日英同時翻訳のための日本語対話文の分割

笠 浩一朗† 松原 茂樹‡ 稲垣 康善§

†名古屋大学大学院情報科学研究科 ‡名古屋大学情報連携基盤センター

§愛知県立大学情報科学部

ryu@el.itc.nagoya-u.ac.jp

1 はじめに

近年の音声翻訳技術の進展にともない，対話を対象とした音声翻訳システムの開発が盛んに行われている [1, 2, 4, 6]．これらのシステムの多くは，ターンや文を処理の基本単位として用いているため，対話の効率や円滑さが低下するという問題がある．それに対して，同時通訳者のように話者の発話に追従して訳出することができれば，対話の効率や円滑さが向上することが期待できる．

計算機で同時翻訳を実現するには，原言語の発話に対して，文よりも短い単語列を独立的に訳出可能な単位（以下，**同時翻訳単位**）として認識し，それを即座に訳出する必要がある。これまでに，独話の日英翻訳における同時翻訳単位として,「節」に着目した研究が報告されている [3]．しかし，対話を対象とした研究は行われていない．

そこで本論文では，日英同時通訳コーパスの日本語対話文に，同時翻訳単位境界を人手で付与し，分析する．分析では，同時翻訳単位と節単位または発話単位との関連性に着目した．また，節境界の種類と同時翻訳単位境界との関係についても調査した．分析結果に基づいて同時翻訳単位境界の判定ルールを作成し，日本語対話文を同時翻訳単位に分割する実験を行った．

2 同時翻訳単位データの作成

2.1 同時翻訳単位

同時翻訳単位とは，原言語において独立かつ即座に訳出可能な単位であるとする．例えば，日本語対話文

(J2.1) 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.

の英語訳を

(E2.1) For now, it is on time, but the departure might be delayed. Please understand it.

とする．このとき，構成要素間の対訳対応とその出現順序を考慮して (J2.1) を同時翻訳単位に分割すると，図 1 のような単位に分割できる．(J2.1) の日本語文を図 1 の同時翻訳単位ごとに翻訳し，出力すれば，同時通訳者のような訳出を実現できる．

同時翻訳単位	対応する英語訳
今のところ	For now
予定通りですが	it is on time
出発が遅れる可能性がありますので	but the departure might be delayed.
ご了承くださいませ。	please understand it.

図 1: (J2.1) の同時翻訳単位の分割例

表 1: 分析データの規模

対話数	7
文数	329
同時翻訳単位境界数 (文境界を除く)	164
発話単位境界数	167
節境界数 (文末除く)	189
形態素数	2677

2.2 対訳データベースとその利用

名古屋大学同時通訳データベース [7] に収録されている対話データを利用して，同時翻訳単位データを作成した．データベースには，英日，日英の二人の同時通訳者を介した英日間対話の音声データとその書き起こしデータが収録されている．ただし，同時通訳には，訳出における厳しい時間的な制約があり，その通訳音声には意訳や省略が多く含まれているため，訳の独立性が確保されず，同時翻訳単位データとしては適さない．そのため，新たに逐語訳データを作成し，それを利用した．逐語訳データは，次のような基準に基づいて作成した．

- 原文の内容を聞き手が理解可能な訳文である (**訳質**)
- 細かい単位で漸進的に訳出するために，原文の語順に準じた訳文である (**漸進性**)
- 文脈への依存を避けるために，意訳や省略を含まない訳文である (**文脈依存度**)

日本語対話文を同時翻訳単位に分割したデータの例を図 2(a) に示す．また，図 2(a) に対応するように逐語訳を分割したものを図 2(b) に示す．

3 同時翻訳単位境界の分析

同時翻訳単位が入力されるごとに訳出するシステムを実現するために，同時翻訳単位を漸進的に認識する必

1-1 お店は 1-2 道路沿いではないんですけれども 1-3 林ビルの二階にあります。
2-1 林ビルはすぐ見つけていただけると思います。
3-1 テレビ塔という大きなタワーのすぐ横ですので。
4-1 多分 4-2 日本語と思いますので 4-3 今からお書きします。
5-1 そうですね。
6-1 せっかくお越しいただいてるので 6-2名古屋城を見られたらと思いますね。

(a) 同時翻訳単位に分割された日本語対話文

1-1 The restaurant 1-2 is not on the street but 1-3 It's on the second floor of Hayashi building.
2-1 You can find Hayashi building easily.
3-1 It's just next to a tall tower called TV tower.
4-1 Perhaps 4-2 it is written in Japanese. 4-3 So I write it for you.
5-1 I see.
6-1 As you took a trouble to come here. 6-2 You should see Nagoya castle.

