日英コードスイッチング音声データの構築*

◎中山佐保子 (NAIST), ド クオック チュオン (NAIST),
サクティ サクリアニ (NAIST/RIKEN), 中村哲 (NAIST/RIKEN)

1 はじめに

バイリンガル話者が会話の中で二つ以上の言語を混ぜることを，コードスイッチングという。日本においてもバイリンガルの数は増加しており，実際に日英コードスイッチングを使用した会話が観察されている [1]。そのため，日英コードスイッチングを認識可能な音声認識が必要である。これまで，中国語と英語のコードスイッチング音声認識 [2] や，フリジア語とオランダ語のコードスイッチング音声認識 [3] はあるが，日英コードスイッチングを扱う研究については，言語ごとに文と話者が異なる音声認識のみとなっている [4]。そこで，我々は日本語と英語のコードスイッチングに焦点を当て，音声認識モデルのための音声データを構築することにした。バイリンガル話者の音声から作成した日本語と英語の音声合成システムを利用することによって，時間と費用がかからない日英コードスイッチング音声データの構築に取り組んだ。

2 日本語におけるコードスイッチング

コードスイッチングは主に文間と文中に分類される。文間コードスイッチングは文の間で言語の変換が起こり，文中コードスイッチングは文の中で変換が行われる。文中コードスイッチングは場所や長さによって様々なものがある。以下に実際のコードスイッチングの例を示す [1]。

- [文間コードスイッチング]:
  ああ，そうだってね。 On the honeymoon, they bought this. (ああ，そうだってね。新婚旅行でこれを買いました。)
- [文中単語レベルコードスイッチング]:
  Trust-してる人にだけ貸してあげるの。(信頼してる人にだけ貸してあげるの。)
- [文中フレーズレベルコードスイッチング]:
  それは that's not his arm. (それは彼の腕じゃない。)

なお，外来語や引用によるコードスイッチングについては，理論的にはコードスイッチングではないかもしれないが，全ての日英会話の認識を目標としているため，それらもコードスイッチングの枠組みの中で扱うこととする。

3 コードスイッチングのデータ構築

データ構築の概要は Fig. 1 に描かれている。データ構築のプロセスは，テキストデータ構築と音声データ構築の二つのプロセスで構成される。

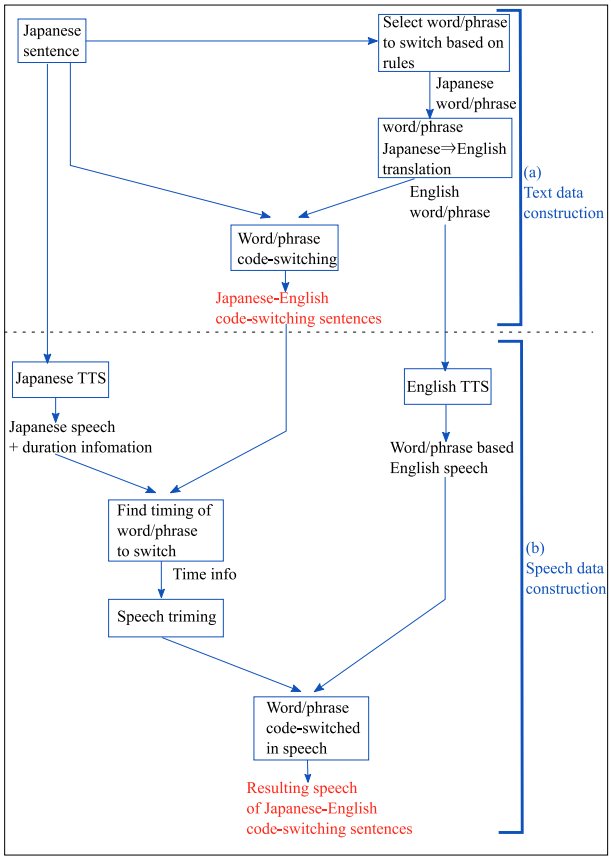

Fig. 1 Overview of Japanese-English code-switching data construction

3.1 テキストデータ構築

最初に BTEC コーパス [5] の日本語テキスト文から，単語またはフレーズを選び，Google 翻訳 API を用いて英語に翻訳した。そして，翻訳した単語を元の日本語文に挿入することによって文中コードスイッチング文を作成した。単語の選択については主にカタカナ文字の単語を選び，フレーズの選択について

