# Construction of Chunk-Aligned Bilingual Lecture Corpus for Simultaneous Machine Translation

**Masaki Murata[†], Tomohiro Ohno[††], Shigeki Matsubara[†] and Yasuyoshi Inagaki[†††]**

[†]Graduate School of Information Science, Nagoya University,
[††]Graduate School of International Development, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
[†††]Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580, Japan
murata@el.itc.nagoya-u.ac.jp, {ohno,matubara}@nagoya-u.jp

**Abstract**

With the development of speech and language processing, speech translation systems have been developed. These studies target spoken dialogues, and employ consecutive interpretation, which uses a sentence as the translation unit. On the other hand, there exist a few researches about simultaneous interpreting, and recently, the language resources for promoting simultaneous interpreting research, such as the publication of an analytical large-scale corpus, has been prepared. For the future, it is necessary to make the corpora more practical toward realization of a simultaneous interpreting system. In this paper, we describe the construction of a bilingual corpus which can be used for simultaneous lecture interpreting research. Simultaneous lecture interpreting systems are required to recognize translation units in the middle of a sentence, and generate its translation at the proper timing. We constructed the bilingual lecture corpus by the following steps. First, we segmented sentences in the lecture data into semantically meaningful units for the simultaneous interpreting. And then, we assigned the translations to these units from the viewpoint of the simultaneous interpreting. In addition, we investigated the possibility of automatically detecting the simultaneous interpreting timing from our corpus.

## 1. Introduction

With the development of speech and language processing, speech translation systems have been developed (Frederking et al., 2002; Arranz et al., 2005; Gao et al., 2006; Nakamura et al., 2006). These studies target spoken dialogues, and employ consecutive interpretation, which uses a sentence as the translation unit. On the other hand, there exist a few researches about simultaneous interpreting (e.g. (Ryu et al., 2006)), and recently, the language resources for promoting simultaneous interpreting research, such as the publication of an analytical large-scale corpus (Matsubara et al., 2002), has been prepared. For the future, it is necessary to make the corpora more practical toward realization of a simultaneous interpreting system.

In this paper, we describe the construction of a bilingual corpus which can be used for simultaneous lecture interpreting research. Simultaneous lecture interpreting systems are required to recognize translation units in the middle of a sentence, and generate its translation at the proper timing. We constructed the bilingual lecture corpus by the following steps to develop a simultaneous lecture interpreting system.

1. We segmented sentences in the lecture data into semantically meaningful units for the simultaneous interpreting.

2. We assigned the translations to these units from the viewpoint of the simultaneous interpreting.

In addition, we investigated the possibility of automatically detecting the simultaneous interpreting timing from our corpus.

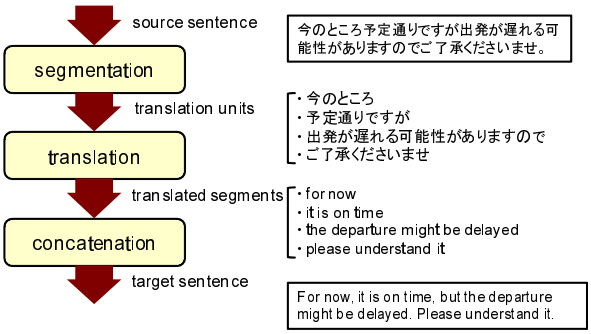


Figure 1: Configuration of a method for simultaneous lecture interpretation

This paper is organized as follows: Section 2 discusses a method of simultaneous lecture interpreting. Section 3 describes the design and the construction of a chunk-aligned bilingual lecture corpus. Section 4 reports corpus-based analyses on automatic detection of interpreting timing.

## 2. Simultaneous lecture interpretation

A simultaneous lecture interpreting system is required to output the translation result simultaneously with the input utterance. Since a sentence in a lecture tends to be long basically, it is necessary for the system to adopt shorter language units than sentences as translation units. We are suggesting a method of simultaneous lecture interpreting, as shown in Figure 1, consisting of the following three steps:

