

# Japanese-English Code-switching Speech Data Construction

Osahoko Nakayama, Takatomo Kano, Quoc  
Truong Do, Sakriani Sakti, Satoshi Nakamura  
Nara Institute of Science and Technology  
Japan

# Code-switching

## ◆ Code-switching (CS):

A speaker switches languages within a conversation

- [Inter-sentential code-switching]:

ああ、そうだってね。 On the honeymoon, they bought this. [Nakamura+2005]  
(*Oh, year, your're right. On their honeymoon, they bought this*)

- [Intra-sentential word-level code-switching]:

確率を与えたので、その確率を使ってrecountingします。 [From NAIST lecture ]  
(Since we gave a probability, we recalculate using that probability.)

- [Intra-sentential phrase-level code-switching]:

わたしね、that's not the pointとか言われたもんね。 [Fujimura+2013]  
(*I was told that that's not the point.*)

## ◆ Loanword insertion:

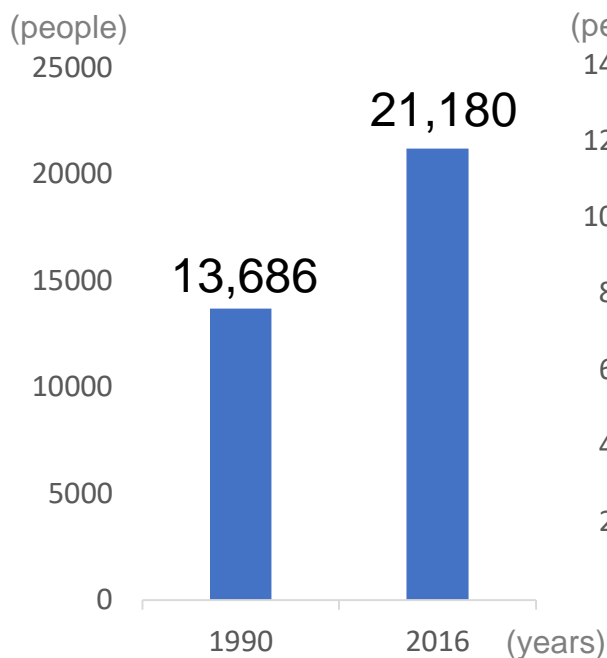
中間言語を使った時のメリットに何があるか？ [From Naist lecture]  
(*What is the merit of using an interlingua?*)

# Bilinguals and Code-switching

**Code-switching plays a vital role in bilingualism.** [McSwan+2000]

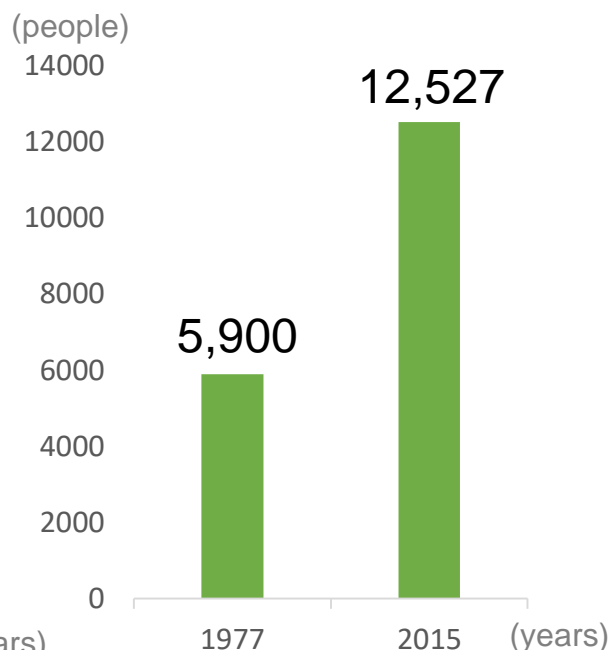
**Bilingual speakers have increased in Japan.**

