同時通訳研究のための対訳コーパスの設計と構築

遠山仁美† 松原茂樹‡ 笠浩一朗† 河口信夫‡ 稲垣康善*

†名古屋大学大学院情報科学研究科社会システム情報学専攻

‡名古屋大学情報連携基盤センター学術情報開発研究部門

*愛知県立大学情報科学部地域情報科学科

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

1 はじめに

近年，音声・言語に関する研究資源としての利用を目的として，様々な研究機関においてコーパスが作成されている (例えば [1, 2])．特に，大規模コーパスの重要性は広く認識されており，音声情報処理，自然言語処理，言語学，日本語教育，辞書編纂など，幅広い領域で利用されている．

名古屋大学統合音響情報研究拠点 (以下，CIAIR) では，マルチリンガルコミュニケーション研究環境の実現を目指し，1999 年度から 2003 年度までの 5 年間にわたり，同時通訳コーパスを構築してきた．全体で約 182 時間の音声を収録し，音声の文字化，視覚化，および，言語分析を完了している．文字化データのサイズは単語数にして約 100 万語に達し，世界最大の同時通訳コーパスと位置付けられる．さらに，コーパスの活用を支援するために，データ分析用ソフトウェアツールを開発している．Web サーバ上で実行可能なソフトウェアとして実現しており，ユーザはブラウザを使用することによって，データを容易に参照できる．

本稿では，名古屋大学 CIAIR 同時通訳コーパスについて，設計，収集，構築，および，利用について詳述する．最後に，認知科学，言語学など，領域を超えた研究分野への応用可能性について触れる．

2 コーパスの設計

近年，世界のグローバル化にともない，異言語間コミュニケーション支援環境の実現が望まれている．同時通訳コーパスは，話し言葉翻訳技術の向上，ならびに，通訳理論の構築を目指し， 名古屋大学 CIAIR における音声言語資源の整備の一環として作成された．

既に対話翻訳システムとしては，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．しかし，これらの通訳スタイルは同時通訳ではなく逐次通訳である．より自然なクロスリンガルコミュニケーション支援を目指すためには，話者が通訳者の訳出状況を考慮せずに，自分のペースで話すことができる，同時翻訳技術の実現が望まれる．そのために，大規模な同時通訳コーパスを構築し，分析することは，効果的な方法の 1 つである．

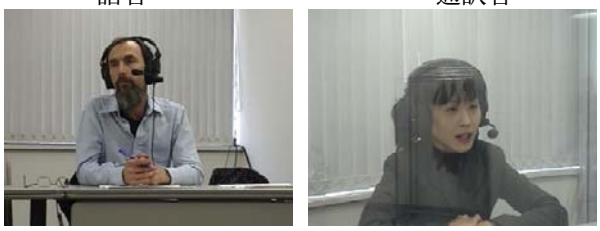

図 1: 音声収録の様子

大規模コーパスの収集は，それに費やされる膨大なコストを勘案すると，将来における幅広い利用可能性を考慮し，汎用性を備えた設計が要求される．本データベースは，多様なデータを収集するために，独話 (monologue) および，対話 (dialogue) の同時通訳音声をいくつかの日常的なトピックを設定して収録した．対象言語は英語と日本語とし，その双方向音声を収録した．

3 データの収集

3.1 収録の環境

名古屋大学 CIAIR では，実音響環境下での音声データを収集することを重視しており，収録は教室レベルの録音環境を採用した．また，同時通訳者にとって，話者の発話だけでなく，表情や振る舞いも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できる通訳専用のブースに入り，通常行われる同時通訳とほぼ同じ環境下で収録を行った (図 1 参照)．収録を通して全て同一のスタンドマイクを使用し，話者と通訳者の音声を，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 (DAT) に複数チャンネル環境で収録した．

