms of verbs or conjunctive particles. Marking up all of the clause boundaries in

a sentence will be helpful for extracting clauses as basic processing units.

### *2.2.2 Clause boundary annotation program*

According to the grammatical classification above, we developed an annotation program, called the Clause Boundary Annotation Program (**CBAP**) that detects and labels every clause boundary in a sentence.

The substance of our program itself is a set of clause boundary annotation rules, described manually. Each clause boundary annotation rule consists of "boundary patterns" to refer to the Part-Of-Speech (POS) tags and find the clause boundaries, and "annotation labels" to represent the types of clause boundaries. The program requires a string of morphemes as an input. Thus, the input sentence must be analyzed morphologically in advance. We used the Japanese morphological analyzer "ChaSen (http://chasen.aist-nara.ac.jp/)" (Matumoto et al., 2001), and formed each morpheme by using the four tags of "Surface form_POS_Conjugation form_ Conjugation type." When a particular string of less than three morphemes is accepted as an input, CBAP compares it with the boundary patterns. If the string of the input matches some boundary pattern, a corresponding annotation label is inserted after the boundary. There are a total of 361 rules included in CBAP. A few examples of the rules are shown as follows:

1. s/(が_助詞-接続助詞__) /$1 \/並列節ガ\/ /g;
If GA_conjunctive-particle appeared in a text, insert the clause boundary marker labelled "Compound clause GA" following the conjunctive particle GA.

2. s/(*_動詞_*_連用形 たら_助動詞_特殊・タ_仮定形) /$1 \/条件節タラ\/ /g;
If concatenation of the infinitive form of any verb and an auxiliary verb TARA appeared in a text, insert the clause boundary marker labelled "Conditional TARA" following the concatenation of words.

3. s/(*_(動詞|助動詞)_*_基本形 (と|って)_助詞-格助詞-引用__) /$1 \/引用節\/ /g;
If concatenation of the basic form of any verb or an auxiliary verb and the case particle of TO or TTE appeared in a text, insert the clause boundary marker labelled "Quotational clause" following the concatenation of words.

## 2.3 Accuracy of clause boundary

We compared the result of annotations by CBAP with manual annotations to examine the performance of CBAP. We chose 500 sentences from each corpus and compared them with the result of manual annotation to calculate the precision and recall. The result shown in Table 1 verifies that CBAP can consistently detect the clause boundaries in a sentence and annotate them with very high accuracy.

Table 1 Precision and Recall

| Precision | Recall |
|---|---|
| 97.49% | 97.07% |

Nevertheless, there are a few boundaries that CBAP cannot detect from the local concatenation of morphemes. For example, it is impossible to detect the boundary of "noun-final clauses," irregular clauses that finish with a noun phrase. The "noun-final clause" is formally outside our definition of a clause. However, notionally it would be better to detect the boundary for the following process. "‖" in the following example marks the boundary of a noun-final clause.

[J:]コンチネンタル式が十ドル ‖ 英国式が十二ドルです。
[E:] The Continental breakfast is ten dollars, ‖ and an English breakfast is twelve dollars.

Since the boundary of the noun-final clause lies between two nouns ``十ドル" and ``英国式," there is no clue to detect the boundary from the surface concatenation of the morphemes. In order to handle these kinds of irregular clauses, we have to consider the syntactic structure of the sentence or the context in which it appears.

The average length of a sentence in a monologue is about 30 words. We examined the average length of each clause divided by CBAP. A detailed distribution of sentence/clause length is shown in Figure 1.

