GCP 同時通訳コーパスの構築

東山翔平　今村賢治　内山将夫　隅田英一郎
情報通信研究機構
{shohei.higashiyama, kenji.imamura, mutiyama, eiichiro.sumita}@nict.go.jp

概要

同時翻訳システムの学習・評価のための GCP 同時通訳コーパスを構築した．本コーパスは，日・英各 100 万単語以上の原発話テキストと英・日への順送りの訳文テキストからなり，原発話・訳文中のチャンク境界の情報と，訳文中の省略可能な表現の情報を含む．本コーパスの構築方法と，チャンクと省略の観点から行った分析について報告する．

1　はじめに

即時性の高い多言語コミュニケーションが求められる状況において，人間の通訳者による同時通訳のように，原発話と同時並行的に翻訳を行う同時機械翻訳（同時翻訳）[1, 2] の実用化が期待されている．

同時通訳では，同時性の制約を克服するため，原発話をチャンクと呼ばれる意味的なまとまりに区切りながら，前のチャンクから順に訳出を行う「順送り」の方略が用いられる．同時通訳と同様に遅延の少ない同時翻訳を実現するには，順送り形式の対訳を基に翻訳システムを構築することが必要となる．そこで我々は，同時翻訳システムの学習・評価のための多言語の GCP 同時通訳コーパス[1)]を構築している．本稿では，そのうち日英・英日方向のコーパスについて報告する．

既存の日英・英日の同時通訳コーパスには，主に以下の三つがある．Tohyama ら [4] は，模擬講演・対話とその同時通訳音声を収録し，CIAIR 同時通訳データベース（SIDB）を構築した．松下ら [5] は，逐次または同時通訳付きの会見動画を書き起こし，英日・日英通訳データベース（JNPC）を構築した．Doi ら [6] は，既存の講演・会見音声を聴きながらの模擬的な同時通訳を実施・収録し，NAIST-SIC（NSIC）を構築した．上記コーパスと本コーパスの原言語データサイズ（時間数 N_{hour}，発話数 N_{utter}，単語数 N_{word}）を表 1 に示す．

1）　GCP は「グローバルコミュニケーション計画」[3] を指す．

表 1　同時通訳コーパスの原言語データサイズ

Corpus	ja-en			en-ja		
	N_{hour}	N_{utter}	N_{word}	N_{hour}	N_{utter}	N_{word}
SIDB [4]	37.9	23.0k	191k	39.6	22.6k	198k
JNPC [5]	4.4	1.3k	43k	56.7	18.7k	444k
NSIC [6]	171.0	N/A	N/A	117.5	N/A	N/A
Ours	–	67.4k	1.5M	–	116.5k	1.2M

これら既存の同時通訳コーパスと比較すると，本コーパスには主に次の特徴がある．(1) 音声を介さずにテキストを翻訳することで作成した対訳で，原発話テキストと順送りの訳文テキストの双方にチャンク境界を付与している．(2) 日本語原発話 150 万単語，英語原発話 119 万単語からなり，テキスト量での規模が大きい[2)]．

以降，§2–§4 にて，本コーパスの設計・構築方法と，本コーパスの性質の分析について報告する．

2　コーパス設計方針

書き言葉の逐次翻訳と比べて，同時翻訳システムに求められる要件として以下が挙げられる．

1. **原発話のスタイルへの対処**：非流暢でくだけた話し言葉の表現に対して適切な翻訳を行う．
2. **訳文のスタイルの制御**：同時通訳と同様の順送りによる低遅延の訳出を行う．
3. **訳出する情報の制御**：読み手・聞き手の負担削減と，読み上げ音声の出力時間削減のため，原発話中の冗長な表現や重要性の低い表現を除いた訳出を可能とする．

これらを満たすコーパスの構築方法として，原発話と同時通訳の音声を収録する方法が考えられる．しかし，通常の同時通訳では作業負荷の大きさから 15〜20 分程度で通訳者が交替するのが一般的 [7] であり，大規模な対訳テキストを確保するには多大な時間や費用を要する．そこで本研究では，作業効率性を重視し，プロの同時通訳者により，原発話テキ

