2U-09

音声翻訳研究のための話し言葉コーパスの構築

松原 茂樹†* 相澤 靖之‡ 河口 信夫‡* 外山 勝彦‡* 稲垣 康善‡

†名古屋大学言語文化部 ‡名古屋大学大学院工学研究科

*名古屋大学統合音響情報研究拠点

1 はじめに

音声，及び言語処理技術の進展にともない，近年，話し言葉翻訳がますます重要な研究テーマとなっている．すでに，対話翻訳の実験システムがいくつか提案されており [3]，特定タスクドメインでの異言語間対話の実現可能性が明らかになりつつある．ただし，対象となる対話は，発話者だけでなく通訳者の発話権をも保障することが前提となっており，今後はより自然なクロスリンガルコミュニケーションを目指し，翻訳技術を高度化することが望まれる．

このような目的に対して，話し言葉のパラレルコーパスを作成し分析することは，効果的な方法の一つである．パラレルコーパスは，通訳付きの言語データであり，通訳者の経験的知識を獲得するための基礎データ，統計的手法に基づく翻訳の対訳用例，さらには，システムを評価するためのテストセットとしての役割を担っており，これまでも機械翻訳技術の向上に重要な役割を果たしてきた [1, 4]．

本稿では，我々が現在，構築している話し言葉コーパスについて述べる．このコーパスの特徴として，

- 異なる言語間での対話だけでなく，講演通訳も収録した英語と日本語のパラレルコーパスである．
- 自然な音声コミュニケーションのため，同時通訳者による講演通訳，及び対話通訳を実現している．
- 同一の講演に対して経験度が異なる複数の通訳者を用意し，複数の講演通訳データを収録している．
- 話者，及び通訳者の発話をポーズで分割し，各々を一発話として収録している．各発話には開始時間，終了時間を付与している．

などがあり，話し言葉翻訳技術の向上はもちろん，通訳会話の分析や通訳理論の構築のための基礎データとしても活用できる．

2 データの収集

これまでに作成された話し言葉のパラレルコーパスでは，対話が収録の中心であった [1, 4]．これは，自動翻訳電話など，対話翻訳の基礎資料とすることが収録の目的であったことが背景となっている．しかし，講義や講演など，片方向の音声コミュニケーションもまた，話し言葉翻訳の対象であり，現在の通訳需要からみても，将来の講演通訳システムに対する期待は大きい．このような観点から，我々は，対話だけでなく，通訳者を介した講演音声も収録している．コーパスの内容を表 1に示す．本節の以下では，収録会話の形態，内容，ならびに，環境について述べる．

表 1: コーパスの内容

会話様式	講演，対話
使用言語	英語，日本語
発話様式	原稿あり発話，模擬対話
通訳様式	同時通訳
データ様式	音声，書き起こしテキスト

表 2: 講演のテーマと対話のトピック

講演	政治，経済，技術，言語，都市，環境 など
対話	旅行（入国審査，空港，ホテル，電話予約など）

2.1 会話の形態

円滑な異言語間コミュニケーションを実現する上で，通訳者の発声タイミングが重要である．話者と通訳者との間の発声の同時性が高いほど，

- 講演では，聴衆が，通訳者の発声と話者の振舞いを結び付けて理解すること
- 対話では，話者が，効率的で結束性の高いインタラクションを遂行すること

が可能となる [2]．同時通訳システムのための基礎データの提供を目指し，同時通訳者を介した英語講演，日本語講演，及び英語日本語間対話を収録している．

2.2 収録の内容

講演では，英語話者の発声と通訳者による日本語発声，ならびに，日本語話者による発声と通訳者による英語発声，をそれぞれ収録している．同時通訳は，豊富な訓練を必要とする高等技術であり，通訳者の熟練度により，結果に大きな違いが生じる．熟練度の影響を調査するために，同一の講演に対して経験年数が異なる複数の同時通訳者を用意し，その通訳結果を収録している．また，通訳者が講演を実際に同時通訳する場合には，原稿やレジュメなどをもとに，あらかじめ講演内容を把握しておくことが普通である．そこで，実際に近い状況を設定するために，さらには，ある程度の通訳の品質を確保するために，通訳者には事前に講演原稿を渡し，内容に精通してもらうようにしている．さらに，社会的にみて頻度の高いテーマ及び内容 (表 2参照) を採用した．話者には，原稿の読み上げでなく，できるだけ講演調で話すように依頼している．

一方，対話では，英語話者と日本語話者の通訳者を介した対話を収集している．通訳の品質を高めるために，英日及び日英の二人の同時通訳者を用いている．対話ドメインとして，これまでに構築されているいくつかの対話データベース [4] と同様，「旅行会話」を採用し，表 2に示すよう

Spoken Language Corpus for Machine Interpreting Research: Shigeki Matsubara, Yasuyuki Aizawa, Nobuo Kawaguchi, Katsuhiko Toyama and Yasuyoshi Inagaki (Nagoya University)
matubara@lang.nagoya-u.ac.jp

表 3: 現在までに文字化されたデータの量

時間数 (延べ)	講演	40 時間 (通訳者： 24 時間)
	対話	32 時間 (通訳者： 16 時間)
	合計	72 時間 (通訳者： 40 時間)
総発話数	講演	13,553 文 (通訳者： 8,656 文)
	対話	8,676 文 (通訳者： 3,813 文)
	合計	22,229 文 (通訳者： 12,469 文)

