日本語対話文における同時通訳単位
－音声対話コーパスを用いた分析－

丁　喆† 笠　浩一朗‡ 松原 茂樹§ 吉川　正俊‡

†名古屋大学工学部 ‡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

ding@dl.itc.nagoya-u.ac.jp

1　はじめに

近年，音声・言語処理技術の向上により，音声翻訳に関する研究が活発に行われている．現在開発されている対話を対象とした音声翻訳システムの多くは，話者の発話の終了を待ってから訳出を開始する逐次通訳の形態を採用している．しかし，対話に出現する発話には，重文や複文などの長文の出現も少なくないため，発話の途中で訳出を開始する同時通訳が望ましい．実際，通訳者を介した対話データの分析においても，対話の効率および円滑さにおける同時通訳の優位性が明らかにされている [1].

同時通訳では，通訳者は話者の発話途中で訳出を開始することから，話者の発話の一部を通訳単位として捉え，その訳を早い段階で訳出している．本研究では，このような発話の一部を**同時通訳単位**と呼ぶ．同時通訳単位とは，同時的にかつ独立的に翻訳可能な言語的単位である．機械翻訳の場合，同時通訳は，話者の発話を同時通訳単位に漸進的に分割し，その単位ごとに訳を出力することによって実現できると考えられる．講演などの独話の翻訳システムにおいて，翻訳の処理対象として適切な言語単位が検討されている [2]．しかしながら，対話における同時通訳単位を形成する適切な言語単位については，これまでに明らかにされていない．

対話における同時通訳単位を明らかにするために，実データの定量的な観察が必要である．一つの方法として，日本語対話文を同時通訳単位に分割する境界（以下，**同時通訳単位境界**）が付与されたコーパスの利用が考えられる．しかしながら，同時通訳単位境界が付与された対話コーパスは未だに存在しておらず，また，人手で大量の同時通訳単位境界を付与することは大きなコストを要するため難しい．

そこで本稿では，日英対訳コーパスの日本語対話文に同時通訳単位境界を自動的に付与する手法を提案する．本手法では，同時通訳単位境界と節境界が類似した関係にあることを利用して，節境界を同時通訳単位境界の重要な手掛かりとして利用する．名古屋大学 CIAIR 同時通訳コーパス [3] の日本語対話文の一部に人手で付与した同時通訳単位境界と節境界との関係を分析した．また上記のコーパスに収録された日本語対話文を用いて，同時通訳単位への分割実験を行い，本手法の有効性を確認した．

2　同時通訳単位境界と節境界の関係

日本語対話文

(2.1)　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．

を同時通訳単位に分割することを考える．分割するうえで，対応する英語訳が手掛かりとなるため，(2.1) の英語訳

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

を利用する．

(2.1) と (2.2) の文の構成要素間の対応を考慮して，(2.1) に対して同時通訳単位境界を示す記号“ // ”を付与すると，

(2.3) 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．

のようになる．(2.3) のような同時通訳単位で翻訳処理を行えば，「今のところ」などの同時通訳単位とそれに対応する英語訳“ For now ”などの生起順

序が同じであるため，同時通訳単位ごとに翻訳可能である．ここで，(2.1) に対して節境界プログラム CBAP[4] により節境界を付与すると，

(2.4)　今のところ予定通りですが/**並列節ガ**/出発が遅れる/**連体節**/可能性がありますので/**理由節ノデ**/ご了承くださいませ．

のようになる．

(2.3) と (2.4) を比較すると，同時通訳単位境界と節境界が類似した単位であることが推測される．また，節境界は CBAP を用いることにより，自動的に付与することが可能になる．そのため同時通訳単位境界を自動的に付与するのに，節境界情報を利用することは効果的であると考えられる．

しかしながら，対話文のあらゆる節境界が同時通訳単位境界になるわけではない．(例えば，(2.4) の「連体節」)．また，節境界以外の箇所が同時通訳単位境界となる場合（例えば，(2.3) の最初の境界）も存在する．そのため，節境界と同時通訳単位境界との関係を明らかにする必要がある．