また，同時通訳者は，第一線で活躍しているプロの通訳者を起用し，高い通訳レベルを保証している．

3.2 独話データの収集

独話 (講演) の収録では，講演者が通訳者の通訳状況を気にせず，自分のペースで発話できるよう，講演者

```
0001 - 00:05:264-00:09:399 N:
The theme for this speech is going to be the American
0002 - 00:09:840-00:11:032 N:
Presidential debate
0003 - 00:11:424-00:13:391 N:
and who would be the
0004 - 00:13:640-00:15:215 N:
better president for America<SB>
0005 - 00:16:272-00:18:327 N:
(F um) Let's see, today is
0006 - 00:18:640-00:20:400 N:
December fifteenth
0007 - 00:20:696-00:24:407 N:
and it's been about a month and a half since
```

図 2: 独話音声の文字化データ (英語話者)

```
0001 - 00:06:660-00:08:355 I:
次のスピーチのテーマは
0002 - 00:09:344-00:10:071 I:
アメリカの
0003 - 00:10:816-00:13:439 I:
大統領選に対するディベートです<SB>
0004 - 00:14:048-00:14:623 I:
誰が
0005 - 00:15:184-00:17:055 I:
アメリカにとって良い大統領なのか<SB>
0006 - 00:19:224-00:21:648 I:
今日が十二月の十五日です<SB>
0007 - 00:22:784-00:25:247 I:
約一か月半経ちました<SB>
0008 - 00:27:152-00:27:983 I:
アメリカ
0009 - 00:28:432-00:32:303 I:
の(W 大(D りょ)統領選;大統領選)が始まってから一か月半です<SB>
```

図 3: 独話音声の文字化データ (英日通訳者)

には通訳者の音声が聞こえないようにした．一方，通訳者は通訳用ブースに入り，講演者の振る舞いが見える中で，ヘッドホンから流れる講演者の音声に対して，同時通訳する．講演の聴衆は，ヘッドホンを利用して，通訳者の音声を聴くことができる．

また，英語あるいは日本語の講演者に対し，複数の同時通訳者が通訳を行った．つまり，1つの講演者発話ソースに対し，複数の通訳データが存在する．従って，個人に特化しない多くの通訳事例を幅広く収集したり，1つの発話に対する，複数の通訳事例を比較することが可能である．また，通訳経験年数の違いによる訳出の特徴分析などにも利用でき，コーパスの汎用性を高めている．

3.3 対話データの収集

対話の収録では，英語話者と日本語話者の異言語間対話に対し，通訳の品質を高めるために，英日，日英の2名の同時通訳者を設置する形態をとった．また，会話における話者の発話権を確保するために，話者は相手話者の発話を通訳した結果のみを聞くことができるようにした．一方，通訳者は，対話全体の流れを把握するために，担当する話者の音声だけでなく，もう一方の話者音声も聞ける環境を設定した．また，可能な限り自然な対話を収集するため，話者役割と対話タスクの設定のみを行い，基本的には自由発話という様式で収録した． 例えば，ドメインがホテルの予約であれ

