〈報告〉 JAITS

英日・日英通訳データベース（JNPC コーパス）の概要

松下佳世[1]　山田優[2]　石塚浩之[3]

（[1]立教大学 [2]関西大学 [3]広島修道大学）

1. はじめに

英日・日英通訳データベース（JNPC コーパス）は、主に通訳翻訳分野の研究と教育のために構築された日英通訳の対訳コーパスである。公益財団法人・日本記者クラブの合意のもと、同法人が YouTube に公開する英日通訳付きの記者会見動画から、原発話と通訳者の訳出を書き起こしてコーパスを作成した。文字情報とオリジナルの動画、音声などは「ELAN」というフリーソフトウェアで視聴・閲覧できるようになっている。収録データは記者会見計約 77 時間分で、特定非営利活動法人・言語資源協会（GSK）を通じて入手できる（https://www.gsk.or.jp/catalog/gsk2020-a/）。本稿では、その詳細を説明する。尚、今後 JNPC コーパスを利用して研究を行う際は、本稿を引用して頂きたい。

2. JNPC コーパスの内容

2.1 オリジナルデータ

2020 年 4 月の公開時点で、JNPC コーパスには、2010 年から 2017 年の間に日本記者クラブで行われた記者会見のうち、79 件分のデータが収録されている。内訳は、同時通訳付きのものが 71 件、逐次通訳付きのものが 8 件である。会見は最短 17 分 35 秒、最長 2 時間 7 分 12 秒で、平均時間は 58 分 21 秒である。

会見は冒頭のスピーチ部分と、その後の質疑応答からなる。収録分に関しては、英語のスピーチが 75 件、日本語のスピーチが 4 件。基本的に同時通訳の場合は 2 名、逐次通訳は 1 名の通訳者が担当している。スピーチ部分は英日または日英一方向の訳

MATSUSHITA Kayo, YAMADA Masaru, ISHIZUKA Hiroyuki, “An Overview of the Japan National Press Club (JNPC) Interpreting Corpus,” *Invitation to Interpreting and Translation Studies*, No.22, 2020. pages 87-94.

出のみで、質疑応答部分に関しては概ね双方向の訳出が行われている。

2.2 データ形式

英語と日本語の文字データは、フリーソフトウェア「ELAN」(https://archive.mpi.nl/tla/elan) を用い、「.eaf」というファイル形式で収録されている。同ソフトフェアを用いることで、コーパスに同梱して提供する動画(.mp4) と音声(.wav)ファイルを同期させ、下図のように表示することができる。音声はステレオで、英語と日本語は独立したチャンネルに収録されている。書き起こしのテクストは、音声波形の下に、英語（en）と日本語（ja）の異なる注釈層で表示される。

図 1: ELAN によりコーパスを開いた様子

2.3 データの構築方法

書き起こしのテクストの構築は、英語と日本語ともに音声認識ソフト（speech to text）にかけて自動生成した。使用したシステムは IBM Watson Speech to Text と Speechmatix の 2 種類。自動生成した後に、誤認識によるエラーを、人手によって修正した。3 名の異なる作業者が 3 段階の確認作業を行ったが、基本的には自動生成テクストをできるだけ活かす方針としたため、2 種類の音声認識システムの記述の不統一や表記揺れ等が残っている。このため、利用の際には、研究の目的に合わせ、利用者自身による修正や調整が必要となる。

トランスクリプトには文字情報のほか、英語の場合は単語ごとに、日本語の場合は、「わかち書き」ないし、形態素ごとに発話時間が付与されている。例えば、“Thank you very much.”であれば、4 単語それぞれについて発話時間（開始と終了の時間）が記録されている。これらのデータは ELAN から任意のフォーマットで書き出し可能である。

3. 校正作業

上述の通り、書き起こしテクストの構築にあたっては、音声認識ソフトを用いた後、人手による 3 段階の校正を行った。校正作業の際には、正確性確保、工数削減、応用可能性を意識した。

正確性確保

原発話者および通訳者が実際に発した言語音を、話者の意図に忠実に文字化した。作業上の問題としては、聞き取りにおける個人の技能差、音声として表れない要素の表記（例：漢字・カナ表記、句読点の仕様）などがある。作業品質のばらつきを低減するために、共通の指針を用いた。