3　音声対訳コーパスを用いた分析

3.1　分析の概要

節境界と同時通訳単位境界の関係を調べるため，名古屋大学 CIAIR 同時通訳コーパス [3] に収録された対話データを用いて分析した．対話データのドメインは，空港やホテルなどの旅行会話である．分析には，コーパスに含まれる全 216 対話から無作為に選んだ 21 対話の日本語対話文 941 文を用いた．フィラーや言い直しなどを含んだデータに対して CBAP を用いると精度が低下するため，日本語対話データからフィラーや言い直しなどを除去してから CBAP を用いて節境界を付与した．また，英語訳を参照して人手により日本語対訳文に同時通訳単位境界を付与した．

3.2　分析の結果

同時通訳単位境界と節境界の総数を図 1 に示す．CBAP により分割された節境界の数は 574 箇所であり，そのうち 292 箇所は同時通訳単位境界であった．また，節境界以外の同時通訳単位境界の数は 95 箇所存在し，同時通訳単位境界の総数は 387 箇所であった．すなわち，節境界でなく同時通訳単位境界であるものが同時通訳境界の総数の約 25%(95/387) を占めている．また，節境界のうち約半数が同時通訳境界になることがわかった．

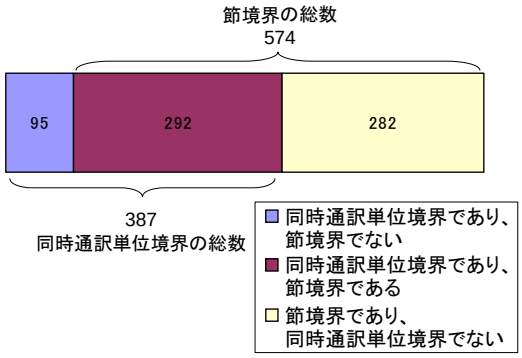

図 1: 同時通訳単位境界と節境界の総数

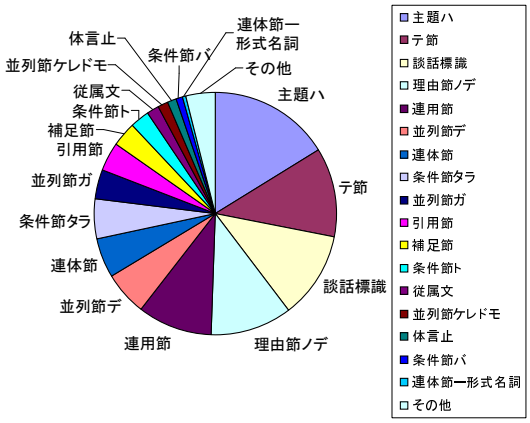

図 2: 節境界の種類と出現頻度

次に，節境界の種類と同時通訳単位境界との関係を分析する．節境界 32 種類の出現頻度を図 2 に示す．32 種類の節境界のうち 4 回以上出現したものは 17 種類あり，それらが同時通訳単位境界になるそれぞれの割合を図 3 に示す．従属文，連用節，連体節，補足節，引用節は同時通訳単位境界になりにくく，並列節テレドモ，並列節ガはほぼ 100%に近い割合で同時通訳単位境界となることがわかった．

4　同時通訳単位への分割

4.1　基本的な考え方

前節の分析結果より，節境界のうち同時通訳単位境界でないものが約半数あり，それらを区別する必要がある．また，同時通訳単位境界のうち節境界でないものが約 25%存在するため，これを無視することはできない．本研究では，節境界が同時通