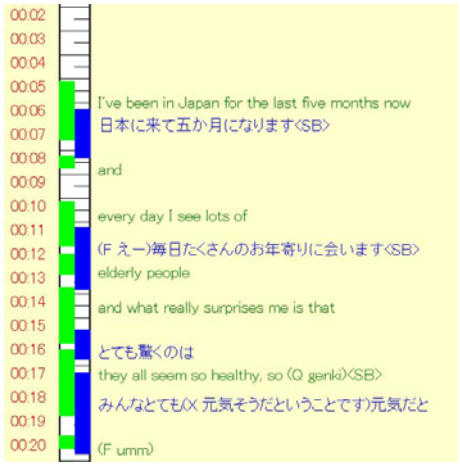

図 4: 音声データの視覚化

#	講演者発話	通訳者発話
0	0001 - 00:05:264-00:09:399 N: The theme for this speech is going to be the American	0001 - 00:06:440-00:08:207 I: (F え)次のテーマですが 0002 - 00:08:944-00:09:783 I: アメリカの
1	0002 - 00:09:840-00:11:032 N: Presidential debate	0003 - 00:10:296-00:12:775 I: (F え)大統領に関するディベート
2	0003 - 00:11:424-00:13:391 N: and who would be the 0004 - 00:13:640-00:15:215 N: better president for America<SB>	0004 - 00:13:096-00:14:424 I: そして誰が 0005 - 00:14:648-00:18:255 I: より良い大統領とアメリカのためになり得るかということですが
3	0005 - 00:16:272-00:18:327 N: (F um) Let's see, today is	0006 - 00:18:728-00:19:263 I: 今日が
4	0006 - 00:18:640-00:20:400 N: December fifteenth	0007 - 00:19:528-00:21:887 I: 十二月の十五日ですので
5	0007 - 00:20:696-00:24:407 N: and it's been about a month and a half since	0008 - 00:22:472-00:24:711 I: そして(F まあ)一ヶ月半ほど 0009 - 00:25:160-00:26:311 I: 経ってると思うんですが

図 5: 話者発話と通訳者発話の対訳対応データ

ば，客を担当する話者には予約したいホテル名と予約人数を，ホテル側を担当する話者には，空室状況のみを提示し，自由発話で対話を進めていく．

4 コーパスの構築

4.1 音声データの文字化

音声データの文字化は日本語話し言葉コーパス (CSJ) の書き起こし基準 [3] に準拠した (図 2,3 参照)．文字化作業は収録した全音声データ 182 時間分に対して行われた．以下にその基準を示す．

- 発話単位
  話者および，通訳者の音声を 200msec 以上のポーズ (無音声区間) で分割し，発話単位を定めた．
- 表記方法
  日本語音声に限り，片仮名で表記する「発音形」と漢字仮名まじりで表記される「基本形」の 2 種類で構成している．
- タグの付与
  - 発話 ID
    全発話に対し通し番号を付与した．
  - 時間情報タグ
    発声の開始時刻と終了時刻を付与した．

– 談話タグ
話し言葉に特有の言語的現象であるフィラーや言い淀みについて，談話タグを付与している．

4.2 音声データの視覚化

音声データを視覚化するために，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開発した． それによって，同時通訳者を介した講演，会話の様相をタイムチャートによって概観できる (図 4 参照)．これにより，音声，および文字データからは分かりにくい様々な通訳特有の現象を視覚的に観察することができる．

4.3 対訳対応データの作成

同時通訳システムの開発，同時通訳方略の構築において，話者発話とそれに対応する通訳者発話の対訳データを大規模に分析することは極めて重要であり，対訳アライメントを自動化するための研究が行われている [4]．しかし，同時通訳発話にはフィラーや言い淀み，意訳や誤訳，訳出の省略などが頻出するため，精度の高いアライメントは困難であり，その作業は人手に頼らざるを得ない．我々は人手による対訳アラインメント作業のコストを軽減するために，それを支援するツールを作成した．人手によって作成された対訳対応データは，ブラウザ上で並べて閲覧することができる (図 5 参照)．

4.4 データの規模

CIAIR 同時通訳コーパスは，現在までに，収録時間にして約 182 時間，単語数 (日本語は形態素数) にして約 100 万単語を収録している．

5 コーパスの利用

我々は，コーパスを利用し，同時通訳における通訳者発声タイミングや，通訳者を介したコミュニケーションの円滑さなど，通訳現象の個別の現象に着目し分析を行い，同時通訳プロセスの解明を進めてきた．

同時通訳システムの実現においては，通訳単位の決定，訳文の生成，その訳を出力するタイミングなどが重要な課題となる．そのためには，実際の同時通訳者の振る舞いを分析することが有効な手法の 1 つであり，さらに，分析対象となるデータが大規模であれば，定性的な分析結果を，さらに定量的に検証することも可能である．

5.1 データ分析による通訳理論の構築

5.1.1 通訳者発声タイミングの分析

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と推測される [5, 6]．我々は，本コーパスの時間情報と対訳対応データを用い，同時通訳の訳文生成タイミングに関する調査を行っている [7, 8]．