Children  
with a foreign parent



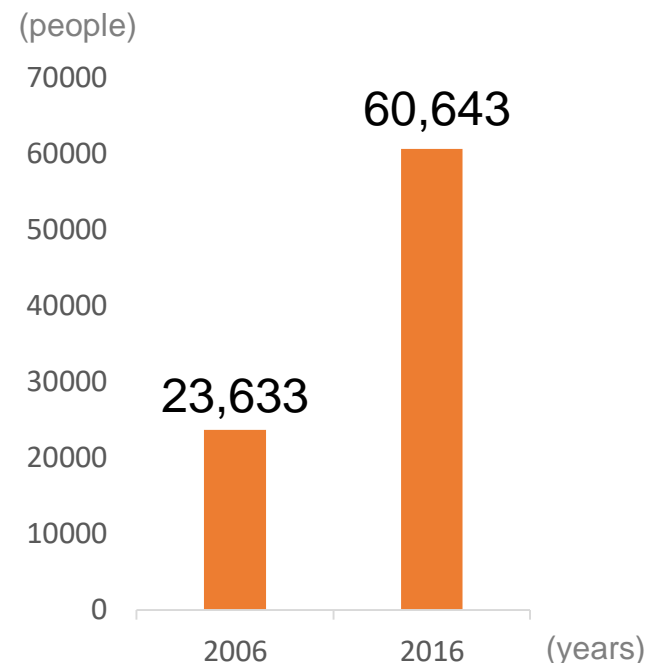
Data : MHLW

Children  
who have lived abroad



Data : MEXT

Students  
who study abroad



Data : JASSO

# Reports of Code-switching in Japanese-English Case

## 1. A Japanese child who lives in the US



By aturukus

**59** code-switching in half an hour

[Nakamura+2005]

## 2. Bilingual children living in Japan, who have foreign parents

**153** code-switching in 4 hours

[Fujimura+2013]



# Speech Recognition

## Requirement

Realize to recognize every conversation

### <Usage scenes >

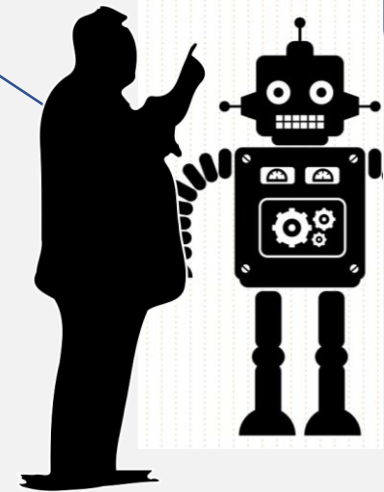
#### 1. Lecture, Meeting

翻訳された文章をtarget sentence  
といいます。 [From NAIST lecture]  
(The translated sentence is called  
target sentence)










#### 2. Dialogue with a robot

えっと、日本語で”coincidence”って  
何ていうんだっけ？  
(Um, what do you say “coincidence”  
in Japanese?)



# Previous Research in ASR

Tried to train speech recognition to recognize code-switching

- Chinese  English  code-switching (GMM-HMM) [Vu+2012]
  - Frisian  Dutch  code-switching (DNN-HMM) [Yilmaz+2016]
- Japanese  English  case has received scant research
- 

**We focus on Japanese-English code-switching.**

# Data Construction

Japanese-English code-switching ASR must be developed, but:

## Problems

- ① No large-scale data exists for training model
- ② Time-consuming and expensive for collecting