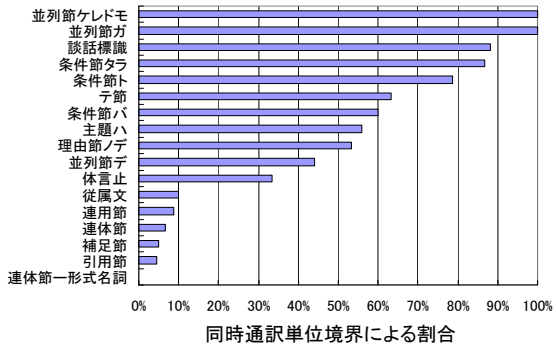

図 3: 節境界の同時通訳単位境界になる割合（4 回以上出現した節境界のみ）

訳単位境界になるか否かの判定について検討することを目的とし，同時通訳単位境界で，かつ，節境界でないものについては，今後の課題とする．

前節の図 3 で示した節境界別の同時通訳単位境界になる割合より，節境界は，同時通訳単位境界に極端になりにくいものと同時通訳単位境界になりやすいものに分けることができる．そこで，同時通訳単位境界になる割合がある閾値 T よりも低い節境界を同時通訳単位境界になりにくいものとみなし，同時通訳単位境界の候補から取り除くことにする．

同時通訳単位境界になる割合が高いの節境界であるものの，実際には，同時通訳単位境界にならない場合の具体例を以下に示す．

(4.1)　空港からですと/条件節ト/このバスに乗っていただくのがいいと思います．

(4.2)　I think you should take this bus from the airport.

ここで，日本語対話文 (4.1) の節境界「条件節ト」より前に出現する単語「空港」と後ろに出現する単語「バス」は，英語訳 (4.2) の英単語 “ airport ” と “ bus ” にそれぞれ対応している．また，このように節境界の前に出現する単語と後ろに出現する単語において，それに対応する英単語の生起順序が入れ替わっている場合，「条件節ト」は同時的に翻訳できないことになるので，同時通訳単位境界ではないといえる．

そこで，節境界の前に出現する単語と後ろに出現する単語のそれぞれに対応する英単語の生起順序が逆になる場合，同時通訳単位境界でないと判定することができる．

以上のことを用いて，同時通訳単位境界を自動的に付与するアルゴリズムを作成した．

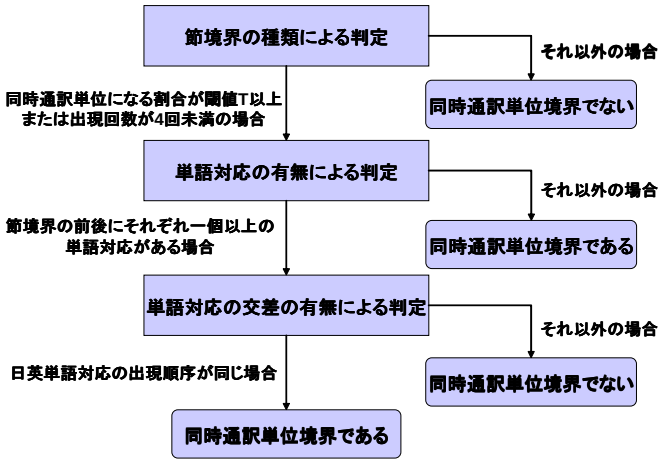

図 4: 同時通訳単位境界の自動付与のアルゴリズム

4.2　自動分割アルゴリズム

節境界に対して，以下のアルゴリズムを適用することにより，同時通訳単位境界であるか否かを判定する．アルゴリズムを図 4 に示す．

1. 節境界の種類による判定

同時通訳単位境界になる割合が閾値 T よりも低い場合，同時通訳単位境界でないと判定する．それ以外の場合は，2. へ進む．ただし，分析結果で出現頻度が 4 回未満の節境界の種類の場合も，数値に信頼性が低いことから，節境界の種類だけにより同時通訳単位境界でないと判定できないので，同様に 2. へ進む．

2. 単語対応の有無による判定

節境界の前と後ろに，それぞれ 1 個以上の単語対応が存在する場合には，3. へ進む．それ以外の場合は，同時通訳単位境界と判定する．

3. 単語対応の交差の有無による判定

節境界の前に存在する単語に対応する英単語が，節境界の後ろに存在する単語に対応する英単語よりも後ろに生起する場合，同時通訳単位境界でないと判定する．それ以外の場合，すなわち節境界の前に存在する単語に対応するすべての英単語が，節境界の後ろに存在する単語に対応する英単語のどれよりも前に存在する場合，同時通訳単位境界であると判定する．

