記者会見通訳の二言語並行コーパスの構築

山田優[1] 松下佳世[2] 石塚浩之[3] 歳岡冴香[4] Michael Carl[5]

1 関西大学 2 国際基督教大学 3 広島修道大学 4 大阪大学 5Copenhagen Business School

1 はじめに

本研究プロジェクトは，我が国の通訳翻訳研究の活性化を目指して，研究者が広く利用可能な日英の通訳の対訳コーパスを構築することを目的としたものである．日本には翻訳テクストからなる単一言語コーパスや外国語学習者向けの英語コーパスは多数存在するが，原文と訳文が対応した「同時通訳・逐次通訳」の対訳コーパスの数は多くない．名古屋大統合音響情報研究拠点(CIAIR)の実験的環境下で記録した同時通訳コーパスは存在したが，自然発話，すなわちプロ通訳者による実現場でのパフォーマンスを記録した対訳データは無い．本研究では，公益財団法人・日本記者クラブによる実際の記者会見における原発話と通訳者の訳出を，音声とその波形，文字情報を組み合わせた形でデータベース化する対訳コーパスを構築し，通訳翻訳研究，メディア研究，自然言語処理研究を横断する学際的な研究や，脳科学・通訳の認知プロセスに関わる応用研究の実施を目指している．

2 背景

もともと，欧州で二十世紀半ば以降に発展を遂げた通訳研究は，通訳行為そのものの即時性と密接に関係した営為であり，通訳者の認知プロセスや訳出プロセスをめぐる研究が盛んであった．通訳プロセスをモデル化し，原発話から訳出開始までのタイムラグを EVS(Ear-Voice Span)として計測して，原発話を心的に処理する際の意味・統語的単位などを解明する手がかりとしてきた．しかし，通訳のパフォーマンスを大量に記録しテクスト化するのは困難であったため，これまで大規模な自然発話のコーパス(authentic corpus)は，10 年ほど前から構築されている欧州議会の通訳コーパス(EPIC: European Parliament Interpreting Corpus)以外に見られない．

翻訳研究においては英マンチェスター大学のモナ・ベーカー教授による「TEC(The Translational English Corpus)」等が有名であるが，これらは必ずしも対訳コーパスではなく，多言語から英語に訳されたテクストからなる単一言語コーパスであった．近年は，統計的機械翻訳や翻訳メモリなどの普及と必要性に迫られて，原文と訳文との対訳コーパスが広く作成され，研究でも多く用いられているが，翻訳研究における関心は静的な「訳文」(Product)の分析から動的な「訳出プロセス」(Process)へと移行しつつあり，翻訳テクスト分析だけでなく，どのようにしてその訳文にたどり着いたのかという，翻訳者の頭の中で起きている認知的・心的プロセスの解明が盛んに進められている．例えば，翻訳者が特定の原文を目で見てから，対訳を入力し始めるまでのタイムラグ(Eye-Key Span)をアイトラッキングやキーボードストロークのデータから測定し，先の訳出単位や作動記憶容量の分析が行われている[1].

このように大規模コーパスが存在する翻訳研究に対して，コーパスが少ない通訳研究の状況は立ち遅れている．この要因の一つは，音声データが主となる通訳の対訳コーパスの構築が，技術的にも，工数的にも，また予算的にも容易でなかったことである．本研究では，このようなハードルを払拭すべく，日本だけでなく世界的な研究に広く活用できる公開型の大規模通訳コーパスを構築することを目指す．

3 データ

本研究でコーパス化するデータは，公益財団法人・日本記者クラブにより YouTube 上の「日本記者クラブチャンネル」(https://www.youtube.com/user/jnpc)で公開されている通訳音声(英語・日本語)を含む映像素材である．首脳演説や記者会見といった開かれた場におけるデータのうち，通訳付きの会見を行っているものが対象となる．

同チャンネルでは過去 5 年分相当の通訳音声が含まれた会見映像が公開されており，そのうち，英語通訳を介して行われた 270 件ほどの映像・音声データが今回の元データとなる．会見は平均 1 時間程度であるので，合計 270 時間相当となる．また毎週数件の会見が新たに行われ，データが蓄積され続けている．会見の通訳は現場の一線で活躍するプロ通訳者による．すでに研究用のコーパス構築に向けた覚書を日本記者クラブと締結し著作権者の了解を得ている．