So we constructed

Japanese-English code-switching speech database by utilizing TTS

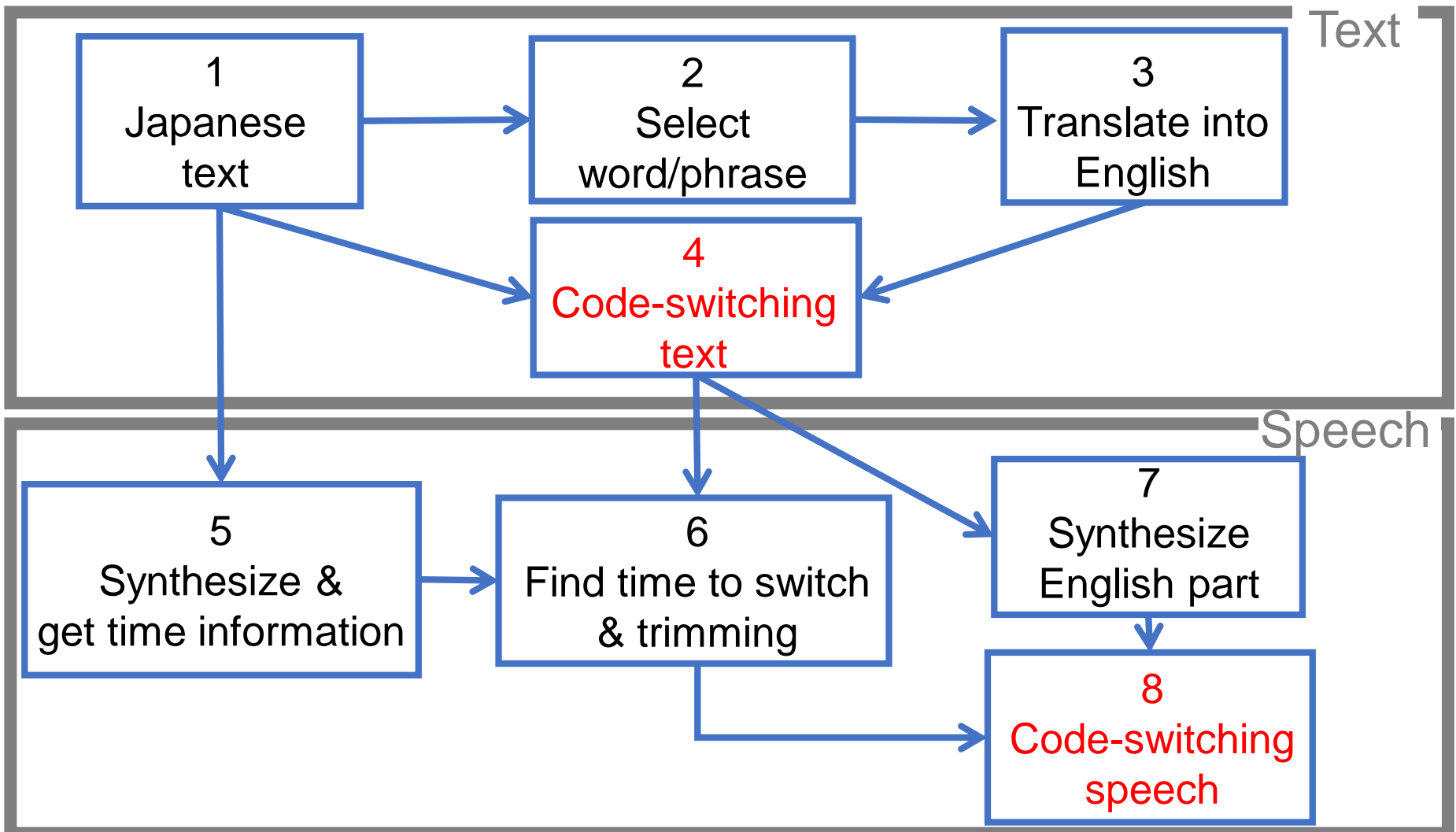
Previous collected  
code-switching speech data

✓ Mainly from conversation

1. Mandarin-English [Lyu+2015]
2. Frisian-Dutch [Heuvel+2016]
3. Spanish-English [Solorio+2008]
4. Turkish-German [Herkenrath+2012]
5. Mandarin-Taiwanese [Lyu+2008]

etc...

