書き言葉向き大語彙辞書を用いたバイリンガル旅行会話コーパスの特徴分析*

竹沢 寿幸 (ATR 先端情報科学研究部)
大山 芳史 (NTT コミュニケーション科学基礎研究所)

要旨

NTT が書き言葉を対象として構築した大語彙辞書「日本語語彙大系」を用いてATR音声翻訳通信研究所が収集した「バイリンガル旅行会話コーパス」の特徴分析を行なった。まず、コーパスに含まれる単語を抽出し、それらが日本語語彙大系の単語体系に含まれているか調査した。次に、コーパスに含まれる用言を中心とするコロケーションを抽出し、それらが日本語語彙大系の構文体系に含まれているか、含まれているならばさらに英訳の妥当性を調査した。書き言葉向きの大語彙辞書を対照データとすることで、会話特有の語彙、単語の連結により構成される表現の特徴を検討する。語彙的な特性に関する数値情報を中心に、その概要を報告する。

1 まえがき

現在の音声翻訳は、限定されたタスクへの適用を想定し、比較的小規模の語彙(約13,000語)と、表現の種類をある程度限定して翻訳実験を行なっている[1, 2, 3, 4]。音声翻訳の適用範囲の拡大を目指すには、大語彙、多様な表現、分野適応性の問題を解決する必要がある。新聞記事等の書き言葉向きの機械翻訳システムは音声翻訳システムに比較して大語彙を扱っているが、テキストの翻訳と会話の翻訳は違った難しさがあるという指摘[5]がある。会話文など話し言葉向きの機械翻訳システムの研究[6]はあるものの、書き言葉向きの機械翻訳システムとは別に研究がなされており、会話文など話し言葉の機械翻訳とテキストなど書き言葉の機械翻訳の扱う内容の違いに関する分析調査はこれまで十分になされていなかった。

近年 NTT が構築した産業経済記事など記述文を対象とする日英機械翻訳システム ALT-J/E [7] の意味辞書の一部が「日本語語彙大系」として出版され[8]、その CD-ROM 版[9]も利用できるようになった。一方、ATR 音声翻訳通信研究所が収集した「バイリンガル旅行会話コーパス」[10, 11]も広く公開されている。実際的かつ定量的な調査分析データは書き言葉向きの機械翻訳システムで会話調の話し言葉を扱う際の指針や有益な知見を与えるものと期待できる。そこで、機械処理可能な大語彙辞書である日本語語彙大系を用いて電子化されたバイリンガル旅行会話コーパスの特徴分析を試みた。まず、コーパスに含まれる単語を抽出し、それらが日本語語彙大系の単語体系に含まれているか調査した。次に、コーパスに含まれる用言を中心とするコロケーションを抽出し、それらが日本語語彙大系の構文体系に含まれているか、含まれているならばさらに英訳の妥当性を調査した。書き言葉向きの大語彙辞書を対照データとすることで、会話特有の語彙、単語の連結により構成される表現の特徴を検討する。本稿では、語彙的な特性に関する数値情報を中心に、その概要を報告する。

2 で分析調査対象である日本語語彙大系とバイリンガル旅行会話コーパスの概要について述べる。**3** で旅行会話に現れた会話特有の語彙について述べる。**4** で旅行会話に現れた会話特有の語連結表現について述べる。最後に **5** で全体をまとめ、関連情報を記す。

2 分析調査対象

2.1 日本語語彙大系

NTT では、産業経済記事など記述文を対象とする日英機械翻訳システム ALT-J/E [7] を実現し、その意味辞書の一部が日本語語彙大系として出版されている[8]。日本語の語彙30万語を3,000種類の意味

*Characteristics of a bilingual travel conversation corpus in comparison with a large vocabulary dictionary for written language by Toshiyuki TAKEZAWA (ATR) and Yoshifumi OOYAMA (NTT)

表 1: 日本語語彙大系の概要 (部分)