(b) 同時翻訳単位に対応する逐語訳

図 2: 同時翻訳単位に分割された日本語対話文とその逐語訳

表 2: 同時翻訳単位境界になる割合

項目	発話単位境界である	発話単位境界でない
節境界である	89.8% (53/59)	50.0% (65/130)
節境界でない	21.2% (23/208)	1.4% (29/2072)

要がある．そこで，漸進的に検出可能な情報として形態素境界，発話単位境界，節境界の利用可能性を検討するため，各境界と同時翻訳単位境界との関係について調査した．

3.1 分析データの概要

分析に用いたデータの規模を表 1 に示す．ここで，**発話単位境界**とは，原則として 200ms 以上のポーズで発話を分割した境界である．また，**節境界**とは，節境界解析プログラム CBAP[5] により分割された境界である．

同時翻訳単位境界と節境界との関係，及び，発話単位境界との関係を調査した．その結果，節境界であるか否か，発話単位境界であるか否かによって同時翻訳単位境界になる割合 (表 2 参照) に違いがみられた．これ以降では，表 2 の 4 つのパターン別に，同時翻訳単位境界になる場合とならない場合の違いについて分析する．

3.2 分析結果

以下に分析結果を示す．ただし例文中では，節境界，発話単位境界，同時翻訳単位境界を示す記号として，それぞれ “**/節境界の種類/**”,“**/PS/**”,“//”を利用する．

A：節境界でかつ発話単位境界である場合

ほとんどが，同時翻訳単位境界と一致した．節境界の種類ごとに同時通訳単位境界になる割合を調べたところ (図 3 参照),「補足節」の場合は同時翻訳単位境界になることはなかった．また,「並列節」「主題ハ」の場合，同時翻訳単位境界にならない場合があった．以下では，「並列節」と「主題ハ」について同時翻訳単位にならない場合を示す．

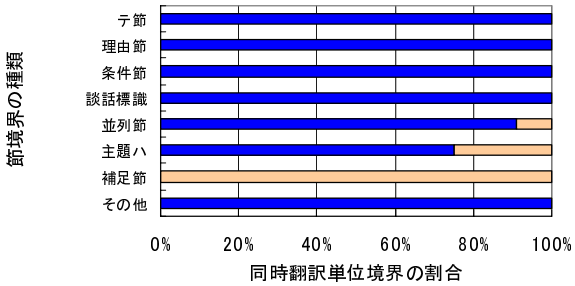

図 3: 同時翻訳単位境界と節境界との関係：A の場合

節境界の種類：「並列節」

- 「けれども」が前置きとして話題を持ち出す役割をしている．
  - 十二月二十日なんですけれども**/並列節ケレドモ/**東京発ロサンゼルスラスベガス行きはすべて満席になってございますが。
  - The flights from Tokyo to Los Angels or to Las Vegas on December twentieth are fully booked.

節境界の種類：「主題ハ」

- 疑問文である．
  - お部屋のほうは**/主題ハ/**どうされますか。
  - What type of room would you like?
- 述部が「存在」を表す動詞である．
  - コアラは**/主題ハ/**いません。
  - You can not see koala bears.

B：節境界でかつ発話単位境界でない場合

約半数が同時翻訳単位境界と一致した．節境界の種類ごとに同時翻訳単位境界になる割合を調査したところ (図 4 参照)，節境界の種類が「談話標識」,「並列節」,「理由節ノデ」の場合はすべてが,「条件節タラ」の場合

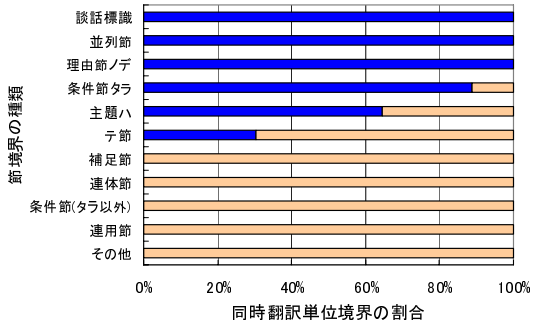

図 4: 同時翻訳単位境界と節境界との関係：B の場合

はほとんどが，同時翻訳単位境界であった.「主題ハ」,「テ節」の場合は，同時翻訳単位境界にならならい場合があった．このうち「主題ハ」は A の場合と同様の特徴であった.「テ節」における同時翻訳単位境界にならない場合には，以下のような特徴がみられた.