2）　本コーパス及び JNPC の単語数計測には §4 で述べる方法を用いた．SIDB の単語数は文献 [4] に記載の値から計算した．

表 2　日本語原発話データの内訳

Domain		N_{utter}	N_{word}	N_{char}
BSDja	シナリオ業務会話	55.9k	682.8k	1,160.5k
SIDBja	模擬講演	2.7k	86.1k	144.1k
CSJ	実・模擬講演	6.9k	694.0k	1,117.2k
JNPC	実講演	2.0k	58.6k	100.2k
Total		67.4k	1,501.9k	2,522.1k

表 3　英語原発話データの内訳

Domain		N_{utter}	N_{word}
BSDen	シナリオ業務会話	55.5k	509.2k
AMI	模擬会議	51.3k	519.4k
TED	実講演	9.3k	159.2k
WPV	実講演	0.5k	7.0k
Total		116.5k	1,194.8k

ストを目視しながら順送りの訳文テキストを作成する方法を採用した．

要件 1 への対応として，講演・会議音声の書き起こしテキストや口語調で作成された業務会話のシナリオテキストを用いた．要件 2 への対応として，順送りの訳文を作成し，原発話及び訳文中にチャンク境界を付与した．要件 3 への対応として，原発話の内容全体を訳出した訳文と，重要性の低い内容を省いた訳文との 2 段階の訳文を作成した．

本方法は，主に音声を介さない点と明示的な時間的制約がない点が通常の同時通訳と異なる．しかし，上述の要件を満たすことで，同時翻訳システムのためのコーパスとして機能すると想定している．

3　コーパス構築方法

3.1　原発話データ

作業データのドメインとして，講演，会議，業務会話を対象とした．日本語原発話テキストには，日本語話し言葉コーパス（CSJ）[8] の一部と，Business Scene Dialogue Corpus（BSD）[9]，SIDB，JNPC の一部を使用した．英語原発話テキストには，BSD，AMI Meeting Parallel Corpus（AMI）[10]，IWSLT 2017 Evaluation Campaign データセット（TED）[11] の一部と，ウェブで公開されている英語講演動画を書き起こしたテキスト（Web Presentation Video; WPV）を使用した．使用した原発話データのサイズを表 2 及び表 3 に示す（N_{char} は文字数を表す）．

3.2　作業体制・方法

2020〜2022 年の間の延べ 11 か月の期間で，日英または英日の同時通訳実務経験 2 年以上を有する通訳者（延べ 63 名）に，以下の作業を依頼した．

表 4　原発話 U の訳文 ST1，ST2 と逐次翻訳文 CT の例

U	先日ご紹介した商品同様、|20 年未満の積立期間だと、| 途中解約した場合、| 戻ってくるお金は積立金の 0.8 倍になります。
ST1	Same as the product I introduced the other day,| if the funding term is less than 20 years,| and if you cancel it before the full term,| 80% of your funds will be returned.
ST2	Same as the product I introduced,| if the funding term is less than 20 years,| and if you cancel it before the full term,| 80% will be returned.
CT	If you cancel your plan before reaching the 20 year saving stage you'll be reimbursed only 80% of your saving, same as the product I showed you the other day.

1. 原発話テキストを，同時通訳において訳出の単位とする意味的なまとまり（チャンク）に分割し，チャンク境界を挿入する．
2. チャンクごとに，原発話を目的言語へ訳した順送りの訳文テキスト ST1 と ST2 を作成する [3]．
   - ST1：原発話の内容全体を訳出する．
   - ST2：原発話の意図の理解に支障がないような重要性の低い語句を省略し，原発話の内容のうち概ね 60〜80%以上を訳出する．
3. 原発話チャンク境界と対応する各訳文中の位置にチャンク境界を挿入する．

表 4 に，BSDja データの原発話 U と，通訳者により作成された同時翻訳文 ST1 及び ST2 の例を示す．原発話と各訳文に挿入されたチャンク境界を「|」，ST2 で省略された ST1 中の表現を波線で表した．参考に BSD 原データに含まれている逐次翻訳文 CT も示した．ST1 及び ST2 では先頭から順にチャンク単位で訳出されているのに対し，CT では冒頭の「先日」が末尾で訳されるなど構造の違いが見られる．