収録データ	件数
単語体系	30 万語
構文体系	6,000 用言 14,000 文型

表 2: バイリンガル旅行会話コーパスの概要

収集会話数	618
異なり話者数	71
異なり通訳者数	23
発話総数	16,107
日本語形態素延べ数	301,961

属性で分類し、さらに6,000 語の用言には日英の文型14,000 パターンが付与されている。**表 1**に日本語語彙大系のうち本稿で調査対象とした項目の概要を示す。分析調査にはその CD-ROM 版 [9] および必要に応じて適宜 ALT-J/E [7] の辞書情報を活用した。関連情報は http://www.kecl.ntt.co.jp/icl/mtg/resources/GoiTaikei/ をご覧いただきたい。

2.2 バイリンガル旅行会話コーパス

ATR 音声翻訳通信研究所では音声翻訳研究のためにバイリンガル旅行会話コーパス [10, 11] を構築した。良質で大量の基礎資料を得るため、通訳者を介したバイリンガル会話を収集した。また実用性を考えて、タスクとしては多くの人に利用可能なホテル予約を中心とした旅行会話を選んだ。旅行会話は日本語話者、英語話者と日英方向、英日方向の 2 名の通訳者の合計 4 名によってなされる。話者の一方はホテルのフロント係であり、他方は外国人の旅行客である。2 名の通訳者は音声翻訳システムの代りとして振る舞う。具体的には、話者の 1 回の発話は 10 秒以内とし、相手が話している間に割り込むことは禁止した。そして、通訳者はそのような 1 回の発話毎に逐次的に通訳を行なう。このような制約を設けることにより、極端に長い発話や発話の重なりを避けることができるため、「近未来の音声翻訳システム」研究の基礎資料として適していると判断できた。**表 2**にバイリンガル旅行会話コーパスの概要を示す。

```
表記形 | 読み | 標準形 | 品詞 |(出現頻度)
の | ノ | の | 格助詞 |(69)
の | ノ | の | 終助詞 |(4)
の | ノ | の | 準体助詞 |(337)
の | ノ | の | 連体助詞 |(4135)
```

図 1: 助詞「の」の例

表 3: 品詞体系の違い (助詞の例)

旅行会話コーパス	日本語語彙大系や ALT-J/E
格助詞	格助詞
準体助詞	なし (形式名詞)
係助詞	なし (副助詞)
副助詞	副助詞
並立助詞	なし (格助詞)
接続助詞	接続助詞
終助詞	終助詞
連体助詞	なし (格助詞)
引用助詞	なし (格助詞)

3 旅行会話に現れた会話特有の語彙

バイリンガル旅行会話コーパスのうち日英方向の発話を対象に、単語の抽出と頻度のカウントを行なった。助詞「の」の例を図 **1**に示す。記号 | で区切られた項目は左から順に「表記形」「読み」「標準形」「品詞」を表し、() 内の数値が出現頻度である。

表 3に助詞の例を示すように、旅行会話コーパスの品詞体系 [12] は日本語語彙大系や ALT-J/E の辞書情報とは異なる。そこで、その違いを考慮して、表 3の () 内に記すように、旅行会話コーパスの準体助詞は日本語語彙大系や ALT-J/E の形式名詞に、旅行会話コーパスの係助詞は日本語語彙大系や ALT-J/E の副助詞に、旅行会話コーパスの並立助詞、連体助詞、引用助詞は日本語語彙大系や ALT-J/E の格助詞に対応させて照合を行ない、日本語語彙大系の単語体系に基づく被覆率を調査した。

図 1の例では、格助詞「の」、準体助詞「の」、連体助詞「の」は日本語語彙大系や ALT-J/E に同等とみなせる語彙項目が含まれていると判断できた。しかし、終助詞「の」は日本語語彙大系や ALT-J/E には含まれていなかった。したがって、異なり語数を基準とする被覆率は 3/4 となり、75.0% である。延べ語数を基準とする被覆率は 4541/4545 となり、99.9% で