に，空港やホテルなど，海外旅行において想定されるトピックをいくつか定めた．遂行される対話のドメインに応じて，対面，非対面の会話環境を作り上げ，収録を行っている．なお，対話様式は模擬的であるものの，できる限り自由な発話を収集するため，対話タスクと話者役割のみを事前に設定している．

2.3 収録の環境

実音響環境下での音声データを収集するために，教室レベルの録音環境で収録を行っている．同時通訳では，対象となる音声だけでなく，その発声者の表情や振舞いもまた，重要な情報となるため，通訳者は，話者をガラス越しに観察可能な専用のブースに入り通訳を行う．講演の収録では，話者には通訳者の音声が伝わらないようにし，聴衆の様子を窺いながら自らのペースで発声できるようにしている．一方，対話では，話者は，相手の話者の発声を通訳した結果のみを聞くことができ，通訳者は，両ネイティブ話者の発声を聞き取ることができる．すべて同一のスタンドマイクを使用し，話者とその通訳者の音声は，サンプリング周波数 16kHz，16 ビットでデジタル化し，デジタルオーディオテープに複数チャネル環境で収録している．

3 データの分析

3.1 音声データの文字化

収集した音声データの文字化は，人手によって行っている．データの言語学的分析として，話し言葉の特徴的現象である，フィラー，言い淀み，言い誤り，言い直しなどにタグを付与している．講演音声と通訳音声データの文字化の例を図 1に示す．現在までに文字化が完了しているデータ量を表 3に示す．

同時通訳技術の実現にあたり，システムが生成する内容とその出力タイミングが重要なポイントである．原言語と目標言語との間の語の生起順序の違いのため，通訳文の品質を確保し，かつ，入力に対する出力の同時性を満たすことは難しい．我々は，通訳者が，どのような発話をどのようなタイミングで発声するのかを調査するため，話者，及び通訳者の発話をポーズで分割し，各々を発話単位と定め，その開始時間及び終了時間を記録している．

3.2 音声データの視覚化

データの文字化作業で付与した発声時間情報をもとに，話者と通訳者の発声タイミングを視覚的に表示するツールを作成した．対話に対する音声の視覚化の例を図 2に示す．図中，右側の発話が，左のグラフ内に記された時間に行われたことを表している．グラフの左から順に，英語話者，英日通訳者，日本語話者，日英通訳者の発声時間であり，話者と通訳者の発声の重なり具合がわかる．

(F えー) 理事長からのご紹介あったように (F えー) 私 (わたし) は長い間外交官として (F ま) 様々な <H> 国で勤務してきたわけですが (F ま) その中でも大使として (F え) ジュネーブでの勤務がもっとも長く (F ま) 多少なりとも国連についてもお話しできるのではないかと思い本日の講演の依頼をお受けした次第であります．

(V_breath) As (V_breath) he (V_breath) mentioned I served as a diplomat for many years and I was stationed in various countries.
(V_breath) In particular I stayed in Geneva as an ambassador of Japan the longest.
(V_breath) So I think I could somewhat talk to you about the United Nations.
That's why I accepted the invitation today.

図 1: 講演音声と通訳音声の文字化

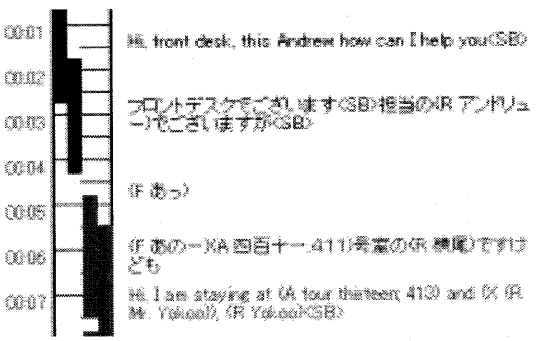

図 2: 対話音声と通訳音声の視覚化

4 おわりに

本稿では，現在，収集している話し言葉コーパスについて述べた．本コーパスは，同時通訳者を介した講演，及び対話データから構成されており，話し言葉処理のための言語データとして，また，同時通訳に有用な知見を獲得するための基礎資料として利用できる．引続き，コーパスの収集を進めるとともに，言語処理に有用な各種タグ付け，ならびに，対訳アライメント作業の実施を予定している．

謝辞

データの作成にあたり，ATR 音声言語通信研究所の竹沢寿幸，隅田英一郎の両氏に有益な御助言をいただきました．(株) インターグループには，データの収録に御協力いただきました．記して感謝します．本研究の一部は，文部省科学研究費補助金 COE 形成基礎研究費 (課題番号 11CE2005「多元音響の統合的理解」代表 板倉文忠) による．

参考文献

[1] 江原，小倉，篠崎，森元，樽松: 電話またはキーボードを介した対話に基づく対話データベース ADD の構築，情報処理学会論文誌, 33(4), pp. 448 456 (1992).

[2] Matsubara, S. and Inagaki, Y.: Incremental Transfer in English-Japanese Machine Translation, *IEICE Transactions on Inf. & Sys.*, E80-D(11), pp. 1122 1129 (1997).

[3] Takezawa, T. et al.: Japanese-to-English Speech Translation System: ATR-MATRIX, *Proceedings of ICSLP-98*, pp. 957–960 (1998).

[4] 浦谷，竹沢，田代，森元，匂坂: ATR の新音声言語データベース，情報処理学会第 48 回全国大会講演論文集 (3), pp. 79–81 (1994).