4　コーパス分析

システムの学習・評価に使用する上で，本コーパスがどのような性質・傾向を持つデータであるか，チャンク及び省略の観点から分析した．

原発話・訳文の単語数計測と，本節の分析の前処理として，日本語形態素解析に `unidic-cwj-3.1.0` [4] [12] と `MeCab` [5] [13] を使用し，英語品詞タグ付けに `flair/upos-english-fast` モデル [6] を使用した．また，句読点・補助記号（と日本語では空白）を削除し，日本語文において連続する「名詞-数詞」及び「記号-文字」の各単語列を 1 単語にまとめ上げた．

3）2021 年度は ST1→ST2 の手順，前年度は逆の手順で作成．
4）https://clrd.ninjal.ac.jp/unidic/back_number.html
5）https://taku910.github.io/mecab/
6）https://huggingface.co/flair/upos-english-fast

表 5　チャンク境界付き原発話データの統計情報

Domain	N_{word}	N_{utter}	L_{utter}	N_{chunk}	L_{chunk}
BSDja	647.0k	54.3k	5.1±3.5	92.0k	3.0±2.1
SIDBja	49.9k	1.6k	14.8±9.7	4.8k	5.1±3.4
CSJ	654.2k	6.6k	47.7±27.9	45.7k	6.8±4.4
BSDen	509.2k	53.4k	9.0±5.7	87.5k	5.5±3.3
AMI	519.4k	48.5k	10.1±12.8	76.8k	6.4±5.5
TED	159.2k	7.7k	17.1±11.9	18.6k	7.1±4.3

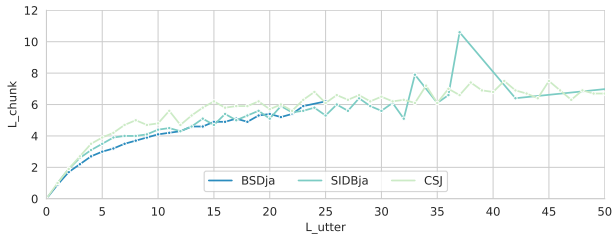

図 1　日本語原発話長ごとの平均チャンク長（内容語数）

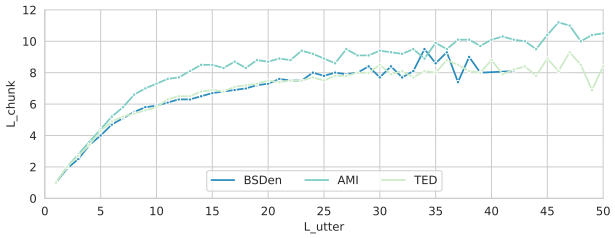

図 2　英語原発話長ごとの平均チャンク長（単語数）

4.1　チャンク長の分布

原発話のチャンク分割の細かさ／粗さは，訳文の構造の違いに影響するとともに，翻訳の遅延の大きさに直結する．そこで，原発話がどのような粒度でチャンクに分割されているかを分析した．

本分析には，システム評価セットを除いた表 5 の原発話データ（上段：日，下段：英）を使用した．表 5 には，原発話の件数 N_{utter} と長さ L_{utter}，通訳者が付与した原発話側チャンクの総数 N_{chunk} と長さ L_{chunk} を示した．発話長とチャンク長は，日本語では内容語数，英語では単語数で測り，平均 ± 標準偏差で示した．発話長は，日・英とも会話・会議（BSD, AMI）では短く，講演では長く，また発話長が長いドメインほど平均チャンク長も長い傾向がある．

そこで，原発話長ごとのグループに分けて，グループ内の平均チャンク長を計測した．結果を図 4 及び図 2 に示す[7)]．日・英の原発話とも，各ドメイン共通の傾向として，発話長が長くなるにつれて平均チャンク長も増加するものの，日本語ではチャンク長 6〜8 前後，英語では 8〜11 前後で概ね飽和し，それ以降は顕著な増減が見られない．チャンク分割作業では，作業記憶上の制約がある同時通訳の状況を想定して境界を定める作業としたため，ある程度の長さでチャンク長が飽和することは直感的な結果である．また，これらのチャンク長は通訳者による同時通訳の遅延の上限に相当すると考えられ，本コーパスから構築したシステムで同時翻訳を行う際も，同程度の遅延に収まると期待できる．

7)　発話数が 5 件以上ある発話長のみ対象とした．BSD は短い発話が多く，BSDja では最大発話長 25，BSDen では 42 であった．他のドメインの発話長 50 以上のデータはグラフから割愛したが，50 以下と比べて特に傾向の変化は見られなかった．