# The Overview of Code-switching Data Construction





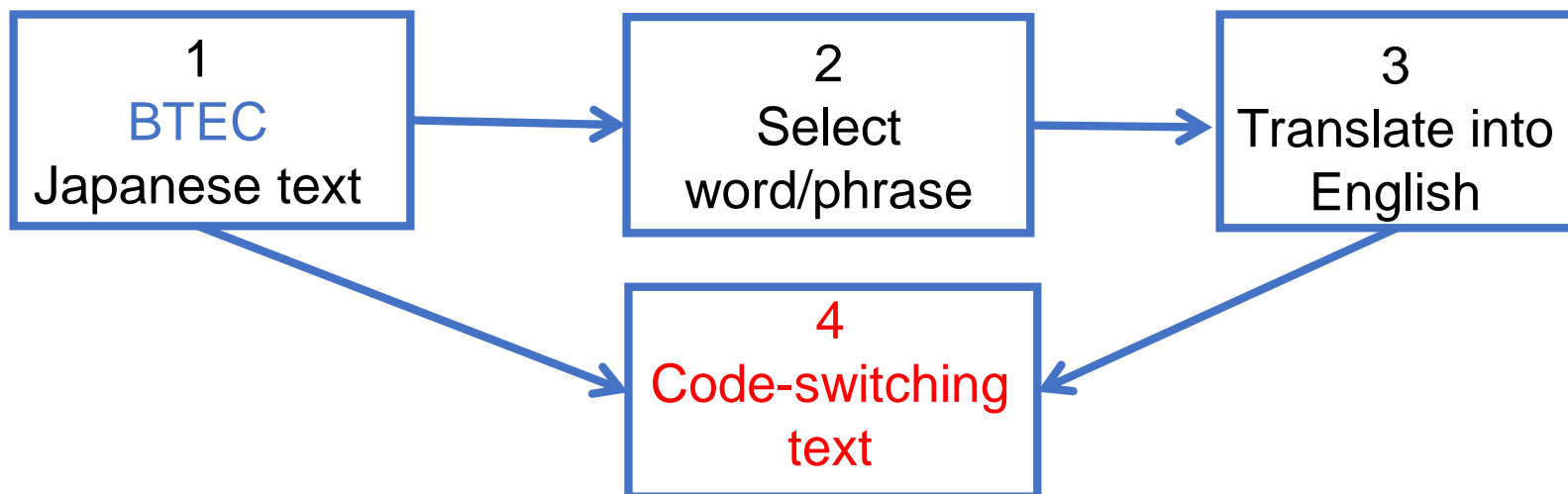
# Data Resources in Text Data Construction

Available data resources:

## Monolingual BTEC text data

- ✓ The ATR Basic Travel Expression Corpus(BTEC)
- ✓ Basic conversations in travel domains

We used Japanese/English BTEC text data called BTEC1,2,3, and 4.



# Data Resources in Speech Data Construction

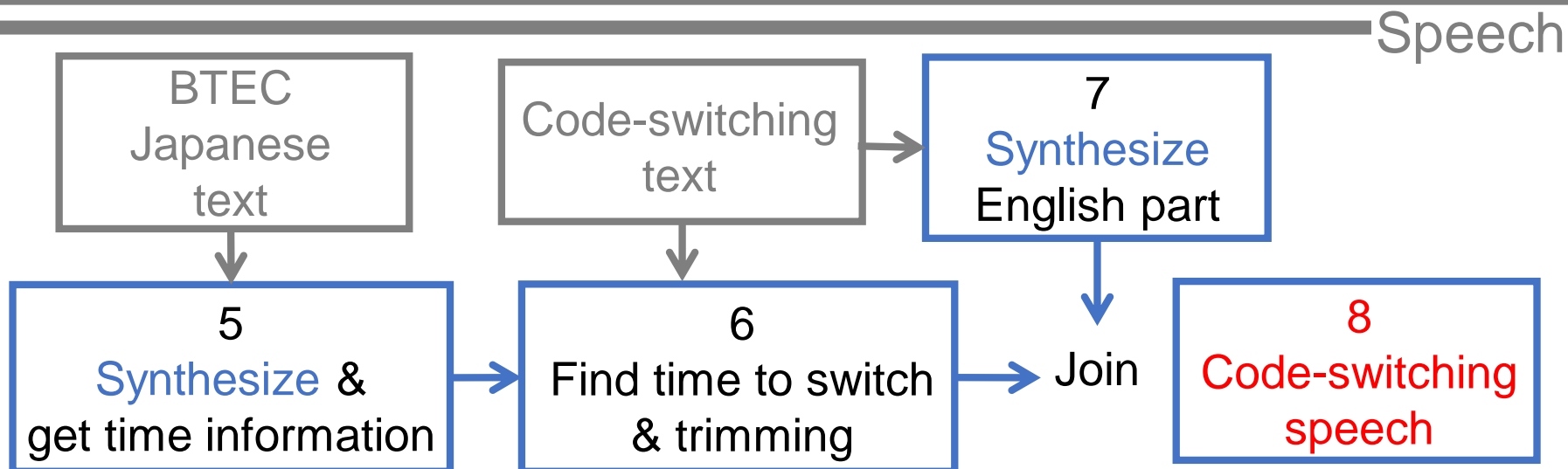
Available data resources:

## Monolingual Japanese and English TTS system

- ✓ Use HMM-based Speech Synthesis System(HTS)
- ✓ Generated based on a [bilingual BTEC speech data](#)

## Bilingual BTEC speech data

- ✓ Originally constructed to emphasize a speech translation study
- ✓ Use the data from 1 bilingual speaker



# Word-level in Text Data Construction

- Intra-sentential word-level code-switching

①

観光 バス の パンフレット は あり ます か  
(*do you have any brochures for the sightseeing bus?*)

※Commonly select katakana characters

# Word-level in Text Data Construction

- Intra-sentential word-level code-switching

①

観光 バスの パンフレット は あり ます か  
(do you have any brochures for the sightseeing bus?)

※Commonly select katakana characters

②

観光 バスの パンフレット は あり ます か  
(do you have any brochures for the sightseeing bus?)

パンフレット ⇒ pamphlet  
by Google Translation  
API

# Word-level in Text Data Construction

- Intra-sentential word-level code-switching

①

観光 バスの パンフレット は あり ます か  
(*do you have any brochures for the sightseeing bus?*)

※Commonly select katakana characters

②

観光 バスの   は あり ます か  
(*do you have any brochures for the sightseeing bus?*)

パンフレット⇒pamphlet  
by Google Translation  
API

③

観光 バスの pamphlet は あり ます か  
(*do you have any brochures for the sightseeing bus?*)

# Phrase-level in Text Data Construction

## ● Intra-sentential phrase-level code-switching

①

Reference [Nakamura+2005]	これ、ウルトラマンティガ <b>か</b> , what do you think this is? (Yes, this (is) Ultraman-Tiga or what do you think this is?)
	それ <b>は</b> , that's not his arm. (Speaking of that, that's not his arm)

ガス入りの炭酸水 **を**二本氷と一緒に持ってきてください  
(Bring me two bottles of carbonic minerals and some ice please)

Trigger is Japanese particles:  
wa, ga, wo, ni, he(e), to, ka, kara, yori, ba, temo, keredo,  
noni, node, kara, nari, nagara, tari, tsutsu

# Phrase-level in Text Data Construction

- Intra-sentential phrase-level code-switching

① ガス入りの炭酸水 **を** 二本氷と一緒に持ってきてください  
(Bring me two bottles of carbonic minerals and some ice please)

② ガス入りの炭酸水 を  
(Bring me two bottles of carbonic minerals and some ice please)

二本氷と一緒に持ってきてください



bring me with two ice

by Google Translation API

# Phrase-level in Text Data Construction

- Intra-sentential phrase-level code-switching

① ガス入りの炭酸水 **を** 二本氷と一緒に持ってきてください  
(Bring me two bottles of carbonic minerals and some ice please)

② ガス入りの炭酸水 を  
(Bring me two bottles of carbonic minerals and some ice please)