表 4: 旅行会話コーパスに対する被覆率

	延べ語数	異なり語数
含まれるもの	80.685	2.493
含まれないもの	18.828	1.269
被覆率	81.1%	66.3%

ある。

バイリンガル旅行会話コーパスには言い直し等の単語の断片が含まれており、それらは形態素の品詞に「その他」と付与されている。また、バイリンガル旅行会話コーパスでは活用する語は語幹と語尾に分割しているが、日本語語彙大系では語幹と語尾に分割していない。そこで、品詞「その他」「語尾」等を除き、被覆率を求めた。結果を表 4に示す。

日本語語彙大系の単語体系が含まない単語は大きく二つに分類できる。一つは会話調の話し言葉特有の表現であり、もう一つは旅行という分野に依存する語彙である。それぞれの例を次に示す。

- 会話調の話し言葉特有の表現: 主に感動詞、副詞、接続詞、助詞、助動詞等

  「ありがとうございました(感動詞)」、「いらっしゃいませ(感動詞)」、「すみませんが(副詞)」、「そうしますと(接続詞)」、「じゃ(格助詞)」、「の(終助詞)」、「ちゃ(助動詞)」等

- 旅行という分野に依存する語彙: 主に普通名詞、固有名詞等

  「さば寿司(普通名詞)」、「カナディアンロッキー(固有名詞)」、「嵐山線(固有名詞)」等

日本語語彙大系の単語体系に含まれない単語のうち、名詞類が異なり語数で 888 語、延べ語数で 7,261 語を占める。大まかに名詞類がすべて旅行という分野に依存する語彙であると近似すれば、日本語語彙大系の単語体系に含まれない単語のうち旅行という分野に依存する語彙の占める割合は、異なり語数基準で約 70%、延べ語数基準で約 40% である。逆に、その残りが会話調の話し言葉特有の表現の占める割合である。

4 旅行会話に現れた会話特有の語連結表現

バイリンガル旅行会話コーパスのうち日英方向の発話を対象に、用言を中心とするコロケーションの抽出を行なった。いわゆる「ダ文」は会話調の言い回しに多用される。その事例の一部を図 2に示す。内容項目は左から順に出現頻度、日本語構文、英語構文である。出現頻度は日本語構文と英語構文のペアを単位として求めた。日本語構文はダ文表現と格要素相当語句から構成される。ダ文表現の前後を / で囲んで示し、ダ文の「だ」の直前に # を付与する。「です」等の表層表現となっているものもすべて「# だ」の形で抽出した。主題を表す係助詞「は」で格助詞「が」に置き換え可能な場合は「が((は))」と記述した。助詞を伴わない格要素相当語句において格要素を示す助詞を補うことができる場合は * で囲んだ形式でそれを記述した。英語構文全体は " で囲んで示し、さらに日本語のダ文表現に相当する部分を / で囲んで記述する。英語に人称代名詞が含まれる場合はそれを one に置き換えた。英語構文として連結する必要はないが、日本語構文の対訳として必要な要素は、| で区切って記述した。

また、会話調の言い回しでは「朝食は別になる」「ツインルームは満室となる」等の「になる」「となる」という言い回しが多用される。これらも書き言葉の訳し分けとは異なる傾向があり、興味深い。これら会話特有の語連結表現に関する分析結果や数値情報の詳細に関しては、改めて別の機会に報告する予定である。

5 むすび

NTT が書き言葉を対象として構築した大語彙辞書である日本語語彙大系を用いて ATR 音声翻訳通信研究所が収集したバイリンガル旅行会話コーパスの特徴分析を行なった。テキストの翻訳と会話の翻訳は違った難しさがあるという定性的ないし思索的な指摘はあったが、会話文など話し言葉の機械翻訳とテキストなど書き言葉の機械翻訳の扱う内容の違いに関する分析調査はこれまで十分になされていなかった。そこで、機械処理可能な大語彙辞書や電子化されたバイリンガル会話コーパスが公開されたことを背景に、実際的かつ定量的な調査分析を試みた。このような調査分析は書き言葉向きの機械翻訳システムで会話調の話し言葉を扱う際の指針や有益な知見を与えるものと期待できる。

