講演の同時翻訳のための対訳データの作成と分析

Construction of Simultaneous Lecture Interpreting Corpus and Its Analysis

村田 匡輝 [1]　大野 誠寛 [1]　松原 茂樹 [1]　稲垣 康善 [2]

Masaki Murata[1]　Tomohiro Ohno[1]　Shigeki Matsubara[1]　Yasuyoshi Inagaki[2]

1　はじめに

音声・言語処理技術の進展を背景に、音声翻訳技術の開発が進んでいる。また、同時通訳に関する研究もいくつか行われ [1, 2]、最近では、大規模データの公開 [3] など、同時通訳研究を推進するための環境も整いつつある。今後は、同時通訳機の実現を目指し、これらのデータをより活用できように整備していくことが望まれる。

本稿では、講演の同時翻訳に利用することを目的とした対訳コーパスの構築について述べる。このコーパスは、既存の講演データに対して、同時翻訳を考慮した対訳文を新たに作成することにより実現する。同時翻訳では、講演の途中の段階で、意味的なまとまりを認識しながら、それに対する訳出を適切なタイミングで実行する必要がある。本研究では、構築したコーパスを用いて、訳出タイミングの検出可能性について検討する。

2　講演の同時翻訳方式

講演の翻訳では、入力音声と同時的にその対訳を出力する必要がある。一般に講演文は長くなる傾向にあるため、同時性の高い翻訳を実現するために、文よりも小さな単位で翻訳処理を実行することが求められる。このため、入力される音声言語文を適切な単位に分け（**分割**）、その単位に対応する訳文を生成し（**翻訳**）、訳出された表現が自然に繋がるように整形する（**連結**）という方法が考えられる。

このようなアプローチに基づく翻訳処理プロセスの例を図 1 に示す。これは、

- 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。

という入力文が、「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」の４つの単位に分割され、それぞれの対訳である “for now”, “it is on time”, “the departure might be delayed”, “please understand it” を連結することにより、最終的に

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

を訳出することを表している。このような処理を実現するために、同時翻訳のための処理単位を、文より短い単位で定め、それを正しく検出する必要がある。

3　対訳コーパスの作成

講演データを用いて同時翻訳のための対訳コーパスを作成した。データとして、名古屋大学同時通訳データベースの独話データ [3] を用いた（1,935 文、60,829 形態素）。このデータには、形態素、文節、節の各境界、及び、係り受け構造が与えられている [4]。

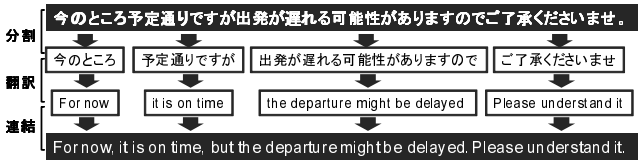

図 1: 同時翻訳の処理プロセス

17	
千九百四十五年に戦争が終わりまして	World War Two ended in 1945 and
それから今日までの五十年間を	
便宜的に分けますと	
私の考えでは	I think that for the sake of discussion, the past fifty years
三つ位に分けられるのではないかという	can be divided into three periods of time.
感じが致します	
18	
第一の期間は	The first period of time is
千九百四十五年から戦後処理	devoted to the elimination of postwar matters after 1945.
戦争によって引き起こされた	The war caused
いろんな問題を処理することに	the elimination of various issues and
努力の中心が払われた	we focused our efforts on that.
そういう時期であったと思います	It was necessary for such a period, I suppose.

図 2: 対訳コーパスのサンプル

なお、このデータベースには、プロの同時通訳者による通訳音声とその文字化データが含まれている。しかし、実環境下で生成された通訳は原文に忠実でない場合があり、現状の翻訳技術に用いるデータとしては必ずしも適さないと考え、新たに訳文を作成した。

3.1　講演テキストの分割

講演テキストを、同時翻訳のための処理単位（以下、**チャンク**）に分割する作業を人手で実施した。チャンクとしては、

- **長すぎない**：　チャンクが長くなると訳出タイミングが遅れ、同時性が損なわれる。また、遅れの程度をある程度一定に保つために、チャンクの長さは均一であることが望ましい。
- **意味的にまとまっている**：　各々のチャンクに対してその対訳を生成する必要があるため、意味的にまとまっていることが望ましい。

という制約を課した。

本研究では、プロの同時通訳者における訳出遅延の様相を考慮し [5]、チャンクの長さを 4.3 秒以内とし、その制限のもとで意味的にまとまる単位に分割した。その結果、8,644 チャンク（1 文あたり平均 4.47 チャンク）に分割された。なお、このチャンク境界は、約８割の精度で自動検出できることを確認している [4]。

3.2　対訳文の作成

各チャンクに対して、対訳を付与することによりコーパスを構築した。対訳作成は、通訳業務に精通するプロの翻訳者により実施した。ただし、必ずしもあらゆるチャンクに対して、それが出現した時点で対訳を生成できるわけではない。このため、コーパスの作成においては、対訳表現を、それが生成されうる時点で出現したチャンクに対して付与することとした。サンプルを図 2 に示す。