1. Segmentation of an input sentence into suitable translation units (**segmentation**).

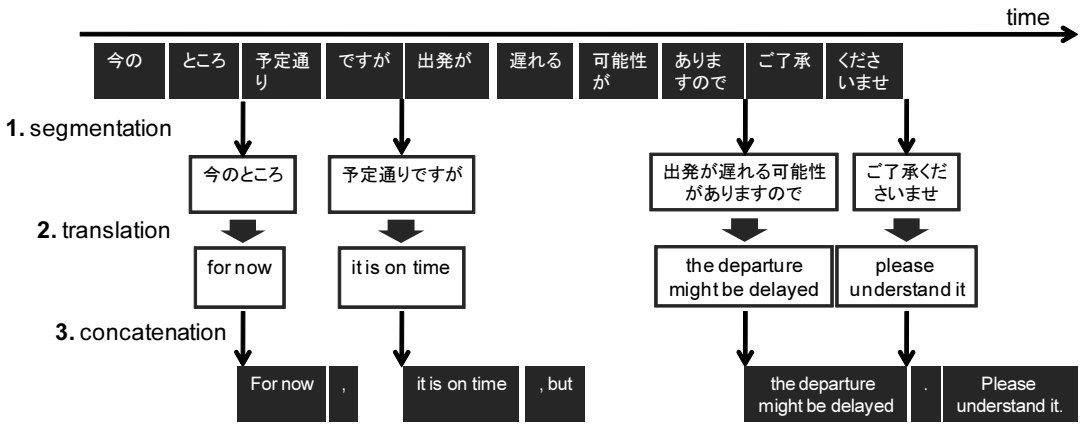


Figure 2: Flow of the simultaneous translation

2. Translation of each translation unit (**translation**).

3. Concatenation of these translated segments so that the translations form a natural English sentence (**concatenation**).

These steps work simultaneously with the speech input. Figure 2 shows an example of simultaneous translation process based on this approach. In this example, the following input Japanese sentence:

- 今のところ予定通りですが出発が遅れる可能性がありますのでご了承くださいませ.

is segmented into four units "今のところ," "予定通りですが," "出発が遅れる可能性がありますので," and "ご了承くださいませ," in the middle of the input. At the same time as segmentation, the system translates the units into English phrases "for now," "it is on time," "the departure might be delayed," and "please understand it," respectively. Then, the system concatenates each translation result and generates the following English sentence:

- For now, it is on time, but the departure might be delayed. Please understand it.

To realize such the process, it is necessary for the system to adopt shorter units than sentences as the translation units and detect such units correctly.

## 3. Construction of bilingual corpus

We constructed the bilingual lecture corpus for simultaneous lecture interpreting research. As the Japanese lecture data, we used Japanese spoken monologue data (1,935 sentences, 60,829 morphemes) in the simultaneous interpretation database (Matsubara et al., 2002). This data is annotated by hands with information on the morphological analysis, *bunsetsu*[1] boundary, dependency analysis, clause boundary (Ohno et al., 2009). Figure 3 shows the sample of the annotated spoken monologue data.

[1] *Bunsetsu* is a linguistic unit in Japanese that roughly corresponds to a basic phrase in English. A bunsetsu consists of one independent word and zero or more ancillary words.

| | |
|---|---|
| sentences | 1,935 |
| morphemes | 60,829 |
| bunsetsus | 23,598 |
| clauses | 9,664 |
| chunks | 8,644 |
| chunks per sentence | 4.47 |
| bunsetsus per chunk | 2.73 |

Table 1: Size of segmented Japanese data

In addition, this database includes the speech of interpretations by professional interpreters and their transcribed texts. However, such the interpretations are not always suitable as the data for current machine translation technologies because simultaneous interpretations under real environment may include loose translations of original sentences. Therefore, we assigned the renewed translations to this data.

### 3.1. Segmentation of the lecture text

We have segmented lecture texts into several shorter units than sentences by hands. In this paper, we call this unit a **chunk**. We set the following concepts as the chunk:

- Not so long: If the length of chunks gets long, the simultaneity is decreased because it takes much time to start the translation process. Also, it is desired that the length of chunks is uniform so that the delay of translation is kept constant.

- Semantically meaningful: It is desired that a chunk is semantically meaningful because the translation needs to be generated for each chunk.

We defined the maximum length of a chunk as 4.3 sec by considering the delay in the actual interpretations by the professional interpreters (Ono et al., 2008), and we segmented sentences into chunks according to this restriction. Table 1 shows the size of the segmented Japanese spoken monologue data. As a result, 1,935 sentences in the database were segmented into 8,644 chunks (4.47 chunks per a sentence). In addition, we have already confirmed that the chunk boundaries can be detected automatically with about 80% of precision (Ohno et al., 2009).