表 6　チャンク境界の作業者間一致率（日英 20 通訳者ペア，英日 19 通訳者ペアの F1 値）

ja-en	76.3	73.0	71.7	71.0	64.8	64.5	64.0
	59.1	58.8	58.4	56.5	56.4	53.2	47.5
	46.8	43.5	43.4	34.0	31.5	15.7	
en-ja	75.7	71.0	68.8	67.6	65.7	64.0	61.5
	59.9	59.4	53.9	53.0	49.9	49.0	49.0
	48.9	46.1	41.5	40.1	33.5		

4.2　チャンク境界の一致率

本コーパスの一部の原発話（日本語 2,755 発話，英語 3,160 発話）について，通訳者 3 名によるチャンク分割・訳文作成を実施した．これを，同一の原発話集合（50 発話以上）を作業した通訳者 2 名のペアごとのデータに分け，各ペアのチャンク境界の一致率を算出した[8)]．ペアの一方の境界を正解とした際の他方の境界についての F1 値を求め，双方向的な一致率の尺度として用いる．日英 20 ペア，英日 19 ペアについての F1 値を表 6 に左右方向に降順で示した．日英・英日合計 39 ペアのうち，約 7 割の 28 ペアについては F1 値 48%以上となり，ペアの一人目から二人目の境界を見た場合と，その逆の場合の両方向において，境界の概ね半数が被覆される状況と解釈できる．残り約 3 割のペアはそれを下回る低い一致率であった．

F1 値の高低と，言語方向，原データ，作業発話数との関連は特に見られなかった[9)]．一方，表 6 で下線で示した F1 値は，日英，英日それぞれ作業者 ID 01，63 を含むペアに関するもので，両作業者は他の作業者と異なる位置で分割する傾向が大きいことが示唆され[10)]，ペアを構成する作業者の影響が大きいことを確認した．コーパス中に非典型的な境界／非境界が混在していると，同時翻訳システムのチャンク分割に関する安定性を損なうことが懸念されるため，それらを検出し，学習時の影響を低減するよう

8)　空白・句読点の前後に挿入された境界は同一とみなした．
9)　原データ及び作業発話数の情報は付録 §A.1 に記す．
10)　詳細は付録 §A.1 に示す．

表 7　日英翻訳の 2 段階訳に関する統計情報

	N_{utter}	$N_{\text{word}}/N_{\text{utter}}$			CR [%]
		U	ST1	ST2	
All	62.5k	21.6	16.7	15.0	89.55
Diff	30.2k	31.9	24.5	20.9	85.28
NoDiff	32.3k	11.9	9.4	9.4	–

表 8　英日翻訳の 2 段階訳に関する統計情報

	N_{utter}	$N_{\text{word}}/N_{\text{utter}}$			CR [%]
		U	ST1	ST2	
All	109.6k	10.1	13.2	11.5	87.28
Diff	44.4k	14.0	19.0	14.8	78.22
NoDiff	65.2k	7.3	9.2	9.2	–

な方法が有効である可能性がある．

4.3　省略の傾向

§4.1 と同一の原発話データとその訳文データを用いて，ST1 と ST2 の 2 段階の訳文の差について分析を行った．2 種類の訳文とそれらの差分に関する統計情報を表 7 及び表 8 に示す．

全発話（All）のうち，ST1 と ST2 で文字列上の差分があった発話（Diff）は日英で約 48%，英日で約 41%であり，省略または他の何らかの編集が行われていることがわかる．差分なしの発話（NoDiff）では発話あたり単語数 $N_{\text{word}}/N_{\text{utter}}$ が少なく，省略の余地が小さかったことが窺える．実際，短い発話に対して省略を行う必要性は低い．差分ありの発話に限ると，英日で 78%，日英で 85%の圧縮率 CR（ST2 の単語数合計 ÷ ST1 の単語数合計）であり，システムによる省略処理を行う際の圧縮率の目安ともみなせる．作業指示（60〜80%以上）に対してはやや高めの割合での訳出となった．