*Corpus Construction of Japanese-English Code-switching by Sahoko Nakayama(NAIST), Quoc Truong Do(NAIST), Sakriani Sakti(NAIST/RIKEN), and Satoshi Nakamura(NAIST/RIKEN)

は実際のコードスイッチング文 [1] を参考に，助詞の後のフレーズを選択した。

3.2　音声データ構築

3.2.1　音声合成

テキストデータを作成した後，バイリンガル話者の音声から作成した音声合成システム [6] を用いて音声を作成した。日本語発話については，コードスイッチングを作成する前の日本語文全体を合成し，同時にその音声の時間情報のファイルを作成した。英語発話は翻訳した英語の単語またはフレーズを取り出して合成した。

3.2.2　トリミングと結合

取得した時間情報に基づいて，言語が切り替わる場所で日本語発話をトリミングし，英語の音声と結合した (Fig. 2)。その結果，日英コードスイッチングの発話音声を作成することができた。

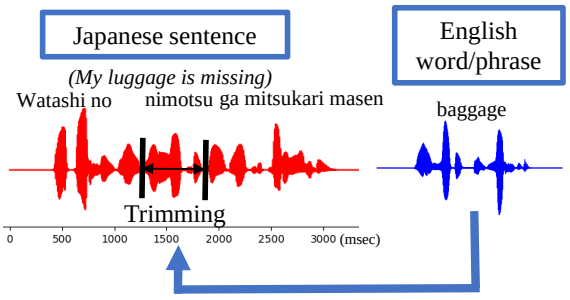

Fig. 2　Speech trimming and concatenation process

4　データ分析

データを構築した後，コードスイッチングに含まれる日本語の割合を調査した。MeCab[1]で解析される形態素を一つの単位として，各言語の数を全体の数で割って割合を計算した。結果は Table 1 に示されている。モノリンガル日本語のコーパスでは日本語が 100%でモノリンガル英語では英語が 100%であるのに対し，単語レベルコードスイッチングでは日本語が 88%で英語が 12%，フレーズレベルコードスイッチングでは日本語が 46%で英語が 56%になっている。

5　おわりに

本論文では，日英コードスイッチングコーパスの構築について紹介した。日英コードスイッチング発話はバイリンガル話者を用いた日本語と英語の音声合成システムから作成した。構築したコーパスは英語を 12%含む 146k の文中単語レベルコードスイッチング

[1]MeCab は京都大学で開発された形態素解析ツール

Table 1　Statistics of Japanese-English code-switching(CS) speech utterances

	Utterances	Japanese
Monolingual Japanese	273k	100%
Word-level CS	146k	88%
Phrase-level CS	146k	46%
Monolingual English	273k	0%

と英語を 54%含む 146k の文間コードスイッチングで構成されている。今後は，自然なコードスイッチングの音声と比較することによって，作成したコードスイッチング音声の質を向上させ，音声認識の開発に役立てる所存である。

謝辞　本研究は科研費 [JP17H06101,JP17K00237] の助成を受けております。

参考文献

[1] Nakamura, “Developeing codeswitching patterns of a Japanese/English bilingual child,” in *Proceedings of the 4th International Symposium on Bilingualism*, 1679-1689, 2005.

[2] Vu *et al.*, “A first speech recognition system for Mandarin-English code-switch conversational speech,” in *ICASSP*, 4889-4892, 2012.

[3] Yilmaz *et al.*, “Investigating bilingual deep neural networks for automatic recognition of code-switching Frisian speech,” *Procedia Computer Science*, col.81, 159-166, 2016, SLTU2016 5th Workshop on Spoken Language Technologies for Under-resourced languages.

[4] Seki *et al.*, “An end-to-end language-tracking speech recognizer for mixed-language speech,” Calgary, Canada, 2018.

[5] Takezawa *et al.*, “Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World,” *The International Conference on Lnaguage Resources and Evaluation*, 147-152, 2002.

[6] Truong *et al.*, “Collection and analysis of japanese-english emphasized speech corpus,” in *Proceedings of Oriental COCOSDA*, 2014.