{PAU}{0132-06:57:568-07:01:276}{0132-06:57:568-07:01:276} utterance_unit_segmentation none none
* 0 4D
それから *sorekara* それから conjunction none none
C-BOU /discourse marker/
* 1 2D
千九百六十四 *senkyuhyakurokujuyon* 千九百六十四 noun 数 none none
年 *nen* 年 noun 接尾-助数詞 none none
に *ni* に particle 格助詞-一般 none none
* 2 4D
なり *nari* なる verb 自立 五段・ラ行 連用形
ます *masu* ます auxiliary-verb 特殊・マス 基本形
と *to* と particle 接続助詞 none none
C-BOU /condition clause -*to*/
{PAU}{0133-07:01:724-07:04:796}{0133-07:01:724-07:04:796} utterance_unit_segmentation none none
[F-え] [F-*e*] [F-え] filler none none
* 3 4D
オーイーシーディー *OECD* オーイーシーディー noun 固有名詞-組織 none none
に *ni* に particle 格助詞-一般 none none
* 4 -1O
加盟 *kame* 加盟 noun サ変接続 none none
し *shi* する verb 自立 サ変・スル 連用形
て *te* て particle 接続助詞 none none
おり *ori* おる verb 非自立 五段・ラ行 連用形
ます *masu* ます auxiliary-verb 特殊・マス 基本形
C-BOU /end of a sentence/

Figure 3: Example of the annotated spoken monologue data

| | |
|---|---|
| chunks | 5,662 |
| words | 50,054 |
| words per chunk | 8.84 |

Table 2: Size of translation data

### 3.2. Translation of chunks

We constructed the bilingual corpus by assigning the translations to each chunk. The translations were provided by professional translators who are familiar with interpretations. Though it is ideal to assign one translation to each chunk, every chunk can not be always translated by itself. The translators provided a translation to each chunk basically, but if a chunk was not able to be translated by itself, the translators translated such chunk together with chunks following it.

Figure 4 shows an example of the bilingual corpus. In this example, a chunk "それから千九百五十六年には" was translated into "Then, in 1956" by itself. On the other hand, for example, a chunk "より強くなったということが" was not translated by itself. So, this chunk was translated into "can be said, I think, to have become stronger." together with a chunk "いえると思います" following it.

Table 2 shows the size of the translation data.

## 4. Analysis of interpreting timing

We tried to assign one translation to one chunk at the construction of our corpus. However, there existed chunks which were not able to be translated by itself. So, it is not always appropriate that the system adopts a chunk as a translation unit. We investigated when the translation of a certain chunk had been generated. Concretely, we measured the number of chunks that had been observed by the time the translations were generated. Figure 5 shows the result. There exist 5,662 chunks which were able to be translated when these were observed, and its percentage of total is 65.50%. Also, 85.67% of all chunks was able to be translated in case that the next chunk was observed.

To identify translation units based on our corpus, it is necessary to decide whether to generate the translation whenever a chunk boundary is detected. We analyzed the timing with which the translation was generated at chunk boundaries. In this paper, we call this timing **simultaneous interpreting timing**. We focused on the pause, clause boundary and dependency relation as the available information in the automatic analysis of simultaneous interpreting timing. Here, 65.50% (5,662/8,644) of all chunk boundaries were simultaneous interpreting timing in our corpus. This is the standard ratio of simultaneous interpreting timing on chunk boundaries.

### 4.1. Pauses and interpreting timing

Pauses can be detected automatically when these were inserted. Since the pauses could correspond to syntactic boundaries. Therefore, pauses might be useful for detecting simultaneous interpreting timing. Table 3 shows the relation between pauses and simultaneous interpreting timing. The ratio that a chunk boundary having a pause was a simultaneous interpreting timing was 75.65% (4,526/5,983), and this ratio was higher than that of chunk boundaries (65.50%). This indicates that pauses are useful to detect