差分の大きさを定量化した指標として，ST2 を参照訳としたときの ST1 の Translation Edit Rate（TER）[14] を `tercom` [11] を用いて算出したところ，表 9 に示すように日英・英日とも全発話（All）に関して 13%程度であった．また，同ツールにより各編集操作の適用回数も算出した．作業指示として「抽象的な表現への言い換えではなく，主に語句を削ることで省略を行うこと」[12] を推奨したことの表れとして，削除 Del が 80〜90%以上を占めたものの，他の編集操作（置換 Sub，挿入 Ins，シフト Shf）も検出された．文法的に妥当な文とするために削除以外の編集が必要であった可能性などが考えられる．

削除された単語は，日本語訳では助詞，代名詞，

表 9　ST2 に対する ST1 の TER 及び編集操作内訳

		TER	Del	Sub	Ins	Shf
ja-en	All	13.70	184,677	10,629	1,061	1,377
	Diff	23.46	(93.4%)	(5.4%)	(0.5%)	(0.7%)
en-ja	All	13.43	113,293	19,708	4,264	2,933
	Diff	18.44	(80.8%)	(14.1%)	(3.0%)	(2.1%)

名詞など，英語訳では verb，adverb，noun などが多かった．詳細は付録 §A.2 に示す．

5　関連研究

同時通訳者により原発話テキストにチャンク境界を付与した研究に，丁ら [15]，清水ら [16] の研究がある．丁ら [15] は，SIDB の日本語対話文に，英語同時通訳文を参照しながらチャンク境界を付与し，自動付与した節境界との関係を分析した．節境界のうちチャンク境界であるものは 51%，チャンク境界のうち節境界であるものは 75%と，両者には，ずれがあることを報告している．清水ら [16] は，話し言葉の自動分割を目的とし，日本語講演文にチャンクに相当する「音声翻訳単位」境界を付与した．作業者 3 名の多数決で正解境界を定め，作業者の F 値を平均 0.9 以上と報告している．

同時通訳における欠落／脱落は，高負荷状況での訳出の失敗として生じる [17] 場合と，時間的制約の克服のための意図的な省略によって生じる場合がある．蔡ら [18] は，訳出の失敗として起こる語の欠落に注目し，SIDB を用いて，講演者話速の速さや訳出遅延時間の長さに応じて欠落率が高くなることや，副詞の欠落率が高いことを示した．遠山ら [19] は，訳出の発話量を減らす方略として短縮による訳出が行われることを挙げ，英語原発話中の「A think B」の主語などが省かれて「〜と思います」や「〜なんでしょう」と訳出される SIDB 中の事例を挙げた．

6　おわりに

本稿では，GCP 同時通訳コーパスについて報告した．コーパスの分析結果として，(1) 長い原発話に対してもチャンク長（訳出遅延）は一定程度で収まること，(2) チャンク境界の作業者間一致率の高低は作業者の組合せに依存すること，(3) 原発話の内容全体を訳した訳文について，その 10〜20%前後の語句が省略可能な非重要情報と判断されていること（圧縮率 80〜90%前後）を示した．今後，本コーパスを用いた同時翻訳システムの構築と性能評価を行う予定である．

11）`https://www.cs.umd.edu/~snover/tercom/`
12）簡易なモデルによる省略処理の実現や，省略されやすい表現の分析を容易にすることを意図してこのような指示とした．

謝辞

本件は，総務省の「ICT 重点技術の研究開発プロジェクト（JPMI00316）」における「多言語翻訳技術の高度化に関する研究開発」による委託を受けて実施した研究開発による成果である．

本コーパスの作成に当たり，日本語話し言葉コーパス，Business Scene Dialogue Corpus，CIAIR 同時通訳データベース，英日・日英通訳データベース，AMI Meeting Parallel Corpus，IWSLT 2017 Evaluation Campaign データセットを利用した．

参考文献

[1] Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Haifeng Wang, Liang Huang, Qun Liu, Julia Ive, and Wolfgang Macherey. Findings of the third workshop on automatic simultaneous translation. In Proceedings of the Third Workshop on Automatic Simultaneous Translation, pp. 1–11, Online, July 2022. Association for Computational Linguistics.

[2] Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. Findings of the IWSLT 2022 evaluation campaign. In Proceedings of the 19th International Conference on Spoken Language Translation, pp. 98–157, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics.

[3] 総務省.「グローバルコミュニケーション計画 2025」の公表,（2023-01-12 閲覧）. https://www.soumu.go.jp/menu_news/s-news/01tsushin03_02000298.html.