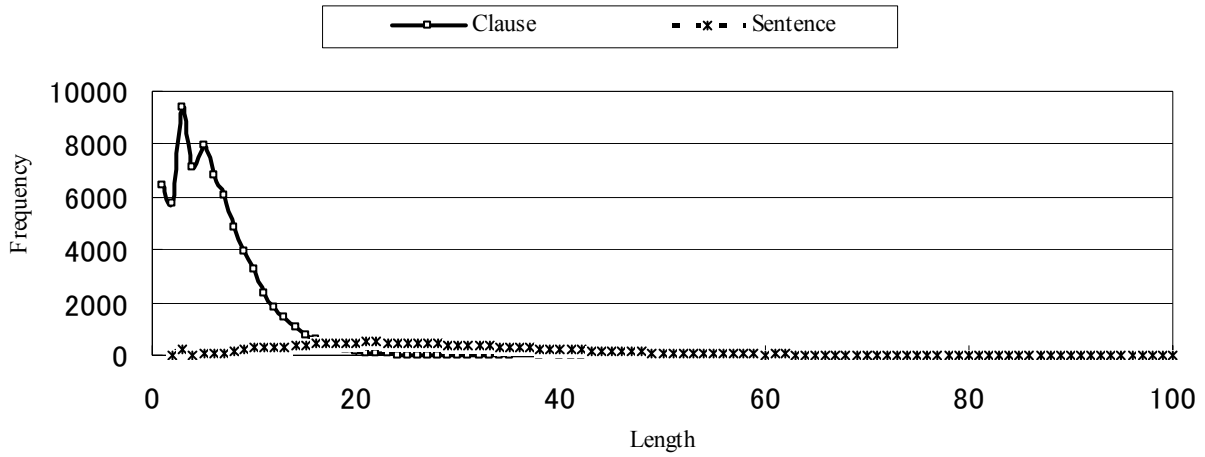


Figure 1 Distribution of sentence/clause length

## 3 Alignment of Translated Data

In this section, we show the process that was used to build a parallel corpus for monologues with "clause alignment." We constructed the parallel corpus using the following steps:

1. Prepare a transcription that has already been morphologically analyzed by ChaSen.

2. Find the clause boundaries by CBAP in the Japanese transcription, reformat the file with one clause in each line, and assign a line number.

3. Have people translate each sentence (paying attention to clause boundaries).

4. Divide the English translated sentence into pieces according to the Japanese clause boundaries by a person who is not a translator.

5. Annotate the corresponding line number for the English piece with the Japanese clause.

In step 2, we use the CBAP that is mentioned in the previous section. Therefore, some lines do not indicate the correct boundaries. Then, a human translator reads the text as shown in Figure 2.

These steps are taken with at least two persons: a translator and a segmentation annotator. Thus, the translator pays a bit of attention to clause boundaries, but not very serious attention.

```
==========================================
 1, 今晩は。/文末/
  (KON BAN WA .)
 2, 私たち人間/体言止/
  (WATASI TACHI NINNGEN)
 3, つまり/談話標識/
  (TUMARI)
 4, 人の遺伝子を解読するという/連体節トイウ/
  (HITO NO IDENSI WO KAIDOKU SURU TO YUU)
 5, 研究が民間企業も参加して/テ節/
  (KENKYUU GA MINNKANN KIGYOU
   MO SANKA SITE)
 6, 激しい競争の中で今進められています。/文末/
  (HAGESII KYOUSOU NO NAKA DE
IMA SUSUME RARETE IMASU.)
 7, 人遺伝子と解読というものが大きな利益に
   結び付く/従属文/
  (HITO IDENSHI TO KAIDOKU TOIUMONO GA
   OOKINA RIEKI NI MUSUBI TUKU)
 8, つまり/談話標識/
  (TUMARI)
 9, 宝の山であると/引用節/
  (TAKARA NO YAMA DE ARU TO)
 10, 考えられているから/理由節カラ/
  (KANGAE RARETE IRU KARA)
 11, です。/文末/
  (DESU.)
==========================================
```

Figure 2 Sample of the text that the translator read.

Human translators translate using sentences with the transcription containing annotated clause boundaries. The human translators were instructed with regard to the following points:

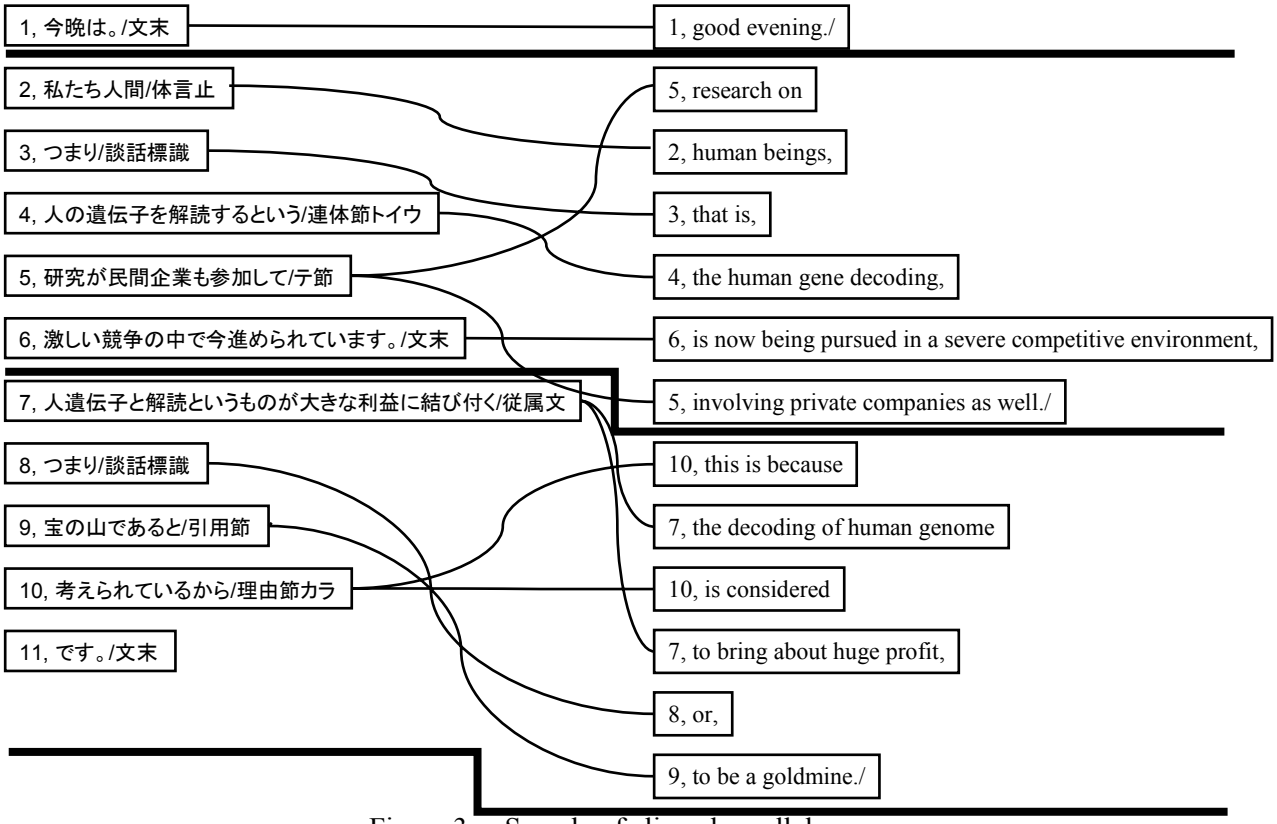


Figure 3 Sample of aligned parallel corpus

- Basically, one sentence should be translated into one sentence.
- The translated sentence should correspond to the Japanese clauses.
- However, the parts of clauses that do not have a very important meaning in the story can be ignored or paraphrased.
- The pronoun in Japanese should be translated with a pronoun or a referential expression rather than a concrete expression.
- A concrete expression in Japanese can be translated with a pronoun or a referential expression.
- The period that appears at the end of a sentence is expressed with "/".
- In the translated text, "/" is not used to indicate anything but the end of a sentence.

In step 4, where the segmentation processing person divides the English sentences into units that correspond to the Japanese clauses, if the corresponding Japanese clause boundary is strange because of a morphological analysis error and the translated English parts cannot be divided according to the Japanese boundary, the mark "@" is assigned following the corresponding line number by him/her, as in the following examples (Figure 4):

==========

***ChaSen + CBAP result:***

338, 否定できませんし/並列節シ/
(HITEI DEKI MASEN SHI)
339, かし普通の企業ですと/条件節ト/
(KASHI FUTSUU NO KIGYOU DESUTO)

***Correct:***

338, 否定できません
(HITEI DEKIMASEN)
339, しかし普通の企業ですと
(SHIKASHI FUTSUU NO KIGYOU DESUTO)

***Translated corpus:***

339,@ but in case of ordinary companies,

==========

Figure 4 Sample of "@" sign in the translation

The translator knows that the translated text will be aligned with the original Japanese clause. To the best of their ability, they should thus include the terms in the translated text. Therefore, as in the following case, the translator would translate the sample alignment corpus as shown in Figure 2.

## 4 Data Analysis

In this section, we describe the characteristics of the clause alignment corpus.