通訳者の発声速度，および，訳出開始の遅れ時間を分析し，実際の通訳者が行っている英日，及び，日英通訳における同時通訳単位と，その発声タイミングについて分析した．その結果，

- 同時通訳において，話者が接続詞を用いたときや，主部が特定できたときには，即座に訳出できる可能性がある．
- 訳出可能な情報の量に応じて発話速度を制御することにより，訳出遅れの少ない訳文生成が可能になる．

ことなどを明らかにしている．

5.1.2 同時通訳と逐次通訳の時間的な特徴分析

同時通訳と逐次通訳という 2 つの通訳スタイルの比較は，通訳理論研究の分野において，様々な議論がなされてきた [9]．我々は，異言語間対話の効率，円滑さという観点に着目し，コーパスを用いて分析を行っている [10]．その結果，

- 対話時間効率は，逐次通訳を介する対話に比べ，同時通訳を介する対話の効率は大幅に上がる．
- 通訳を介した対話におけるターンごとの話者待ち時間の平均は，英語話者，日本語話者の場合ともに，同時通訳を介することにより，対話の円滑さが大幅に向上する．

ことを明らかにしている．これらによって，より自然なクロスリンガルコミュニケーション支援環境における同時翻訳技術の有用性を確認している．

5.1.3 同時通訳方略の収集

同時通訳は，人間にとって極めて高度な言語処理活動である．同時通訳における重要なポイントとして，訳出が話者の発話に追従して遂行されなければならないという訳出タイミングに関する制約 (when-to-say) の問題と，原発話に対してどのような訳を生成するか (how-to-say) という問題ある．通訳者は膨大な訓練によって蓄積された訳出方略を駆使している．実際の通訳者の訳出方略を詳細に，かつ大量に調査することにより，同時通訳に有用な訳出パターンを収集し，それを通訳ルールとして利用することができる．我々は，英日対訳対応データを使用し，同時性を重視するための方略である，順送りによる訳出 (文末まで待たずに，文頭からどんどん訳出していく手法)，および，短縮 (省略) による訳出に該当する訳出パターンを収集している [11]．さらに，同一の英語講演に対し，複数 (最大 4 人) の通訳者の同時通訳データを用いることにより，全く同一の英文に対し，複数の訳出パターンを抽出している．

5.2 多分野への利用可能性

近年，様々な分野で，同時通訳を対象とした研究が行われている．そこでは，ある同時通訳データを綿密に分析する定性的研究がなされている．主な事例を以下に示す。

- 研究における事例
  - 認知科学・認知言語学
    同時通訳は，聞き取った発話を保持しながら，それを他の言語に変換し，さらに，変換した内容を聴き手に伝えなければならない．その言語処理の複雑さから，同時通訳は，ワーキングメモリ (一時的な記憶機能) の極限レベルでの作業事例として扱われるなど，認知科学，認知言語学の分野において，そのメカニズムが広く研究されている [12, 13, 14].
  - 言語学
    通訳者の通訳訓練として用いられる，シャドーイング (shadowing) やサイト・トランスレーション (sight translation) などのメソッドを，外国語学習に導入する有益性が研究されている [15, 16]．同時通訳者は聞き取った発話を語順に従って順送りに訳出していくことから，人間が母国語を聴解するプロセスと類似しており，第二言語獲得への効果があるとされる．
- 教育における事例
  近年，大学において，実践的で，高度な専門的資質を有する人材育成を目指し，通訳者を養成するためのカリキュラムを設ける大学が増えている．そこでは，通訳理論，通訳技術論などの講座が設けられている [17].