節境界の種類：「テ節」

- テ形の動詞「歩く」に接続助詞「て」が接続する.
  - 歩いて/**テ節**/十分ですね。
  - It is ten minutes' walk.
- テ形複合動詞になっている.
  - 林ビルはすぐに見つけて/**テ節**/いただけると思います。
  - You can find Hayashi building easily.

C：節境界でなくかつ発話単位境界である場合

境界の直前の形態素を調査したところ,「も」「を」のときは同時翻訳単位境界と一致し,「で」「が」「に」のときに同時翻訳単位境界になりやすいことが分かった.また，直前の形態素が「で」「が」「に」で同時翻訳単位境界にならない場合には次のような特徴があった.

直前の形態素：「で」「が」「に」

- 境界の直後に述部がくる
  - 二つまでですと無料で/**PS**/お預かりできます。
  - We can keep up to two bags for free.

また，直前の形態素が「で」「が」「も」「を」「に」以外で同時翻訳単位境界になる場合には，次のような特徴があった.

直前の形態素：「で」「が」「も」「を」「に」以外

- 発話の途中で相手発話が挿入されたために,発話を途中で打ち切り,「はい」などの相槌を行っている
  - 駅から/**PS**/はい。
  - From the station. Yes.

D：節境界でなくかつ発話単位境界でない場合

直前の形態素を調べたところ，直前の形態素が「で」「が」「に」のときに，同時翻訳単位境界なる割合が比較的高く，さらに以下のような特徴がみられた.

直前の形態素：「が」「で」

- C と同様，境界の直後に述部が続く場合には，同時翻訳単位境界になりにくかった.

直前の形態素：「に」（副詞）

- 文頭の副詞 (「最後に」,「次に」など) である
  - 最後に//クレジット番号と連絡先をお願いします。
  - Lastly, could I have your credit number and contact address, please?

直前の形態素が「で」「が」「に」以外の場合，下記の場合に同時翻訳単位境界になることがあった.

直前の形態素：その他 (「で」「が」「に」以外)

- 文頭の副詞である
  - 多分//日本語と思いますので今からお書きします。
  - Perhaps, it is written in Japanese. So I write it for you.

4 同時翻訳単位境界の判定ルール

前節で示した分析結果より，節境界，発話単位境界,及び,形態素情報によって同時翻訳単位境界を判定する.

4.1 同時翻訳単位境界の判定ルール

同時翻訳単位境界の判定ルールを図 5 に示す.A,B,C,D ごとの判定ルールの詳細を以下に示す.

A：節境界でかつ発話単位境界である場合

- 節境界の種類が「主題ハ」「補足節」以外である
- 節境界の種類が「主題ハ」で，以下の条件を満たす
  - 直後の形態素が疑問詞でない
    (疑問詞：「何」,「どこ」,「いつ」など)
  - 直後の形態素が「存在」を表す動詞ではない
    (「存在」の動詞：「いる」,「ある」，など)

B：節境界でかつ発話単位境界でない場合

- 節境界の種類が「談話標識」,「並列節」,「理由節ノデ」,「条件節タラ」である
- 節境界の種類が「主題ハ」で，以下の条件を満たす
  - 直後の形態素が疑問詞でない
    (疑問詞：「何」,「どこ」,「いつ」など)
  - 直後の形態素が「存在」を表す動詞でない
    (「存在」の動詞：「いる」,「ある」など)
- 節境界の種類が「テ節」で，以下の条件を満たす
  - テ形の動詞 (「歩いて」など) でない
  - テ形複合動詞 (「〜ている」など) でない