4 コーパス構築

本コーパスの現在の内容と仕様は次の通りである．

4.1. コーパスの内容

コーパスには，動画，音声（英語: L チャンネル，日本語: R チャンネル），トランスクリプション（原文・訳文の書き起こしテクスト）が含まれる．それに加えて，書き起こしたテクストを音声波形に載せたデータと，原言語と目標言語を単語レベルで対応付けしたデータを提供する．

4.2. コーパス構築の流れと仕様

コーパス構築の流れは，以下の通りである．

① テクストの書き起こし

原発話と訳出の音声テクストの書き起こしを外注業者なども活用し効率的に行う．書き起こしの際は，「あー」「えー」などの躊躇・フィラーなども 500msec 以上のものは，明示的に書き起こすことを基本的な取り決めとした．

② 書き起こしテクストの音声波形への搭載

フリーソフトウェア ELAN（下図）を使用し，書き起こしたテクストデータを音声波形に載せる．英語は単語，日本語は分節を最小単位として波形に合わせてゆく．フィラー等については，[ ]で囲んで簡易アノテーション処理をしておき，後にフィラーのみを容易に削除できるよう施した．

ELAN 入力を終えると，csv などの形式でファイルをエクスポート出来る(図 2)．ここには時間情報と原発言・訳出のテクストが含まれる．この段階では，原発話（原言語）と訳出（目的言語）は対応付されていないので便宜的に時間経過ベースで対応付けをし，セグメント番号を付加することによって，大まかアラインメントを施す．

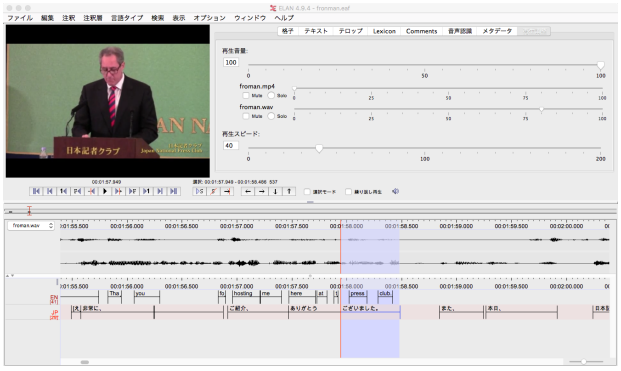

図 1: ELAN での作業の様子

例えば，下図において，英語の原発話（En-an）は，01:50.5（1 分 50 秒 5）から，始めのフィラー[Well,] [ah]から，“good afternoon and…” が発話される．この原発話に対応する訳出（JP-an）は，01:53.6 から，「皆さま方，こんにちは…」と始まる．これらを対応セグメントと見做し，共通セグメント番号 seg#11 を付加しグループ化する（オレンジで色分けしたセル）．これだけの情報でも，同時通訳における原発話から訳出開始までの遅延時間（EVS）を明らかにするなど，一定の研究に資する．