二本氷と一緒に持ってきてください



bring me with two ice

by Google Translation API

③ ガス入りの炭酸水 を bring me with two ice.  
(Bring me two bottles of carbonic minerals and some ice please)



# Speech Data Construction

Japanese sentence

English word/phrase

*(Does that include breakfast and dinner?)*

昼食と 夕食 つきですか



dinner

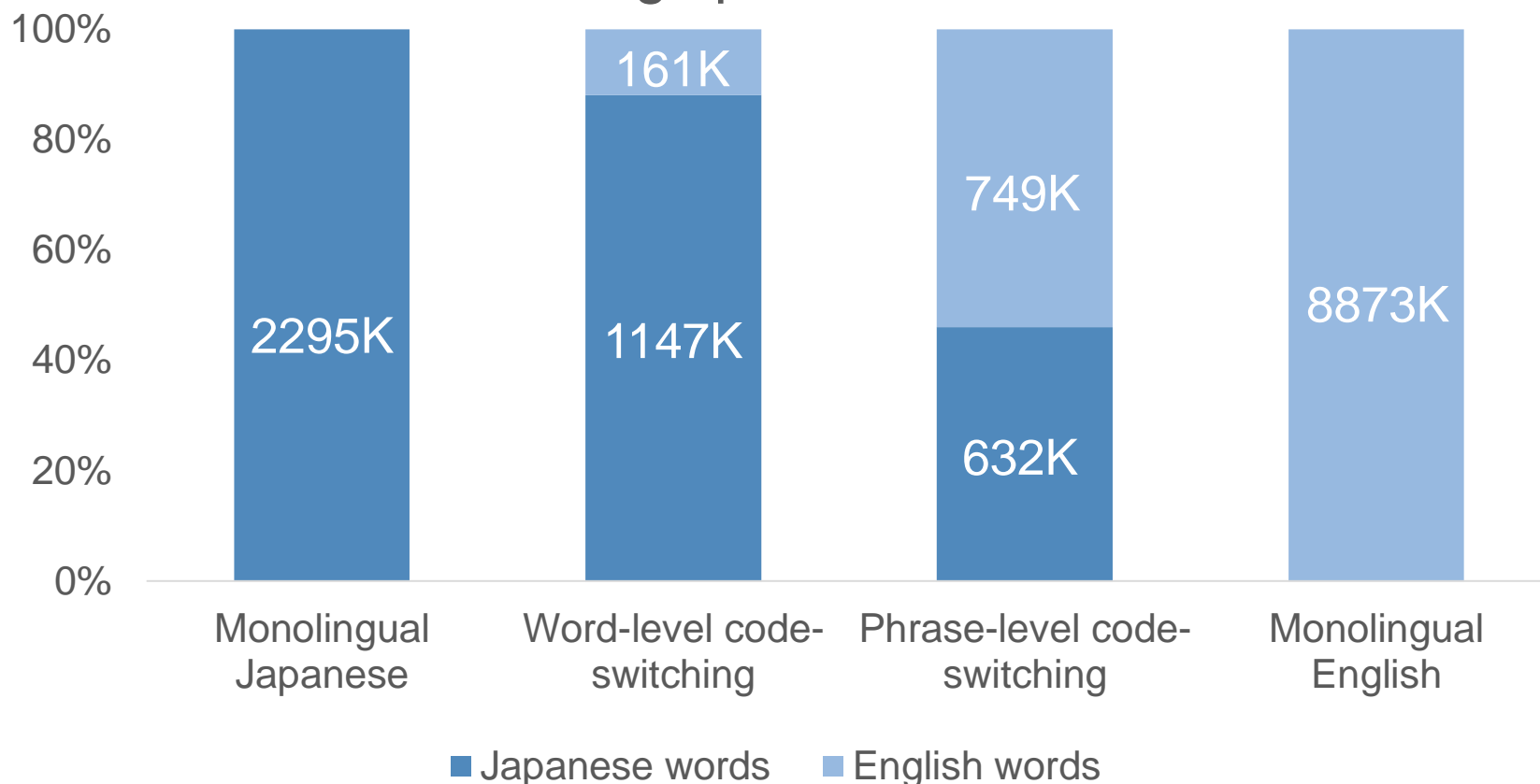


Trimming



# Data Analysis

Statistics of Japanese-English  
code-switching speech utterances



※The number in the bar is the number of words.