[1]名古屋大学

[2]豊橋技術科学大学

[3]http://slp.el.itc.nagoya-u.ac.jp/sidb/

表 1: ポーズと訳出タイミングとの関係

	訳出可	訳出不可	合計
ポーズ有	4,452	1,544	5,996
ポーズ無	1,210	1,438	2,648

表 2: 節境界と訳出タイミングとの関係

	訳出可	訳出不可	合計
節境界有	4,597	1,434	6,031
節境界無	1,065	1,548	2,613

4 訳出タイミングの分析

対訳コーパスでは、チャンクごとに対訳表現を付与することを試みたが、チャンクが出現した時点で訳出できるものもあれば、できないものもある。コーパスを調査したところ、出現した時点で対訳を出力できるものは、全体の 65.50%に相当する 5,662 チャンクであった。

本研究では、訳出可能性の自動判別の実現を目的に、実際に利用できる情報として「ポーズ」「節境界」「係り受け構造」に着目し、分析を与えた。

4.1 ポーズと訳出タイミング

ポーズは、それが挿入された時点で機械的に検出することができる。ポーズの挿入位置は、文法的なまとまりと大いに関連するため、訳出タイミングの検出に利用できる可能性がある。

表 1 に、ポーズの存在と訳出可能性との関係を示す。ポーズが存在するチャンク境界が訳出タイミングとなる割合は 74.25%(4,452/5,996) であり、全体の割合である 65.50%よりも高く、これは、訳出タイミングの検出において、チャンク情報が有用であることを意味している。

4.2 節境界と訳出タイミングの関係

節は、述語を中心としたまとまりであり、単文に相当する文法的単位である。長さの分布のばらつきは（例えば「文」と比べて）小さく、また、その境界は、局所的情報のみを用いて高い精度で検出できることから、訳出タイミングの判別に利用できる可能性がある [6]。

表 2 に、節境界の存在と訳出タイミングとの関係を示す。チャンク境界のうち、節境界のある位置で訳出できるものは 76.22%(4,597/6,031) であり、節境界情報の有用性が確認された。

なお、単に「節境界」といってもいくつかの種類があり、その種類によって「節」の言語的役割は異なる。そこで、節境界の種類（上位 10 種類）ごとの訳出タイミングとなる割合を調査した。結果を表 3 に示す。文末以外にも、並列節ケレドモ、並列節ガ、条件節トなど、高い割合で訳出可能となる節境界が存在することがわかる。その一方で、補足節や連体節など、訳出タイミングとなりにくい節境界が存在することも明らかになった。

4.3 係り受けと訳出タイミング

文節間の係り受け関係は大きく、隣接文節間上のものとそうでないものが存在する。係り受け関係にある隣接文節は、両文節で意味的なまとまりを形成する場合が多く、そのような文節間にあるチャンク境界は訳出タイミングになりにくい可能性がある。

表 4 に、係り受け関係の種類と訳出可能性との関係を示す。隣接文節間に係り受け関係がない場合の 71.66%(5,145/7,192) で訳出可能となっており、係り受け構造がタイミング検出に利用できることが示された。

表 3: 節境界の種類と訳出タイミングとの関係

節境界	割合 (%)
文末	96.39 (1,839/1,939)
主題ハ	70.69 (509/720)
テ節	68.41 (483/706)
補足節	43.66 (117/268)
連用節	72.89 (164/225)
連体節	33.89 (61/180)
並列節ケレドモ	84.02 (205/244)
並列節ガ	91.77 (223/243)
条件節ト	84.88 (146/172)
引用節	43.80 (53/121)

表 4: 係り受けと訳出タイミングとの関係

隣接文節間	訳出可	訳出不可	合計
係り受け関係有	508	944	1,452
係り受け関係無	5,154	2,038	7,192

5 まとめ

本稿では、講演の同時翻訳研究に利用するための対訳コーパスについて述べた。コーパスは、意味的単位に分割された日本語講演データに、同時翻訳に適した対訳表現を与えることにより作成した。全ての意味的単位のうち、それが出現した時点で訳出を完了できるものは約 65%であり、同時性の高い翻訳処理を実現する上で有用なデータであることを確認した。

今後は、このデータを用いて、講演データの意味的単位への分割、訳出タイミングの決定、及び、対訳表現を結合する訳出テクニックの獲得、について検討を進める予定である。

謝辞 本研究は、一部、科学研究費補助金基盤研究 (B)「入力文の分割・翻訳・連結に基づく同時通訳システム」(No. 20300058) により実施されたものである。

参考文献

[1] 松原：同時通訳の工学と科学 ―次世代自動通訳技術の実現に向けて，情報処理, Vol. 49, No. 6, pp. 617–623 (2008).

[2] 笠, 松原, 稲垣：英日同時翻訳のための依存構造に基づく訳文生成手法, 信学論, Vol. J92-D, No. 6, pp. 921–933 (2009).

[3] Matsubara, S., Takagi, A., Kawaguchi, N., Inagaki, Y.: Bilingual Spoken Language Corpus for Simultaneous Machine Interpretation Research, *Proceedings of LREC-2002*, I, pp. 153–159 (2002).

[4] 村田, 大野, 松原：読みやすい字幕生成のための講演テキストへの改行挿入，信学論, Vol. J92-D, No. 9 (2009).

[5] 小野, 遠山, 松原：大規模音声コーパスを用いた日英・英日同時通訳における訳出遅延の比較分析, 通訳研究, No. 7, pp. 51–64 (2007).

[6] Kashioka, H., Maruyama, T., Tanaka, H.: Building a Parallel Corpus for Monologue with Clause Alignment, *Proceedings of MT Summit IX*, pp. 216–223 (2003).