CIAIR 同時通訳データベース －設計・構築・分析－

遠山 仁美*, 松原 茂樹, 笠 浩一朗, 河口 信夫（名古屋大学）
稲垣 康善（愛知県立大学）

Design, Construction, and Analysis of CIAIR Simultaneous Interpretation Corpus
Hitomi Toyama, Shigeki Matubara, Koichiro Ryu, Nobuo Kawaguchi (Nagoya University)
Yasuyoshi Inagaki (Aichi Prefectural University)

1. まえがき

名古屋大学統合音響情報研究拠点（以下、CIAIR）では、異言語間コミュニケーション支援環境の実現を目指し、同時通訳データベースの構築を進めてきた(1)。全体で約 182 時間の音声を収録し、文字化データのサイズは単語数にして約 100 万語に達しており、世界最大の同時通訳データベースと位置付けられる。

本稿では、名古屋大学 CIAIR 同時通訳データベース（SIDB）について、設計、収集、構築、及び、分析について概説する。

2. データベースの設計

独話、及び、対話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、自由発話で収録を行った。収録様式を Table 1 に示す。本データベースは音声データファイル、文字化データファイル、環境データファイルの３つから構成されている。

- 独話データの収集

  １人の講演者に対し、複数の同時通訳者が通訳を行った。すなわち、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しており、複数の通訳事例を比較することが可能である。

- 対話データの収集

  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した。

3. データベースの構築

音声データの文字化は CSJ の書き起こし基準に準拠した(2)。話者および、通訳者の音声を 200msec 以上のポーズで分割し、発話単位を定めた。また、時間情報タグ（発話開始・終了時刻）、および、フィラーや言い淀みについて談話タグを付与している。データベースの規模を Table 2 に示す。

- 音声データの視覚化

  音声データを視覚化するためのツールを開発し、話者と通訳者の発声タイミングをタイムチャートによって閲覧できる（Fig 1 参照）。これにより、様々な通訳現象を観察することができる。

- 対訳対応データの作成

  人手による対訳アライメント作業を支援するツールを開発し、対訳対応データを蓄積している（Fig 2 参照）。

4. データベースの分析

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、訳を出力するタイミングなどが重要な課題となる。我々は、本データベースを分析することにより、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、同時通訳プロセスの解明、及び、同時通訳方略の蓄積を進めている(3)(4)(5)。

5. まとめ

本稿では、名古屋大学 CIAIR 同時通訳データベースの設計、収集、構築、および、分析について述べた。

大規模音声言語データベースの需要は、音声言語処理の分野のみならず、認知科学や音声学、言語学に及ぶ諸分野において増大しつつある。本データベースにおいても、より広範な研究分野で多面的に活用されるために、データの公開を進めている。詳細については、http://www.el.itc.nagoya-u.ac.jp/sidb/ を参照されたい。

Table 1. CIAIR 同時通訳データベースの収録様式

談話形態	独話　対話
対象言語	英語　日本語
通訳スタイル	同時通訳
メディア	音声　文字

Table 2. データベースの基礎統計データ

項目		単語数	発話数	収録時間（分）
話者	英語	198099	22645	2373
	日本語	190536	23014	2275
	合計	388635	45659	4648
通訳者	英日	382826	40793	3317
	日英	219734	29802	2943
	合計	602560	70595	6260
合計		991195	116254	10908

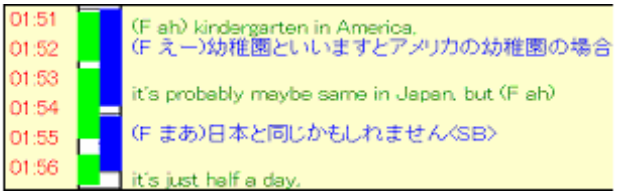

Fig 1. タイムチャートの例 （独話）

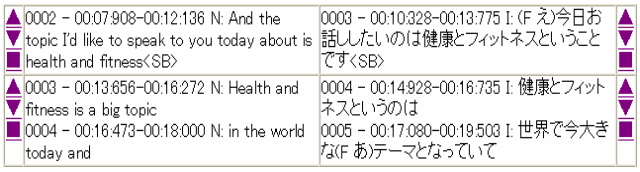

Fig 2. 対訳対応支援ツールの例

謝辞

日頃、ご指導下さる名古屋大学教授の渡邊豊英先生に感謝致します。本研究の一部は、文科省科研費COE形成基礎研究費によります。

文　献

(1) 松原. 他 : 通訳研究, No.1, pp. 85-102, 2001
(2) 前川 : 平成15年度国立国語研究所公開研究発表会, pp.1-8, 2003
(3) 高木. 他 : 言語処理学会第8回年次大会発表論文集, pp.383-386, 2002
(4) 大原. 他 : 通訳研究, No. 3, pp. 35-53, 2003
(5) 遠山. 他 : 信学技報, Vol.103, No.487, pp.13-18, 2003