| | |
|---|---|
| それから千九百五十六年には | Then, in 1956 |
| ソ連との共同宣言によりまして | under the Japan-Soviet Union joint declaration |
| 日ソ間の国交が回復しております | diplomatic relations between Japan and the Soviet were normalized. |
| | |
| 平和条約の締結は | The conclusion of the Peace Treaty |
| その後に残されておりますが | was later postponed, but |
| 一応 | provisionally, |
| 戦争状態は終わったわけでございます | the state of the war ended. |
| | |
| それから千九百六十年には | Then, in 1960 |
| 日米の新しい安保条約が締結されまして | the Japan-U.S. Security Treaty, the new one, was concluded. |
| 安保条約の上でわが国の発言権が | Our nation's right to speak during the Security Treaty |
| より強くなったということが | |
| いえると思います | can be said, I think, to have become stronger. |
| | |
| それから千九百六十四年になりますと | And in 1964 |
| オーイーシーディーに加盟しております | Japan joined the OECD. |
| | |
| これは純粋に戦後処理というよりは | It wasn't purely the disposal of the postwar period, |
| その次の新しい飛躍の時期への | rather |
| 助走と申しますか | an approach run for the next step up, or |
| 準備の時期だったと思います | I suppose, a preparation period. |
| | |
| それから千九百六十五年には | In 1965 |
| 韓国との国交が正常化しております | diplomatic relations between Japan and Korea were normalized. |
| | |
| そして千九百六十八年には | In 1968 |
| それまでアメリカが占領しておりました | |
| 小笠原が返還され | Ogasawara, having been occupied by the U.S., was returned. |
| 千九百七十二年には沖縄の返還が行われ | In 1972, Okinawa was returned from the U.S. to Japan. |
| 先程申し上げましたように | As I mentioned before, |
| 中国との国交正常化もできた | the diplomatic relations between Japan and China were also normalized. |
| そういう時期でございます | That was exemplary of such a period. |

Figure 4: Example of the bilingual corpus

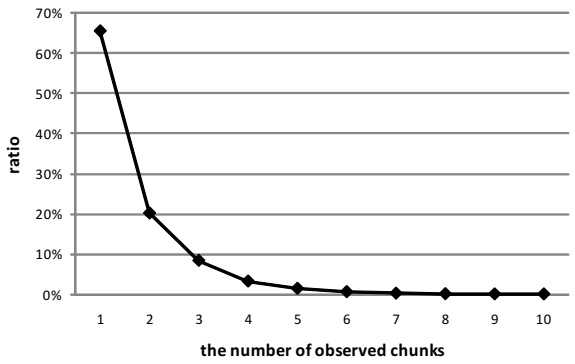


Figure 5: Relation between the number of observed chunks and generation of translations

| pause | translated | not translated | total |
|---|---|---|---|
| exists | 4,526 | 1,457 | 5,983 |
| not exist | 1,136 | 1,525 | 2,661 |

Table 3: Relation between pauses and interpreting timing

simultaneous interpreting timing.

### 4.2. Clause boundaries and interpreting timing

A clause is one of semantically meaningful language units including one verb phrase and corresponds to a simple sentence. The variation in the length of clauses is smaller (for example, than that of sentences), and clause boundaries can be detected using the local morphological information with high accuracy (Kashioka et al., 2003). Therefore, the clause boundaries may be useful for detecting simultaneous interpreting timing. Table 4 shows the relation between clause boundaries and simultaneous interpreting timing. Among chunk boundaries which were also clause boundaries, 77.61% of them were the simultaneous interpreting timing. Thus, we confirmed the usefulness of clause boundaries for detecting simultaneous interpreting timing.

However, there exist several types of clause boundary, and the role of each clause on a sentence is different by the types of clause boundaries. We investigated the ratio that a chunk boundary which was also a clause boundary was a simultaneous interpreting timing. Table 5 shows the top 10 clause boundary types about the occurrence frequency and their

| clause boundary | translated | not translated | total |
|---|---|---|---|
| exists | 4,668 | 1,347 | 6,015 |
| not exist | 994 | 1,635 | 2,629 |

Table 4: Relation between clause boundaries and interpreting timing

| type of clause boundary | ratio of translation (%) | |
|---|---|---|
| end of a sentence | 98.55 | (1,907/1,935) |
| topicalized element *-wa* | 70.35 | (503/715) |
| compound clause *-te* | 69.46 | (489/704) |
| supplement clause | 39.78 | (107/269) |
| continuous clause | 72.89 | (164/225) |
| adnominal clause | 31.69 | (58/183) |
| compound clause *-keredomo* | 94.19 | (227/241) |
| compound clause *-ga* | 94.21 | (228/242) |
| condition clause *-to* | 86.39 | (146/169) |
| quotational clause | 42.62 | (52/122) |