[4] Hitomi Tohyama, Shigeki Matsubara, Koichiro Ryu, Nobuo Kawaguch, and Yasuyoshi Inagaki. CIAIR simultaneous interpretation corpus. In Proceedings of Oriental COCOSDA 2004, 2004.

[5] 松下佳世, 山田優, 石塚浩之. 英日・日英通訳データベース（JNPC コーパス）の概要. 通訳翻訳研究への招待, No. 22, pp. 87–94, 2020.

[6] Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. Large-scale English-Japanese simultaneous interpretation corpus: Construction and analyses with sentence-aligned data. In Proceedings of the 18th International Conference on Spoken Language Translation, pp. 226–235, Bangkok, Thailand (online), August 2021. Association for Computational Linguistics.

[7] 小松達也. 同時通訳. 応用言語学事典, pp. 396–397. 研究社, 2003.

[8] Kikuo Maekawa. Corpus of spontaneous Japanese: Its design and evaluation. In ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.

[9] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the business conversation corpus. In Proceedings of the 6th Workshop on Asian Translation, pp. 54–61, Hong Kong, China, November 2019. Association for Computational Linguistics.

[10] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Document-aligned Japanese-English conversation parallel corpus. In Proceedings of the Fifth Conference on Machine Translation, pp. 639–645, Online, November 2020. Association for Computational Linguistics.

[11] Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. Overview of the IWSLT 2017 evaluation campaign. In Proceedings of the 14th International Conference on Spoken Language Translation, pp. 2–14, Tokyo, Japan, December 14-15 2017. International Workshop on Spoken Language Translation.

[12] 伝康晴. 多様な目的に適した形態素解析システム用電子化辞書 (<特集>日本語コーパス). 人工知能, Vol. 24, No. 5, pp. 640–646, 2009.

[13] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[14] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, pp. 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas.

[15] 丁, 笠浩一朗, 松原茂樹, 吉川正俊. 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 − . 言語処理学会第 12 回年次大会, 2006.

[16] 清水徹, 中村哲, 河原達也. 同時通訳者の知識と韻律情報を用いた講演文章のチャンキング. 情報処理学会研究報告音声言語情報処理, Vol. 2008-SLP-72, No. 68, pp. 81–86, July 2008.

[17] 水野的. 同時通訳の理論 認知的制約と訳出方略. 朝日出版社, 2015.

[18] 蔡仲熙, 笠浩一朗, 松原茂樹. 同時通訳における語の欠落に影響を及ぼす要因の分析. 通訳翻訳研究, No. 18, pp. 13–18, 2018.

[19] 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. 133–146, 2003.

A 付録

A.1 チャンク境界の一致率の詳細評価

作業者間チャンク境界一致率の詳細を述べる．通訳者ペアごとの一致率を，後述する MPR でも算出し，F1 値（表 6 と同一）とともに表 10 に示した．「X-P1-P2」の形式のペア ID は，作業データ X（表 2 及び表 3 に示したドメイン名の 1 文字目）に対する通訳者 P1 と通訳者 P2 による作業結果であることを意味し，N_{utter} は 2 名が作業した発話数を表す．

MPR について，ペアの一方の境界を正解とした際の他方の適合率 P 及び再現率 R のうち値の大きい方 $\max(P, R)$ で定義する．MPR は，より粗く（細かく）分割する作業者の境界を正解とした際の細かく（粗く）分割する作業者の境界の再現率（適合率）と解釈でき，一方から見た他方への一方向的な一致率または被覆率とみなせる．日英・英日合計 39 ペアに関して，最も低い場合でも MPR 56%で過半数の境界が一致し，約 8 割の 32 ペアについては MPR 70%を超え，F1 値と比べて高い値となった．

また，表 6 と表 10 で下線で示した作業者 01 と 63 を含むペアについて，低い F1 値となった要因を分析した．ペア内 2 名の分割回数の比 SR（分割回数が少ない作業者の回数を分子とする）を算出すると，全 39 ペアの平均 SR = 0.59 に対し，作業者 01 を含む 6 ペアでは SR = 0.14〜0.64 で分布し，作業者 63 を含む 4 ペアでは SR = 0.28〜0.67 で分布し，作業者 01 または 63 を含む 10 ペア中 7 ペアが SR 下位 25%に位置した．つまり，作業者 01 はペアの他方より少なく分割し，作業者 63 はペアの他方より多く分割する特徴が見られた．