このような諸専門分野において，個々の研究成果に対し，さらに定量的な分析を行ったり，より多くのサンプルデータを収集するために，本コーパスを活用することができると考えられる．

6 まとめ

本稿では，名古屋大学 CIAIR 同時通訳コーパスについて，設計，収集，構築，さらに，利用について述べた．

本コーパスの作成目的は，同時翻訳技術の向上と通訳理論の構築であり，今後も引き続き言語処理に有用な各種タグ付けや対訳アライメント作業を行い，コーパスの高度化を行う予定である．

また，話し言葉に関する研究の進展にともない，大規模音声言語コーパスの需要は，認知科学や音声学，言語学に及ぶ諸分野においても大いに高まりつつある．本コーパスにおいても，より広範な研究分野で多面的に活用し，研究領域を超えた意見交換を行い，総合的に進展していくことが望ましい．

名古屋大学 CIAIR 同時通訳コーパスの配布については以下を参照されたい．

http://www.el.itc.nagoya-u.ac.jp/sidb/

謝辞 日頃，ご指導下さる名古屋大学大学院教授の渡邊豊英先生に深く感謝致します．本研究の一部は，文部科学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」) によります．

参考文献

[1] 前川 喜久雄, “『日本語話し言葉コーパス』の設計と実装,” 平成 15 年度国立国語研究所公開研究発表会論文集, 話し言葉のデータベース－『日本語話し言葉コーパス』－，pp.1-8, 2003.

[2] 柏岡 秀樹, 竹沢 寿幸, 中村 篤, 隅田 英一郎, “ATR の会話音声翻訳研究用データベース，”音声研究, Vol.4, No.2, pp.16-23, 2000.

[3] 小磯 花絵, 間淵 洋子, 西川 賢哉, 斉藤 美紀, 前川 喜久雄, “『日本語話し言葉コーパス』の書き起こしの仕様について,” 平成 15 年度国立国語研究所公開研究発表会論文集, 話し言葉のデータベース－『日本語話し言葉コーパス』－, pp.27-28, 2003.

[4] 高木 亮，松原 茂樹，稲垣 康善, “同時通訳コーパスの対訳アライメント手法とその評価,” 情報処理学会第 64 回全国大会講演論文集, 2002.

[5] 船山 仲他, “同時通訳における処理単位について,” 通訳理論研究, Vol.16, No.1, pp.4-13, 1996.

[6] 新崎 隆子, “同時通訳と逐次通訳における情報処理,” 通訳理論研究, Vol.14, No.2, pp.40-46, 1994.

[7] 高木 亮, 松原 茂樹, 稲垣 康善, “同時通訳コーパスを用いた通訳者の発声タイミングの分析,” 言語処理学会第 8 回年次大会発表論文集, pp.383-386, 2002.

[8] S. Matsubara, A. Takagi, N. Kawaguchi and Y. Inagaki, “Bilingual Spoken Monologue Corpus for Simultaneous Machine Interpretation Research,” Proceedings of LREC-2002, Vol.1, pp. 153-159, 2002.

[9] D. Gile, “Consecutive vs. Simultaneous: which is more accurate?,” 通訳研究, No.1, pp.8-20, 2001.

[10] 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.85-102, 2003.

[11] 遠山 仁美, 松原 茂樹, “同時通訳コーパスを用いた通訳者の訳出パターンの分析,” 信学技報, Vol.103, No.487, pp.13-18, 2003.

[12] 玉井 健, “通訳作業制御要因としての作動記憶,” 同時通訳における情報フローの認知言語学的検証, 平成 10-11 年度科学研究補助金研究成果報告書, pp.27-46, 2002.

[13] 船山 仲他, 笠原 多恵子, 西村 友美, “同時通訳における対訳遅延のメカニズム,” 同時通訳における対訳遅延の認知言語的研究, 平成 12-13 年度科学研究補助金研究成果報告書, pp.3-24, 2002.

[14] 苧阪 満里子, “脳のメモ帳　ワーキングメモリ,” 新曜社, 2003.

[15] 永田 小絵, “通訳訓練の外国語学習への応用,” 通訳理論研究, Vol.11, 1996.

[16] 染谷 泰正, 玉井 健, 鳥飼 玖美子, “はじめてのシャドーイング,” 学習研究社, 2003.

[17] 鳥飼 玖美子, “日本における通訳教育の可能性―英語教育の動向をふまえて,” 通訳理論研究, Vol.13, pp.39-52, 1997.