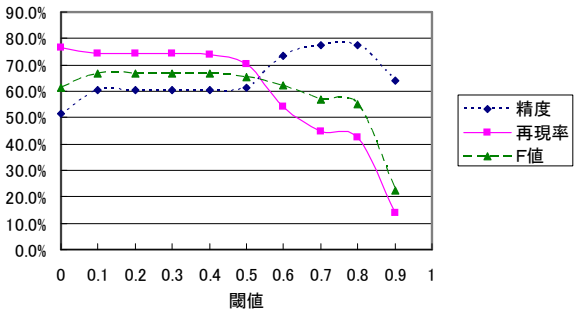

図 5: 閾値と節境界の種類による判定の関係

表 1: 単語対応情報による判定と性能評価

	精度 (%)	再現率 (%)	F 値
単語対応情報による判定をしない場合	60.6	74.3	66.8
単語対応情報による判定をする場合	66.6	72.2	69.3

5 評価実験

5.1 実験の概要

本手法の有効性を確認するため，対訳データに対して同時通訳単位境界を付与する実験を行った．実験には，CIAIR 同時通訳コーパスに収録されている日本語対話データ 14 対話に含まれる 529 文を用いた．ただし，このデータは 3 節の分析で用いたものとは重複しないように選んだ．また，閾値 T により同時通訳単位境界の候補を絞るときに，3 節の分析結果より得られた節境界の種類別の同時通訳単位境界になる割合を用いた．単語対応の情報は，対話文とその英語訳の間に人手により付与されたものを用いた．

5.2 実験結果

前節で示したアルゴリズムの 1. で用いる閾値を 0.1〜0.9 まで 0.1 刻みで変更した実験結果を図 5 に示す．図 5 より，閾値 T を 0.1，0.2，0.3 のいずれかのときに，F 値が最大となった．これは，同時通訳単位境界になる割合が極端に低い種類の節境界を同時通訳単位境界でないと判定することに効果があることを示している．また，閾値 T を 0.1 としたときの，単語対応情報による判定を行わない場合と行う場合の結果を表 1 に示す．表 1 より，単語対応情報による判定を行うことによって F 値が 2.5 ポイント向上しており，ある節境界が同時通訳単位境界になるか否かを判定するのに単語対応情報が有用であることがわかった．

6 まとめ

本稿では，日英対訳コーパスの日本語対話文に同時通訳単位境界を自動付与する手法を提案した．本手法は，節境界の種類と単語対応情報により節境界が同時通訳単位境界になるか否かの判定を行う．本手法の有効性を確認するため，対訳コーパスを用いて評価実験を行ったところ，同時通訳単位境界になる割合が 10%以下になる節境界の種類を同時通訳単位境界の候補から除外した場合 F 値が最大となった．また，単語対応情報を用いることにより F 値が 2.5 ポイント上昇した．また，分析結果より，同時通訳単位境界と判定できなかったものは全体の約 25%を占めることが判明した．この原因は次の二つの場合に分けられる．一つは，節境界ではないものが同時通訳単位境界となる場合であり，もう一つは，本来は節境界で，かつ，同時通訳単位境界であるが，CBAP によって節境界と分割されない場合である．

参考文献

[1] 大原誠, 松原茂樹, 笠浩一朗, 河口信夫, 稲垣康善, “同時通訳を介した異言語間対話の時間的特徴 ― 逐次通訳との比較に基づく対訳コーパスの分析”, 通訳研究, No.3, pp.34-52 (2003).

[2] Hideki Kashioka, Takehiko Maruyama, Hideki Tanaka, “Building a Parallel Corpus for Monologues with Clause Alignment,” MT Summit IX, pp.216–223 (2004).

[3] Hitomi Tohyama, Shigeki Matsubara, Nobuo Kawaguchi, Yasuyoshi Inagaki, “Construction and Utilization of Bilingual Speech Corpus for Simultaneous Machine Interpretation Research,” Proceedings of 9th European Conference on Speech Communication and Technology (Eurospeech-2005), pp. 1585-1588 (2005).

[4] 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, Vol. 11, No3, pp. 39-68 (2004).