# Conclusion

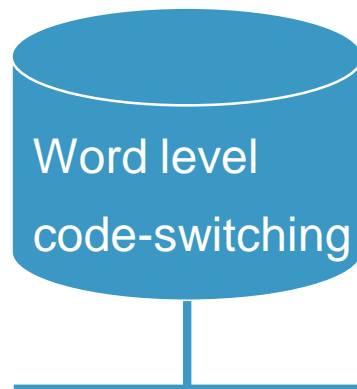
## Problem

For Japanese-English  
code-switching ASR

- ① No large-scale Japanese-English code-switching data exists
- ② Time-consuming and expensive for collecting

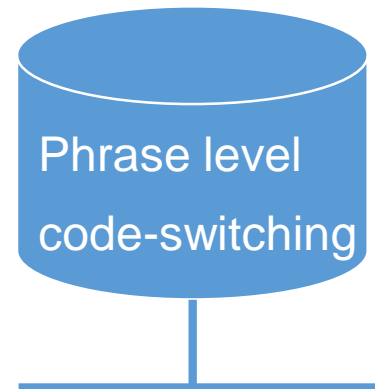
Utilizing  
Japanese and English  
bilingual TTS

**We constructed  
a Japanese-English  
code-switching corpus**



146k

(12% English)

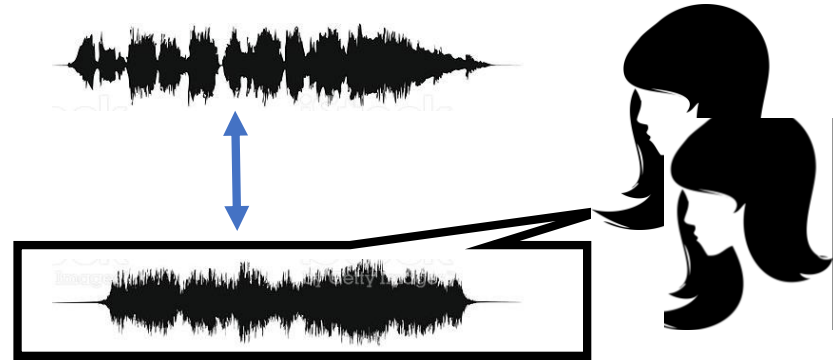


146k

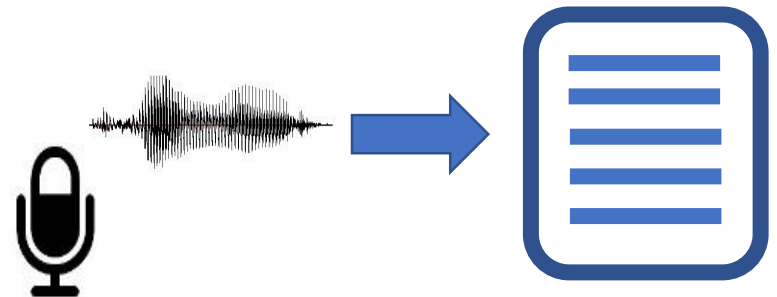
(54% English)

# Future works

- ✓ Investigate the quality by **increasing speakers and comparing with natural code-switching speech**



- ✓ Utilize it to enhance our **speech recognition system** for the bilingual community



# The End

# Appendix

# Reference

- M.Nakamura, “Developing codeswitching patterns of a Japanese/English bilingual child,” in Proceedings of the 4<sup>th</sup> International Symposium on Bilingualism, 2005, pp.1679-1689
- K.Fujimura, “Inevitable Language Outcome: The Use of Code- switching and Code-mixing by Japanese People Living in London, England,” The bulletin of Yasuda women university, 2013, 41, 23-32
- Jeff McSwan, “The architecture of the bilingual language faculty: Evidence from intrasentential codeswitching,” Bilingualism: Language and Cognition, vol.3, no. 1, pp.37-54, 2000.