		start	end	duration	EN-an	JP-an	seg#	
140	EN	01:50.5	01:50.7	00:00.2	[Well,]		11	
141	EN	01:51.0	01:51.1	00:00.1	[ah]		11	
142	EN	01:51.2	01:51.3	00:00.1	good		11	
143	EN	01:51.3	01:51.8	00:00.5	afternoon		11	
144	EN	01:51.8	01:51.9	00:00.1	and		11	
145	EN	01:51.9	01:52.5	00:00.6	arigatou		11	
146	EN	01:52.5	01:53.1	00:00.6	gozaimasu.		11	
2853	JP	01:52.9	01:53.6	00:00.7		皆さま方、		11
2854	JP	01:53.6	01:54.2	00:00.6		こんにちは。		11
147	EN	01:54.0	01:54.2	00:00.2	Thank		12	
148	EN	01:54.2	01:54.3	00:00.1	you		12	
149	EN	01:54.4	01:54.4	00:00.0	for		12	
2855	JP	01:54.4	01:55.3	00:01.0		ありがとうございます。		12
150	EN	01:54.4	01:54.6	00:00.2	that		12	
151	EN	01:54.8	01:55.0	00:00.2	kind		12	
152	EN	01:55.1	01:55.7	00:00.6	introduction		12	
2856	JP	01:55.5	01:55.6	00:00.1		[え、]		12
2857	JP	01:55.6	01:56.2	00:00.7		非常に、		12
153	EN	01:55.7	01:55.7	00:00.1	and		12	
154	EN	01:55.8	01:56.0	00:00.2	thank		12	
155	EN	01:56.0	01:56.3	00:00.3	you		12	
2858	JP	01:56.2	01:56.8	00:00.6		ご親切な		12
156	EN	01:56.8	01:56.9	00:00.1	for		12	
2859	JP	01:56.8	01:57.4	00:00.6		ご紹介、		12
157	EN	01:56.9	01:57.3	00:00.3	hosting		12	
158	EN	01:57.3	01:57.4	00:00.1	me		12	
2860	JP	01:57.4	01:58.5	00:01.1		ありがとうございました。		12
159	EN	01:57.5	01:57.7	00:00.3	here		12	
160	EN	01:57.8	01:57.8	00:00.1	at		12	
161	EN	01:57.8	01:57.9	00:00.1	the		12	
162	EN	01:57.9	01:58.2	00:00.3	press		12	
163	EN	01:58.3	01:58.4	00:00.2	club.		12	
2861	JP	01:58.8	01:58.9	00:00.0		また、		12
2862	JP	01:58.9	01:59.2	00:00.4		また、		12
2863	JP	01:59.2	01:59.9	00:00.7		本日、		12
164	EN	01:59.7	01:59.7	00:00.0	I		13	
165	EN	01:59.7	01:59.8	00:00.1	have		13	
166	EN	01:59.9	01:59.9	00:00.1	been		13	
167	EN	01:59.9	02:00.1	00:00.1	to		13	

図 2 ：ELAN からエクスポートしたファイル

③ アラインメントから分析データ生成

この後，上のファイルを変換し CRITT TPR-DB[1]のプラットフォーム[2]に移管し，付属のツール（YAWAT）を使って原発言と訳出をさらに細かい言語単位でアラインメント（対応付け）していく．アラインメントは，単語やフレーズ単位で行う．

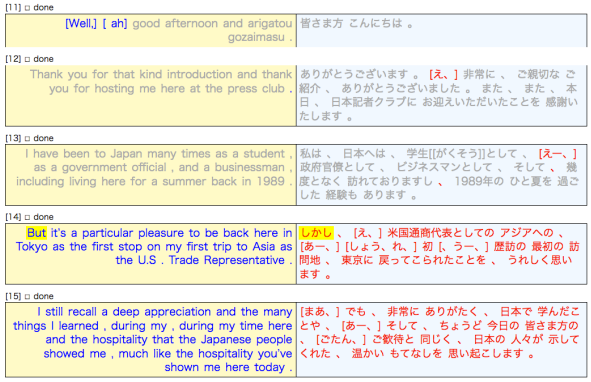

図 3：YAWAT でのアラインメント作業

CRITT TPR-DB は元々，翻訳者の訳出プロセスデータを収集・分析するためのツールで，翻訳者のアイトラッキングデータやキーボードストロークログを時間経過に合わせて記録するものだ．これにより，例えば，翻訳者が翻訳を開始してから何分何秒何 msec 後に，原文のどの単語を読んでいて(視線の位置情報)，何を訳しているのか(キーボード情報)を同定することができる．更に詳細な研究対象項目を指定すれば，分析用のテーブル(ファイル)を生成することができる．

[1] The Center for Research and Innovation in Translation and Translation Technology Translation Process Database

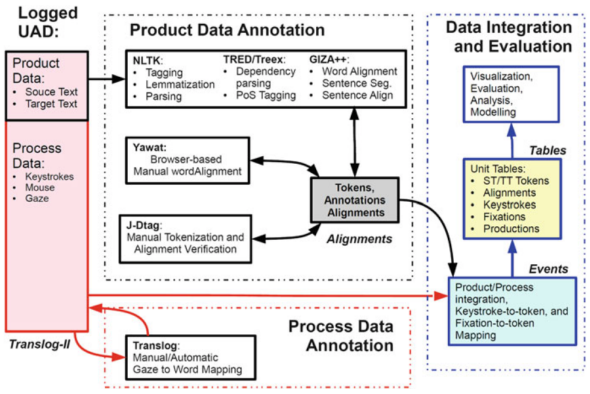

図 4：CRITT TPR-DB ツールの概要

作業の流れを図 4 に即して説明する。図左の Logged UAD（User Activity Data）工程は，翻訳者の訳出プロセスデータを収集する段階になる．本研究では，ここに通訳データを使用する．収集データに対して，図中央の Product Data Annotation 段階で，POS 情報，係り受け情報，形態素解析などのアノテーション付与を行い,YAWAT で手動の原文・訳文のアラインメントを行う．アノテーション処理が完了すれば，図右の Data Integration and Evaluation 段階で，時間情報を含む生データ UAD と Annotation Data を統合し，分析・評価用のテーブルを生成する準備が整う．生成したテーブルは R-Studio®など用いてグラフ化・可視化したりもできる．