なお、本稿で述べたバイリンガル旅行会話データベースに関する問合わせ先は次の通りである。

```
1 / 雨 # だ / "it /rain/"
1 / 雨 # だ / "/rain/"
3 / 初めて # だ / "/be/ one's first time"
1 / 初めて # だ / "/be/ one's first visit"
2 / 初めて # だ / "/be/ the first time"
1 奈良が ((は)) / 初めて # だ / "this /be/ one's first trip | to Nara"
1 着物が ((は)) / 初めて # だ / "/be/ one's first time | to put kimono on"
1 着物が ((は)) / 初めて # だ / "/be/ one's first time | wearing a kimono"
1 それは / お困り # だ / "that /be/ a problem"
1 バスを / ご利用 # だ / "/take/ the bus"
1 こちらが / チケット # だ / "here /be/ one's tickets"
1 サービス料 * が */ 込み # だ / "/include/ service charges"
1 東京成田空港を / 出発 # だ / "/leave/ Tokyo Narita Airport"
1 一つ目が / 東寺駅 # だ / "the first stop /be/ Toji Station"
1 茶室が / 入り用 # だ / "/need/ a tearoom"
3 / ご存じ # だ / "/know/"
6 / お泊まり # だ / "/stay/"
```

図 2: いわゆる「ダ文」の例 (一部)

〒 619-0288 京都府相楽郡精華町光台 2-2
(株) 国際電気通信基礎技術研究所 (ATR) 開発室
電話: (0774) 95 1192
ファクシミリ: (0774) 95 1179
電子メール: deliv@ctr.atr.co.jp
URL: http://www.atr.co.jp/results/

謝辞

分析作業に協力いただいた NTT アドバンステクノロジ株式会社 サービスシステム事業部 言語処理システム部の皆様に感謝する。

参考文献

[1] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto, "A Japanese-to-English speech translation system: ATR-MATRIX," *Proc. International Conference on Spoken Language Processing*, pp. 2779–2782, 1998.

[2] T. Takezawa, F. Sugaya, A. Yokoo, and S. Yamamoto, "A new evaluation method for speech translation systems and a case study on ATR-MATRIX from Japanese to English," *Proc. Machine Translation Summit*, pp. 299–307, 1999.

[3] E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai, "Solutions to problems inherent in spoken-language translation: the approach of ATR-MATRIX," *Proc. Machine Translation Summit*, pp. 229–235, 1999.

[4] F. Sugaya, T. Takezawa, A. Yokoo, and S. Yamamoto, "End-to-end evaluation in ATR-MATRIX: speech translation system between English and Japanese," *Proc. EUROSPEECH*, pp. 2431–2434, 1999.

[5] 長尾真, "情報技術の新時代に向けて," 情報処理, vol. 41, no. 1, pp. 48–49, 2000.

[6] 古瀬蔵, 山本和英, 山田節夫, "構成素境界解析を用いた多言語話し言葉翻訳," 自然言語処理, vol. 6, no. 5, pp. 63–91, 1999.

[7] 八巻俊文, 大山芳史, 白井諭, 横尾昭男, "機械翻訳特集: 日英機械翻訳システム ALT-J/E の研究開発," *NTT R&D*, vol. 46, no. 12, pp. 1391–1398, 1997.

[8] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦 (編), "日本語語彙大系," 岩波書店, 1997.

[9] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦 (編), "日本語語彙大系 *CD-ROM* 版," 岩波書店, 1999.

[10] T. Morimoto, N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi, and Y. Yamazaki, "A speech and language database for speech translation research," *Proc. International Conference on Spoken Language Processing*, pp. 1791–1794, 1994.

[11] T. Takezawa, "Building a bilingual travel conversation database for speech translation research," *Proc. 2nd International Workshop on East-Asian Language Resources and Evaluation — Oriental COCOSDA Workshop '99 —*, pp. 17–20, 1999.

[12] 竹沢寿幸, "音声言語データベースの日本語形態素情報マニュアル — 最終版 —," *ATR* テクニカルレポート, TR-IT-0315, 1999.