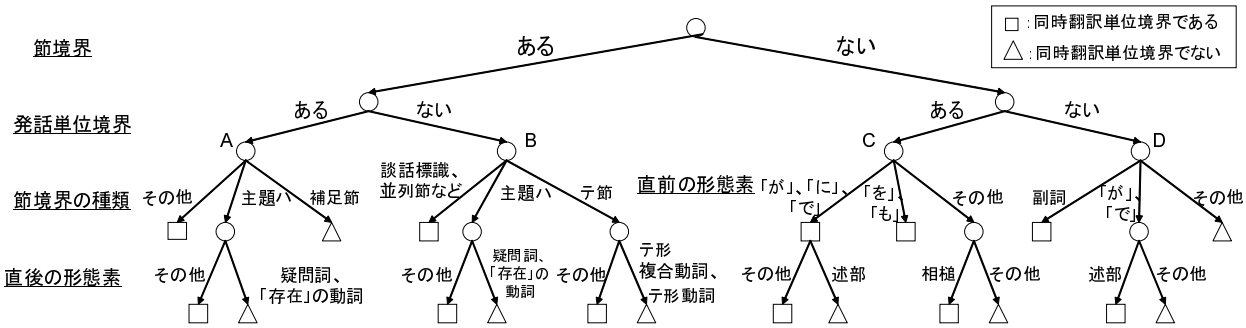

図 5: 同時翻訳単位境界の判定ルール

表 3: 実験結果

場合	精度	再現率	F 値
全体	61.2% (256/418)	80.0% (256/320)	69.4
A	87.2% (68/ 78)	97.1% (68/ 70)	91.9
B	69.5% (137/197)	85.1% (137/161)	76.5
C	45.7% (16/ 35)	61.5% (16/ 26)	52.5
D	41.0% (34/ 83)	54.0% (34/ 63)	46.6

C：節境界でなくかつ発話単位境界である場合

- 直前の形態素が「を」,「も」である
- 直前の形態素が「が」,「に」,「で」で，以下の条件を満たす
  - 直後の形態素が述部でない
- 直後の形態素が相槌（「はい」など）である

D：節境界でなくかつ発話単位境界でない場合

- 直前の形態素が「が」,「で」で，以下の条件を満たす
  - 直後の形態素が述部でない
- 文頭の副詞（「最後に」,「多分」など）である

5 同時翻訳単位境界の分割実験

前節で示した同時翻訳単位境界の判定ルールを用いて日本語対話文を同時翻訳単位境界に分割する実験を行った．実験では，名古屋大学同時通訳データベース [7] のうち，3 節の分析に用いなかった 14 対話分の対話データを用いた．実験データには，320 個の同時翻訳単位境界が存在した．

表 3 に A,B,C,D の各場合の結果とともに境界判定の精度と再現率を示す．A の場合はもともと同時翻訳単位境界になりやすいということもあり，高い精度と再現率になった．B の場合も比較的再現率は高いが，精度がやや低かった．その大きな要因は，節境界の種類が「主題ハ」であるときの判定誤りであり，誤り全体の半数以上を占めていた．C の場合は，精度と再現率ともに低かった．直前の形態素が「を」の場合に同時翻訳単位境界と判定したことが原因であった．本実験では，直前の単語が「を」の場合が 7 境界存在したが，いずれも同時翻訳単位境界にならなかった．D の場合も，精度, 再現率ともに低かった．直前の単語が「で」「が」のときに判定を誤る場合が多かった．

6 おわりに

本論文では，同時的な日英対話翻訳のための翻訳単位について検討するために，日英同時通訳コーパスの日本語対話文を独立かつ即時的に訳出可能な単位に人手で分割し，その境界の特徴を分析した．また，節境界，節境界の種類，発話単位境界，及び，境界の前後の形態素情報の特徴に基づいて作成した，同時翻訳単位境界を判定するルールについて述べた．さらに，そのルールを用いて日本語対話文を同時翻訳単位境界に分割したところ，精度が 61.2%，再現率が 80.0%であった．

今後は，本実験により得られた知見を利用して，ルールの改善，及び，統計的手法の利用を検討する．

参考文献

[1] R. Frederking, A. Blackk, R. Brow, J. Moody, and E. Stein-brecher, “Field Testing the Tongues Speech-to-Speech Machin Translation System,” Proceedings of the 3rd International Conference on Language Resources and Evaluation, pp. 160-164, 2002.

[2] R. Isotani, K. Yamada, S. Ando, K. Hanazawa, S. Ishikawa and K. Iso, “Speech-to-Speech Translation Software PDAs for Travel Conversation,” NEC Research and Development, 44, No.2, pp. 197-202, 2003.

[3] H. Kashioka, T. Maruyama, H. Tanaka, “ Building a Parallel Corpus for Monologues with Clause Alignment,” Proceedings of MT Summit IX, pp. 216-223, 2004.

[4] F. Liu, Y. Gao, L. Gu and M. Picheny, “Noise Robustness in Speech to Speech Translation,” IBM Tech Report RC22874, 2003.

[5] 丸山岳彦， 柏岡秀紀，熊野正， 田中英輝, “ 日本語節境界検出プログラム CBAP の開発と評価，” 自然言語処理, 11, 3，pp. 39-68, 2004.

[6] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo and S. Yamamoto, “A Japanese-to-English Speech Translation System:ATR-MATRIX,” Proceedings of 5th International Conference on Spoken Language Processing, pp. 957-960, 1998.

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