NAIST 同時通訳コーパスの構築：翻訳字幕との比較と通訳経験年数に基づく分析

土肥 康輔（奈良先端科学技術大学院大学 D）、須藤 克仁（奈良先端科学技術大学院大学）、
中村 哲（奈良先端科学技術大学院大学）

本発表では、我々が構築した大規模な英日・日英同時通訳コーパス（NAIST 同時通訳コーパス [1]；Shimizu ら, 2014；Doi ら, 2021）について述べるとともに、本コーパスを用いて行った同時通訳と翻訳字幕の比較、通訳者の経験年数に基づく分析について報告する。

近年の機械翻訳や音声処理技術の発展により、自動音声翻訳システムの研究が加速している。自動音声翻訳システムの構築には、原言語音声とその書き起こしと、対応する目的言語の翻訳テキストや翻訳字幕に基づく音声翻訳コーパスが用いられる。我々は自動同時通訳システムの研究におけるシステム構築や評価のための言語資源として、講演や記者会見を通訳者が実際に同時通訳した音声とその書き起こしを収録した NAIST 同時通訳コーパスを構築した。本コーパスは、通訳・翻訳研究においても、同時通訳の特徴やパフォーマンスの分析に用いることが期待できる。

同時通訳データの収録には、経験年数が異なるプロの同時通訳者が参加した。通訳者は、同時通訳の経験年数に基づき 3 段階にランク分けされている。通訳者には、発話の要約または書き起こしを事前に提示し、収録にあたっては、通訳者はヘッドセットを装着し、コンピュータでビデオを視聴しながら通訳を行った。2018～2020 年の期間に合計で 300 時間以上の英日・日英同時通訳データを収録し、その一部には、ランクが異なる 3 名の同時通訳者が同一の原文を訳出した、通訳結果を比較分析可能なデータ（通訳比較用データ）を含んでいる。

本発表における分析は、この通訳比較用データに含まれる 14 本の講演に対する英日同時通訳データを対象に、訳出遅延、品質、語順の観点から行った。原言語（英語）の文を基準として、原文と同時通訳の対応付けを人手で行った。さらに、このうちの 3 本については、同時通訳と翻訳字幕の対応付けを文節レベルで行った。これらのデータをもとに、訳出遅延、品質、語順に関する評価指標を算出した。また、プロの翻訳者 3 名による品質評価を行った。分析の結果、経験を積んだ通訳者は、遅延時間と品質のバランスをよりよく保てていることが明らかになった。また、遅延時間があまりに長くなると、同時通訳の品質に悪影響が生じていることが明らかになった。

【参考文献】

Shimizu, H., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2014). Collection of a Simultaneous Translation Corpus for Comparative Analysis, In *Proceedings of LREC*, pp. 670-673.

Doi, K., Sudoh, K., and Nakamura, S. (2021). Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data, In *Proceedings of IWSLT*, pp. 226-235.

【注】

1. コーパスの一部はウェブ上で公開している。 https://dsc-nlp.naist.jp/data/NAIST-SIC/