工数削減

本データベースには、計約 77 時間分の記者会見の通訳記録が収められているが、校正作業の対象となった音声データの総量は、のべ 450 時間分に達した。限られた人員で、大量のデータを処理するため、作業工数の削減が必要であった。このため、自動音声認識ソフトを活用し、そのアウトプットをできるだけ活かす形で校正作業を行った。作業の実行可能性に関わる要因（作業人員の入れ替わり、大学間での連携）も考慮すると、校正作業は簡素なものとせざるを得なかった。

応用可能性

通訳を対象とする研究には、さまざまな手法・目的があり、求められるデータの仕様は多様である。本データベースの構築に当たっては、できるだけ幅広い研究に対応するため、あえてデータの加工は最小限に抑え、データベースとして提供することとした。利用者は、各自の関心に応じ、このデータベースを加工して使用されたい。

4. おわりに

本データベース構築の成果は、通訳・翻訳の専門職に必要な知識の明示化のための基礎研究として位置付けられる。今後、JNCP コーパスを活用した様々な研究、例えば、AI 分野を含む機械学習や認知的メカニズムの解明、通訳・英語教育への活用など、新たな研究の領野を切り拓くことが期待される。

.....

【著者紹介】

松下佳世（MATSUSHITA Kayo）立教大学異文化コミュニケーション学部・研究科准教授。主たる研究テーマはニュース・トランスレーション、コーパス通訳研究（CIS）、通訳訓練など。近著に *When News Travels East: Translation Practices by Japanese Newspapers* がある。

山田優（YAMADA Masaru）関西大学 外国語学部・外国語教育学研究科教授。研究の関心は、ポストエディットなどの翻訳テクノロジー論（MTPE、CAT）、翻訳プロセス研究（TPR）、

翻訳教育論（TILT）、翻訳とメタ言語研究、翻訳コンピテンス研究、翻訳社会論。
石塚浩之（ISHIZUKA Hiroyuki）広島修道大学人文学部英語英文学科教授。主たる研究テーマは同時通訳における認知プロセスの明示化。近年は、サイト・トランスレーション、リプロダクションなど、通訳訓練法の外国語教育への応用にも関心を持っている。

.....

【註】

1. 日本記者クラブは主に日本メディア向けの記者会見を行っていることから、通訳が付く場合は外国語話者がスピーカーである場合がほとんどであるため、このような偏りが生じた。
2. 当初は IBM Watson Speech to Text のみを利用していたが、アップデートが頻繁で、そのたびに ELAN に取り込むためのスクリプトの書き直しが必要となったため、途中から Speechmatix に切り替えた。

【謝辞】

JNPC コーパスの構築は、2016 年度から 4 年にわたり、日本学術振興会の科学研究費助成事業「記者会見通訳の二言語並行コーパスの構築と応用研究」（基盤研究 B、16H02915）の支援を受けて行われました。本研究プロジェクトに参画頂いた日本通訳翻訳学会の船山仲他元会長、水野的元会長、染谷泰正元理事、歳岡冴香会員には多大なご尽力を賜りました。この場をお借りして、御礼申し上げます。またデータを取りまとめてくれた平岡裕資氏（関西大学大学院修了生）他、トランスクリプトの修正・校正にご協力いただいた立教大学、関西大学、広島修道大学の学部生、大学院生の皆様にも感謝の意を表します。

Report JAITS

An Overview of the Japan National Press Club (JNPC) Interpreting Corpus

MATSUSHITA Kayo[1] YAMADA Masaru[2] ISHIZUKA Hiroyuki[3]

([1]Rikkyo University [2]Kansai University [3]Hiroshima Shudo University)

1. Introduction

The Japan National Press Club (JNPC) Corpus is a Japanese-English bilingual interpreting corpus that has been developed mainly for the purpose of research and education in the field of Translation and Interpreting Studies (TIS). The corpus was developed by transcribing source speeches and the interpreters' renderings from videos of press conferences made publicly available by the JPNC on YouTube. These transcripts, along with the original video and audio, can be viewed using ELAN, a free open source software. The recorded data consists of approximately 77 hours of press conference broadcasts and is available on the website of *Gengo Shigen Kyokai* (GSK or the Language Resources Association, https://www.gsk.or.jp/catalog/gsk2020-a/). In this report, we offer a detailed description of the JNPC corpus. Those who wish to use this corpus for their own research projects are requested to cite this report.