Table 5: Relation between clause boundary types and interpreting timing

ratio. There existed clause boundaries which were simultaneous interpreting timing with the ratio over 85%, such as "compound clause *-keredomo*," and "condition clause *-to*," besides "end of a sentence." On the other hand, in case of "supplement clause" and "adnominal clause," the ratio that such clause boundaries which simultaneous interpreting timing was less than 40%. This means that the likelihood that the chunk boundaries were simultaneous interpreting timing is different according to the types of the clause boundary.

### 4.3. Dependency structure and interpreting timing

A dependency relation is a modification relation in which a modifier bunsetsu depends on a modified bunsetsu. We focused on the dependency relation in which a bunsetsu depends on the next bunsetsu. In case that a busnetsu depends on the next bunsetsu, the chunk boundaries existing between them may be hard to be a simultaneous interpreting timing because the sequence of such the bunsetsus forms a semantically meaningful unit. Table 6 shows the relation between dependency structure and simultaneous interpreting timing. In case that a bunsetsu did not depend on the next bunsetsu, the ratio that the chunk boundaries between them were simultaneous interpreting timing was 72.93% (5,235/7,178). This indicates that the dependency structure could be used for detecting the simultaneous interpreting timing.

## 5. Conclusion

This paper has described the construction of a simultaneous lecture interpreting corpus. We have constructed the corpus by assigning translations for simultaneous interpreting to chunks, which are made by segmenting the Japanese lecture data. Among all chunks, about 65% of them were translated when they occurred. Therefore, we confirmed that this data is useful for developing simultaneous lecture interpreting systems.

| modified bunsetsu | translated | not translated | total |
|---|---|---|---|
| next bunsetsu | 427 | 1,039 | 1,466 |
| other bunsetsu | 5,235 | 1,943 | 7,178 |

Table 6: Relation between dependency structure and interpreting timing

In the future, we will study on techniques for deciding simultaneous interpreting timing and concatenating the translation result by using our corpus.

## 6. Acknowledgments

This study was partially supported by the Grant-in-Aid for Scientific Research (B) (No. 20300058) of JSPS and by the Continuation Grants for Young Researchers of The Asahi Glass Foundation.

## 7. References

Victoria Arranz, Elisabet Comelles, and David Farwell. 2005. The FAME speech-to-speech translation system for Catalan, English and Spanish. In *Proceedings of the 10th Machine Translation Summit*, pages 195–202.

Robert E. Frederking, Alan W Black, Ralf D. Brown, John Moody, and Eric Steinbrecher. 2002. Field testing the tongues speech-to-speech machine translation system. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 160–164.

Yuqing Gao, Bowen Zhou, Ruhi Sarikaya, Mohamed Afify, Hong-Kwang Kuo, Wei zhong Zhu, Yonggang Deng, Charles Prosser, Wei Zhang, and Laurent Besacier. 2006. IBM Mastor System: Multilingual automatic speech-to-speech translator. In *Proceedings of the 1st International Workshop on Medical Speech Translation*, pages 57–60.

Hideki Kashioka, Takehiko Maruyama, and Hideki Tanaka. 2003. Building a parallel corpus for monologue with clause alignment. In *Proceedings of the 9th Machine Translation Summit*, pages 216–223.

Shigeki Matsubara, Akira Takagi, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2002. Bilingual spoken monologue corpus for simultaneous machine interpretation research. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 153–159.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Genichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, Jin-Song Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. 2006. The ATR multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376.

Tomohiro Ohno, Masaki Murata, and Shigeki Matsubara. 2009. Linefeed insertion into Japanese spoken monologue for captioning. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*, pages 531–539.

Takahiro Ono, Hitomi Tohyama, and Shigeki Matsubara. 2008. Construction and analysis of word-level time-aligned simultaneous interpretation corpus. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 3383–3387.

Koichiro Ryu, Shigeki Matsubara, and Yasuyoshi Inagaki. 2006. Simultaneous English-Japanese spoken language translation based on incremental dependency parsing and transfer. In *Proceedings of the 21th International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 683–690.