CRITT TPR-DB ツールを使った収集と分析は，これまでに多くの実績がある．中国語，デンマーク語，ドイツ語，ヒンディー語，スペイン語，英語，日本語のいずれかの組合せで 150 人以上の翻訳者，計 110 時間以上の翻訳プロセスデータを収集してきた．2015-16 年にも英語→日本語の翻訳者 36 名分の翻訳，ポストエディット，音声認識翻訳の3パターンのプロセスデータ収集・分析を行った実績がある[2]．これらのデータ(ENJA15)および多言語データはすべて Creative Commons ライセンスで一般公開されている．

本研究では翻訳ではなく，同時通訳データを分析対象として使用する．上述した②の ELAN でタグ付けした音声データとトランスクリプションを，UAD 生データとして CRITT TPR-DB ツールにインポートする．翻訳プロセスも同時通訳も，時間経過と共に訳出が進行するプロセスであるので，共通のツールを活用できる．本研究の通訳コーパスも一般公開予定である．これまでの研究蓄積と比較研究できる点でも意義がある．

5 コーパスを使った分析例

同時通訳の難しさは，原発話を聞きながら訳出を行わなければならない同時性にある．通訳理論研究者 Gile[3]の努力モデル（Effort Model）よれば，同時通訳（SI = simultaneous interpreting）という作業は，主に 3 つの作業で構成される．内容は，L: リスニング，P: 訳出，M: 記憶である．それぞれのタスクに割り振られる通訳者の努力（effort）を C: 調整する必要がある．式で示すと，SI = L + P + M (+ C) となる．

原文を聞いてから，訳出する（L＋P）ためには，脳内の作動記憶で L の内容を保持する（M）が必要だ．M への負担を極小化するために，同時通訳で順送り訳という方略がとられることが多い．原言語で聞いた内容を，できるだけ早く目標言語に訳出する漸進的（incremental）訳出である．

例えば，下の原文の①I'm almost positive の部分を 2 通りの訳出方略で訳すと，〈逆送り】では，①に対応する確かだが最後に訳されているのに対して，〈順送り】では最初に訳される．〈逆送り】だと，①を記憶保持しながら残部分②のリスニング／訳出作業を続けることになるので，M への負担が〈順送り〉よりも増大する．

原文
I'm almost positive ① I've heard those names. ②
〈逆送り〉
こういう名前を聞いたことがあるのは ② 確かだ ①.
〈順送り〉
確か, ① こういう名前を聞いたことがある ②.

理論上、〈順送り〉は同時通訳で多く使われる方略であり，他方〈逆送り〉は，翻訳（書き言葉の翻訳）で多く見られると考えられる．この違いを定量的に分析するには，原文と訳文の統語的な交差量（crossover）を見ることで算出できる．上の場合だと，〈逆送り〉と〈順送り〉の比較では，〈順送り〉のほうが統語的交差量は少なくなると予測される．すなわち翻訳と同時通訳で，同じ原文を訳出した時でも，統語的交差量は，同時通訳のほうが小さくなるのである．

このような分析も，詳細なアノテーションを付加する CRITT TPR-DB ツールで可能になろう．2017 年 1 月現在では，まだ分析用の通訳データが揃っていないが，類似分析を 2015−16 年に行っているので[4]，その暫定結果を記しておく．

英日翻訳プロセス実験で，キーボード入力を使って普通に翻訳する「T モード」と音声認識ソフトを使って口頭翻訳する「D モード（translation dictation）」を比較した．D モードは，同時通訳とは異なるが，視訳（サイトトランスレーション）に似ており，同時通訳の訓練などでも広く活用される訳出法である．同じ原文を異なるモードで翻訳した結果の統語的交差量を比較した．

結果として、統計的有意差こそ無かったが，予測した通

り D モードでは統語交差量は T モードよりも少なかった. 〈順送り〉方略的な訳出が高頻度で使用されと推察される.