2. Contents of the JNPC corpus

2.1 Original data

At the time of its release in April 2020, the JNPC corpus included 79 sets of data, generated from press conferences held at the Japan National Press Club between 2010 and 2017. 71 data sets were based on simultaneously interpreted press conferences and 8 on those that were interpreted consecutively. The press conferences ranged in length from 17 minutes and 35 seconds to 2 hours 7 minutes and 12 seconds, with the average length being 58 minutes and 21 seconds.

Each press conference consisted of an opening speech and the subsequent question and answer session. Of the original speeches recorded for the corpus, 75 were in English and 4 were in Japanese.[1] Generally, two interpreters were assigned for simultaneous interpretation and one interpreter for consecutive interpretation. The speeches were interpreted in one direction only, either from English to Japanese or Japanese to English, with question and

answer sessions generally being interpreted in both directions.

2.2 Data format

English and Japanese text data is contained in ".eaf" files created using free software called ELAN (https://archive.mpi.nl/tla/elan). The same software was used to synchronize this data with video (.mp4) and audio (.wav) files provided in the corpus and display them as shown in the figure below. Audio is in stereo format, with English and Japanese audio recorded on separate channels. The transcribed text is displayed beneath the speech waveform indicator in two different annotation tiers: English (en) and Japanese (ja).

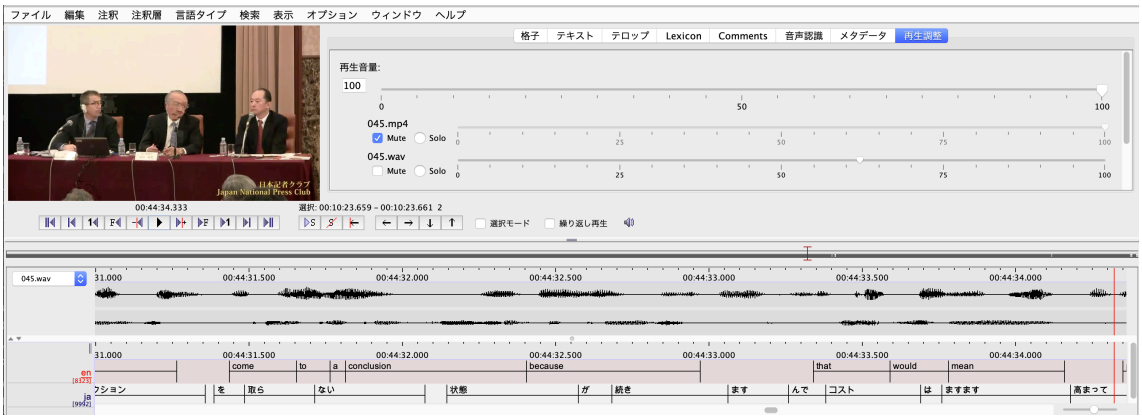

Figure 1. ELAN screenshot showing data from the JPNC corpus

2.3 Methodology for constructing text data

We generated the transcribed text automatically using speech-to-text audio recognition software for both English and Japanese audio. Two different systems were used: IBM's Watson Speech to Text and Speechmatix.[2] After we had generated the texts automatically, recognition errors were corrected manually. Three separate individuals engaged in a three-step editing process. However, we tried to leave the automatically-generated text unchanged wherever possible for reasons explained later in this report. Consequently, some discrepancies between the transcriptions of the two voice recognition systems and inconsistencies in annotation remained. Therefore, users will need to edit and customize the database as necessary, depending on their research objectives.

In addition to text information, transcripts were tagged to denote the length of utterances at the word level for English and at the segment (*wakachikaki*) or morpheme level for Japanese. For example, if the text includes the phrase, "Thank you very much," then the length of utterance (from start to finish) is recorded for each of the four words in that phrase. This data can then be exported from ELAN in several formats depending on the user's intended purpose.