### 4.1 Length of sentence and clause/piece

The Japanese transcriptions of 250 programs as mentioned in section 2 included 15,313 sentences in which 70,989 clause boundaries were detected. The English translation was divided into 73,755 pieces from 15,275 translated sentences. The distribution of the length with translated sentences and pieces in English is shown in Figure 5. The relationship between the English sentences and the pieces is similar to that between the Japanese sentences and the clauses in Figure 1.

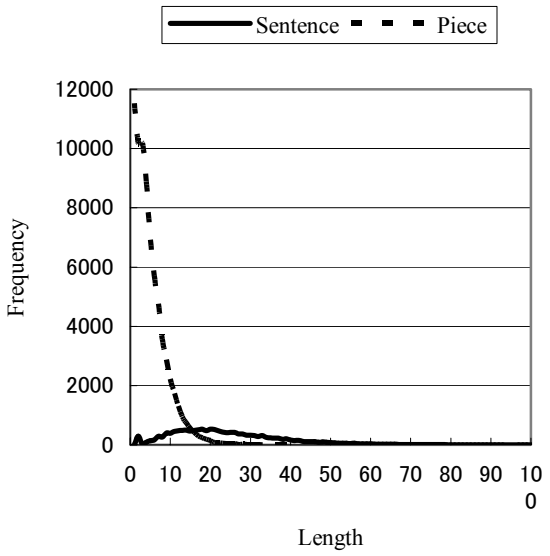


Figure 5 Distribution of lengths for translated sentences/pieces

Figure 6 shows the distribution of the lengths for Japanese clauses and English pieces. The distribution of the clause length is similar to that of the piece length. From the similarity between these two distributions, it is likely that one Japanese clause is almost always translated into one piece and each corresponding pair has almost the same information.

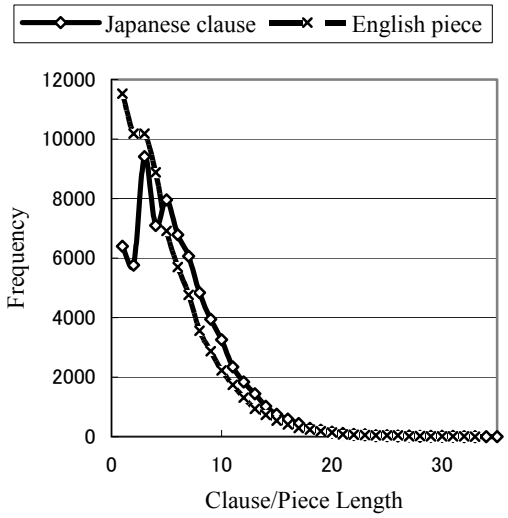


Figure 6 Distribution of lengths for Japanese clauses/English pieces

### 4.2 Clauses that have no corresponding piece

Some 6,280 Japanese clauses have no corresponding piece in the English translation. This means that 8.8% of the Japanese clauses have no corresponding piece. Table 2 shows the Japanese clauses that have no corresponding piece.

Table 2：Clauses that have no translation

| | | |
|---|---|---|
| 思います。<br>(omoi masu) | I think ... | 文末 |
| これは<br>(korewa) | This is ... | 主題ハ |
| わけです。<br>(wake desu) | The reason is ... | 文末 |
| ことです。<br>(koto desu) | It is ... | 文末 |
| です。<br>(desu) | It is ... | 文末 |
| こと<br>(koto) | | 体言止 |
| なります。<br>(nari masu) | It became ... | 文末 |

| | | |
|---|---|---|
| ものです。 (mono desu) | It is ... | 文末 |
| そして (sosite) | And ... | 談話標識 |
| また (mata) | Also ... | 談話標識 |

Almost half of these clauses (2,973 clauses) have a "文末" label. This result shows that about 20% of all sentence end clauses have no piece in the translation. Also, the content words of these clauses consist of only one verb or one verb and a formal noun (e.g., koto, mono, wake), so the clauses are short.

### 4.3 Construction of translated pairs

Each Japanese clause is almost always translated into one piece. A total of 7,872 clauses are translated into two pieces, 511 clauses into three pieces, 15 clauses into four pieces, and 1 clause into five pieces. The clauses that are translated into multiple pieces are almost all long clauses and we were able to find the clause boundaries in them.