このように, 本プロジェクトのコーパスが完成すると, 既存の翻訳データとの比較や,より詳細な分析ができると期待される.

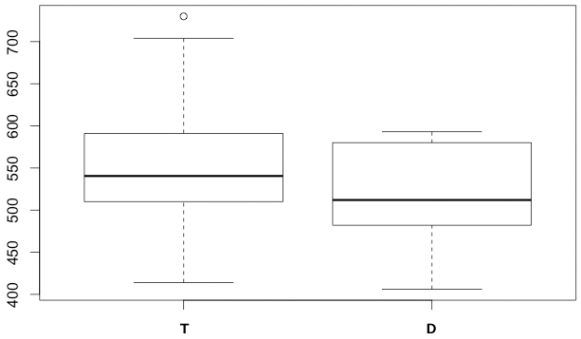

図 5：T と D モードの統語的交差量の比較

6 応用研究・使用可能性

現時点で想定可能な研究・使用への応用としては, 以下の例が挙げられる.

1. 自然言語処理の資源としての活用

SMT/NMT の資源としての活用はもちろん, 音声通訳機・漸進的翻訳機などの基礎研究の資料として活用できる.

2. 通訳の認知プロセスの明示化

同時通訳記録をディスコース処理の時系列に沿って分析することにより, 通訳者の発話理解の展開を跡付けることが可能となる. これにより, 通訳の認知プロセスの明示化に寄与する. ここから得られた知見は一般の発話理解のモデルに応用できる.

3. 通訳・翻訳におけるレトリックの訳出処理に関する言語学的研究

通訳者・翻訳者は多くの場合, 起点言語の修辞表現を文字通りには目標言語に置き換えない. この理由については翻訳テクストを対象とし研究が進められてきたが, 通訳コーパスを分析することで, レトリックの訳出処理についての新たな知見に基づく研究が可能となる.

4. 報道現場における通訳のメディア研究的考察

本コーパスを活用することで, これまで十分な学術的研究が進んでこなかった国際報道の現場における通訳翻訳行為(ニュース・トランスレーション)についての研究を進めることができる. 具体的には, 訳出遅延とリスクの相関分析や, 訳出と報道された記事の比較分析が考えられる.

5. コーパスベース翻訳研究・翻訳プロセス研究との比

先に示した通り, コーパスベース翻訳研究のうち, 翻訳の訳出プロセスと通訳プロセスとを比較し, 人間の「訳出プロセス」の解明に手がかりとなる.

6. 教育現場での活用

大学など通訳教育の現場で, コーパスから実際の通訳者の訳出例を紹介し, 学生の指導に活用する. 通訳研究で卒業研究や学位論文の作成を希望する学部生・大学院生には, 分析資料としてこのコーパスの活用を許可する. このコーパス構築を契機として, 多様な研究分野を背景とする実証的な通訳研究が加速し, 関連する研究領域にも広範な貢献を果たすことが期待される.

7 まとめ

本稿では, 記者会見通訳の二言語並行コーパスについて, 背景, 設計, 収集, 構築, 研究例, さらに, 利用可能性について述べた. 本コーパスの作成目的は, 通訳研究の向上と, 通訳教育への貢献, そして分野を横断して, 自然言語処理, 認知科学, 脳科学など多面的な活用を目論んでいる. 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う予定である. そして広範な研究分野で多面的に活用を促し, 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.

謝辞

本研究は, JSPS 科研費 16H02915 の助成を受けている.

参考文献

[1] Gopferich, S., Jakobsen, A. & Mees, I. (2008) Looking at eyes: Eye-tracking studies of reading and translation process. Copenhagen: Copenhagen Business School.

[2] Carl, M., Lacruz, I., Yamada, M., Aizawa, A. (2016). Comparing spoken and written translation with post-editing in the ENJA15 English → Japanese Translation Corpus. 言語処理学会第 22 回年次大会 (NLP2016). pp. 1209 - 1212.

[3] Gile, D. (1995). Basic concepts and models for interpreter and translator training. Amsterdam/Philadelphia: John Benjamins.

[4] Yamada, M., Carl, M. (2016). An Investigation into the Efficiency of Translation Dictation. TAUS Executive Forum in Tokyo 2016. 2016 年 4 月 19 日.