# Reference

- Japanese Ministry of Health, Labour and Welfare, “Overview of the population statistics in 2016 [in Japanese],” <http://www.mhlw.go.jp/>, 2016.
- Japanese Ministry of Education, Culture, Sports, Science, and Technology, “School basic survey in 2015 [in Japanese],” <http://www.mext.go.jp/>, 2015.
- Japan Student Service Organization, “Survey on japanese student abroad situation in 2016 [in Japanese],” <http://www.jasso.go.jp/>, 2016.











# Reference

N. T. Vu, D. C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E. S. Chng, T. Schultz, and H. Li, “A first speech recognition system for mandarin-english code-switch conversational speech,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4889–4892.

Emre Yilmaz, Henk van den Heuvel, and David van Leeuwen, “Investigating bilingual deep neural networks for automatic recognition of code-switching Frisian speech,” *Procedia Computer Science*, vol. 81, pp. 159 –

# Corpus sample

	Text	This study	Google TTS
Word level	私の baggage が 見つかり ません。 ( <i>I can't find my baggage</i> )		
	これは declaration して ません ね？ ( <i>You haven't declared this have you?</i> )		
Phrase level	私の 荷物 を please look for it. ( <i>Please look for my luggage</i> )		
	申告 する ものが Do you have? ( <i>Do you have anything to declare?</i> )		

# Available Data Resources

## 1. Monolingual BTEC text data

- ✓ The ATR Basic Travel Expression Corpus (BTEC)
- ✓ Basic conversations in travel domains
- ✓ Collected by bilingual travel experts from Ja/En sentence pairs in travel domain phrasebooks

We used Japanese-English BTEC text data called BTEC1,2,3, and 4:

**Table 1.** Basic statistics of BTEC text data

	<b>BTEC 1</b>	<b>BTEC 2</b>	<b>BTEC 3</b>	<b>BTEC 4</b>
# Sentences	172k	46k	198k	74k
# Word tokens	1,174k	341k	1,434k	548k
# Word types	28k	20k	43k	22k

# Available Data Resources

## 2. Bilingual BTEC speech data

- ✓ Constructed to emphasize a speech translation study
- ✓ 1015 Japanese/English sentence pairs were selected from 16,000 BTEC sentences
- ✓ Recording was done with 3 bilingual speakers
- ✓ WAV audio was recorded with a frequency of 16KHz, 16 bits and a single channel

We only use the data from 1 bilingual speaker.

**Table 2.** Sample of bilingual BTEC speech data

Language	Transcription
Japanese	Kamera desu. Suteki na hi ne?
English	That's a camera. Beautiful day, isn't it?

# Available Data Resources

## 3. English and Japanese HMM-Based speech synthesis

Speech synthesis (TTS)	Engine	HMM-based Speech Synthesis System(HTS)
	Text	Japanese-English BTEC
	Speaker	One bilingual speaker
	Training data	960 utterances
	Sampling rate	16kHz
	Frame size/shift	25ms / 5ms
	Feature vector	40-dimensional MGC, log F0
	Vocoder	WORLD vocoder

# SEAME(Mandarin-English) corpus

- South-East Asia Mandarin-English corpus
  - 156 speakers(36.8% are Malaysian while the rest are Singaporean)
  - 192 hours of wav audio records
  - Speaking style: conversation and interview
- In total 110,145 segments
  - 28,655(English)
  - 24,438(Mandarin)
  - 57,052(Code-switch)

# FAME!: Frisian radio broadcast database

- The regional public broadcaster of the province Frisian
- The total duration of the manually annotated radio broadcasts sums up to 18 hours extracted from radio programs almost 50 years
- Topic is about culture, history, literature, sports, nature, agriculture, politics, society and languages.
- 309 identified speakers and 233 unidentified speakers
- Total number of code-switching case is 3837