### 4.4 Order of the piece sequence

From another point of view, the sequence of the piece is not in the same order in the Japanese clauses as in Figure 2. The line between the Japanese clause and the English piece indicates the relationship of the translation pair. If the line does not cross with other lines, the system can output the translated piece without waiting for other translated pieces around the Japanese clause translation. In other words, when the system outputs the translated piece in correct order, the system waits for the output of the clause whose line does not cross previous lines. In Figure 2, the first clause can translate with no waiting. The system waits for the output from the second clause to the sixth clause. Therefore, the clauses in Figure 2 are segmented into three parts for correctly ordered translation. The first part is 1 clause, the second part includes 5 clauses, and the third part includes 5 clauses (or 4 clauses because the last clause has no corresponding piece). The number of clauses like this in all of the data is shown in Figure 5. Figure 5 shows the distribution of the number of pieces that constructed the parts of crossing lines.

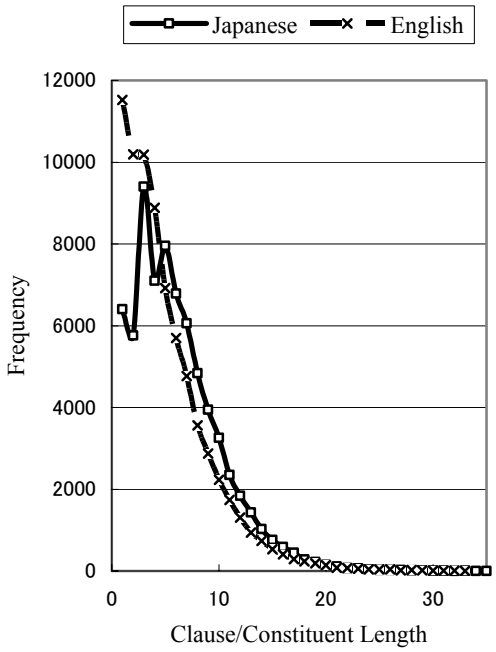


Figure 7 Number of clauses for ordered translation

Consequently, these two distributions are almost the same. It should be noted that more than half of them are pairs of one Japanese clause and one English piece and about 22% of them are pairs constructed with two Japanese clauses. Therefore, we are able to incrementally translate clause by clause.

## 5 Discussion and Conclusion

In Japanese, the clause boundary can be detected using the local information of word and POS tag sequences with very high accuracy. Therefore, if we had a system that could translate a clause into the appropriate forms corresponding to the role shown by the clause label in the sentence, we would be able to achieve simultaneous translation. This is indicated by the observation that half of the clauses are translated into one piece in the previous section. From this point of view, we need a translation system that can output the appropriate form as the role in the sentence.

It is likely that simple and short sentence translations with SSMT systems for dialogues or with MT systems for text, would achieve sufficient

accuracy. The Japanese clauses that can be defined as a meaningful constituent including one verb phrase would be considered as simple and short sentences. Therefore, we would be translating the clause as a sentence. A technique for paraphrasing from a simple sentence to a clause or phrase form that would express the role in the full long sentence is lacking. The parallel corpus that is described in this paper is useful for developing this technique because, using this parallel corpus, we can check the relationship between the role of the clause and the form of the constituent in the sentence. We are now trying to analyze these relationships.

We described three main points in this paper. The first is clause boundary detection and the clause label. The second is how to build a parallel corpus with clause alignment. And the third is the characteristics of the parallel corpus with regard to clause alignment. However, we have not completed the development of the SSMT system for monologues. We need to discuss the timing for output, and a translation mechanism for using the clause to construct the full sentence.

In the future, we will further study the relationship between the Japanese clause and the English piece.

## 6 Bibliographical References


Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world, *Proceedings of LREC 2002*, Canaria, pp. 147-152.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous Speech Corpus of Japanese. *Proceedings of LREC 2000*, Athens, pp. 947-952.

Katsuya Takanashi, Takehiko Maruyama, Kiyotaka Uchimoto, and Hitoshi Isahara. 2003. Identification of "Sentence" in Spontaneous Japanese – Detection and modification of clause boundaries –, *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. Tokyo, pp.183-186.

Takashi Masuoka, and Yukinori Takubo. 1992. *Kiso Nihongo Bunpou -Kaiteiban-*. Kuroshio-Shuppan, Tokyo, Japan.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. 2001. *Morphological Analysis System ChaSen version 2.2.8 Manual*. Nara Institute of Science and Technology.