A.2 2 段階の訳文で省略された表現の詳細

2 段階の訳文 ST1 と ST2 の差分のうち，削除 Del として検出された表現について，1 単語単位及び連続する複数単語単位で集計した．頻度上位 10 件の品詞（列）を，全品詞（列）の合計頻度に対する頻度の割合（Rate，%）とともに，表 11 及び表 12 に示す（「代」は代名詞，「助」は助詞，「副」は副詞，「感」は「感動詞」，「接」は接続詞，「名」は名詞，「形」は形容詞を表す）．複数単語単位（POS Seq）については，頻度上位 2〜3 件の表層の例も示した．

頻繁に削除された表現には，談話標識にあたるような接続詞，副詞，感動詞（日本語訳・英語訳共通），「主語にあたる代名詞＋動詞」（英語訳），「代名詞＋助詞」（日本語訳）などがあった．これらは，作業指示の結果，削除しやすい表現として選択された可能性があり，任意の編集による圧縮を許容した場合には，異なる傾向となることが考えられる．

表 10 チャンク境界の作業者間一致率（MPR，F1 値）

ja-en				en-ja			
Pair ID	N_{utter}	MPR	F1	Pair ID	N_{utter}	MPR	F1
S-18-33	64	86.8	76.3	T-16-30	176	76.3	75.7
B-54-60	343	78.2	73.0	B-20-29	1,020	76.7	71.0
C-54-58	104	71.7	71.7	T-16-18	155	89.4	68.8
S-08-33	64	81.2	71.0	B-10-29	988	74.1	67.6
S-08-17	104	79.9	64.8	T-09-30	392	73.3	65.7
B-54-66	343	77.7	64.5	B-10-20	988	76.7	64.0
B-68-20	764	72.5	64.0	T-18-28	412	72.4	61.5
B-66-20	481	82.3	59.1	T-25-30	214	71.6	59.9
B-60-66	343	76.9	58.8	T-18-30	257	92.0	59.4
J-02-03	789	82.3	58.4	A-66-60	452	60.8	53.9
S-08-18	168	86.9	56.5	T-09-28	392	77.4	53.0
S-17-18	104	77.3	56.4	T-25-28	214	72.4	49.9
S-01-06	176	78.8	53.2	T-16-28	331	87.9	49.0
C-58-66	104	68.2	47.5	W-65-63	474	61.2	49.0
C-54-66	104	67.1	46.8	T-28-30	1,039	87.3	48.9
B-01-18	764	56.0	43.5	W-66-65	474	74.7	46.1
J-01-03	789	80.1	43.4	A-70-63	452	79.0	41.5
B-01-20	1,245	59.6	34.0	A-66-63	452	90.8	40.1
J-01-02	789	93.1	31.5	W-66-63	474	73.0	33.5
B-01-16	481	65.7	15.7	–	–	–	–

表 11 日英翻訳の 2 段階訳において削除された表現

Rate	POS	Rate	POS Seq	Examples
15.8	verb	15.8	adv	so, then, also
14.2	adv	5.1	cconj	and, but, so
13.9	noun	4.7	det	the, a, some
13.7	adp	4.6	noun	today, people, section
12.3	pron	3.8	pron_verb	i_think, it_'s, it_is
11.7	det	3.2	adj	right, such, various
5.4	adj	3.0	verb	is, have, are
4.9	cconj	3.0	adp	that, in, of
2.4	intj	2.8	intj	yes, well, please
1.7	propn	2.6	pron	it, i, what

表 12 英日翻訳の 2 段階訳において削除された表現

Rate	POS	Rate	POS Seq	Examples
35.0	助詞	15.3	代_助	私_は, それ_は, あなた_は
16.0	代名詞	7.2	副	つまり, とても, もし
10.5	名詞	5.1	感	はい, ああ, ええ
7.8	助動詞	4.2	代_接_助	私_たち_(は|が|の)
6.6	動詞	3.9	接	そして, しかし, また
4.7	感動詞	3.7	名_助	本当_に, 実際_に, 実_は
4.5	接尾辞	2.7	名	実際, オッケー, 今
2.8	接続詞	2.2	助	の, を, は
1.7	形容詞	1.3	助_助	で_は, で_も, と_か
1.5	連体詞	1.2	形_助	ええ_と, 少なく_とも