3. Editing

As mentioned above, the transcribed text was generated using voice recognition software before undergoing a three-step editing process. The editing process we employed focused on maintaining accuracy, reducing work hours, and enhancing applicability.

Maintaining accuracy

We faithfully rendered the utterances of the original speakers and the interpreter(s) as text, while preserving speaker intentions. Issues we faced during this process included differences in individual ability to comprehend the audio input as well as the way of denoting orthographic elements that are not evident just by listening to the audio (e.g., Chinese characters, kana, and punctuation requirements). In order to minimize disparities in work quality, we adopted a joint policy on such matters.

Work hour reduction

A total of approximately 77 hours of interpreted press conferences are recorded in the JNPC corpus, but the total quantity of audio data that required editing exceeded 450 hours. In order for such a large quantity of data to be processed by a limited number of people, we needed to reduce the number of work hours required. To accomplish this, we used automatic voice recognition software and carried out editing in a way that preserved the automatically-generated output wherever possible. Editing had to be undertaken in a simplified format that took into account factors affecting the feasibility of carrying out the work (i.e., handovers between the different individuals involved in the project and coordination between the contributing universities).

Enhanced applicability

Research in TIS is carried out using various methodologies and for a variety of objectives, all of which require different data specifications. The JNPC corpus was constructed in order to be applicable to a wide range of research projects. Therefore, we have kept data processing to a minimum. The aim is for users to customize the database in accordance with their individual research interests.

4. Conclusion

We believe that the JNPC corpus will prove its worth by positioning itself as a resource for fundamental research in TIS that illustrates the specialist knowledge required by interpreters and translators. It is our hope that this corpus will be used to research a diverse range of topics, such as machine learning, the cognitive mechanisms of interpreting/translating, and interpreter/translator training.

.....

About the Authors:

MATSUSHITA Kayo: Associate Professor of Interpreting and Translation Studies at the College/Graduate School of Intercultural Communication at Rikkyo University. She specializes in news translation research, Corpus-based Interpreting Studies (CIS), and interpreter training. She is the author of *When News Travels East: Translation Practices by Japanese Newspapers*.

YAMADA Masaru: Professor in the Faculty of Foreign Language Studies at Kansai University. He specializes in Translation and Interpreting Studies with a focus on translation process research (TPR), including translation technology and post-editing, translation in language teaching (TILT), and (neuro-)scientific approaches to TPR.

ISHIZUKA Hiroyuki: Professor in the Faculty of Humanities and Human Sciences at Hiroshima Shudo University. His main research interest is modeling the cognitive process of simultaneous interpreting, drawing on cognitive linguistics and cognitive psychology. His current research also explores translation in language teaching (TILT).

.....

Notes

1. Press conferences held by the JNPC are primarily directed toward the Japanese media. Therefore, press conferences that require interpreting generally involve speakers of foreign languages. This accounts for the imbalance of directionality within the corpus.

2. Our research originally used only IBM Watson Speech to Text, but frequent updating of this software required us to rewrite the scripts used to export data to ELAN on each occasion. For this reason, we decided to switch to Speechmatix partway through the project.

Acknowledgments

We developed the JNPC corpus over four years from FY 2016, with the assistance received from the Japanese Society for the Promotion of Science under its grants-in-aid for scientific research program for “Construction of Japanese-English parallel corpora of interpreter-mediated press conferences and applied studies” (Grant-in-Aid for Scientific Research, Category B, 16H02915). We would like to take this opportunity to express our sincere appreciation to former Presidents of JAITS (the Japan Association of Interpreting and Translation Studies) Chuta Funayama and Akira Mizuno, former JAITS Director Yasumasa Someya, and JAITS member Saeka Toshioka for their participation and the contributions they made to this project. We also extend our gratitude to Yusuke Hiraoka (graduate of Kansai University’s graduate school) for his assistance in collating the data, and to all the undergraduate and graduate students at Rikkyo University, Kansai University, and Hiroshima Shudo University who corrected and edited the transcripts.