Parallel Speech Corpora of Japanese Dialects

Koichiro Yoshino[1], **Naoki Hirayama**[2,†], **Shinsuke Mori**[3],
Fumihiko Takahashi[4,†], **Katsutoshi Itoyama**[5], **and Hiroshi G. Okuno**[5,6]
[1]Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, 630-0192, Japan
[2]Industrial ICT Solutions Company, Toshiba Corporation, 3-22, Katamachi, Fuchu, 183-8512, Japan
[3]Academic Center for Computing and Media Studies, Kyoto University, Sakyo, Kyoto, 606-8501, Japan
[4]Yahoo Japan Corporation, Midtown Tower, 9-7-1 Akasaka, Minato-ku, Tokyo, 107-6211, Japan
[5]Graduate School of Informatics, Kyoto University, Sakyo, Kyoto, 606-8501, Japan
[6]School of Creative Science and Engineering, Waseda University, Ohkubo, Shinjyuku, Tokyo 169-0072, Japan
koichiro@is.naist.jp, naoki2.hirayama@toshiba.co.jp, forest@i.kyoto-u.ac.jp,
ftakahas@yahoo-corp.jp, itoyama@kuis.kyoto-u.ac.jp, okuno@aoni.waseda.jp

Abstract

Clean speech data is necessary for spoken language processing, however, there is no public Japanese dialect corpus collected for speech processing. Parallel speech corpora of dialect are also important because real dialect affects each other, however, the existing data only includes noisy speech data of dialects and their translation in common language. In this paper, we collected parallel speech corpora of Japanese dialect, 100 read speeches utterance of 25 dialect speakers and their transcriptions of phoneme. We recorded speeches of 5 common language speakers and 20 dialect speakers from 4 areas, 5 speakers from 1 area, respectively. Each dialect speaker converted the same common language texts to their dialect and read them. Speeches are recorded with closed-talk microphone, using for spoken language processing (recognition, synthesis, pronounce estimation). In the experiments, accuracies of automatic speech recognition (ASR) and Kana Kanji conversion (KKC) system are improved by adapting the system with the data.

Keywords: Speech, Transcription, Dialect, Japanese

1. Introduction

Texts and their read speeches are necessary for natural and spoken language processing (NLP and SLP). Tireless efforts of data collection of speech data and their transcriptions from the early stage of NLP and SLP researches drastically improved accuracies of a variety of NLP and SLP tasks. However, we still do not have enough resources for some minor languages, and it causes less accuracies in minor languages (Kominek and Black, 2006).

Especially, we do not have enough speech data of dialects. The lack of dialect speech data disturbs the use of NLP and SLP applications of dialect speakers. For example, it is very difficult to recognize dialect speeches accurately by using a model trained in common language. Automatic speech recognition (ASR) system is now generally used in public services, for example, taking minutes of national congresses and city councils (Akita et al., 2009). Such applications are expected to be used for not only current formal situations but also more casual situations, for example, taking minutes of courts, and dialect speech recognition is necessary for this purpose.

If we have a small parallel set of dialect and common language, SLP technologies developed in common language can be applied to the dialect rapidly. Japanese dialect speeches are collected and transcribed in linguistics area (National Institute for Japanese Language and Linguistics, 2001 2008), however, the sound data of the corpus is not collected for using on speech processing. It is recorded on analog recording, and the recording situation is not close talk. Even worse, the corpus does not have parallel data of common language in speech, which is important in speech processing.

Furthermore, rapid progression of globalization blur boundaries between dialects. It is not rare to live in several areas in the process of growing, and the speaking style of such person is affected by several dialects. To process such mixed dialect speech, it is necessary to construct parallel corpora between not only common language and dialect but also dialect and dialect. Such corpora realize to build a system that assumes mixed dialect as an input.

In this work, we collected parallel speech data of common language and several dialects in Japanese. We recorded reading speech of 100 Japanese sentences of balanced texts on their vocabulary, which is read by 5 common language (= Tokyo dialect in Japanese) speakers. The common language text consists of segmented words and their pronunciations. We requested several dialect speakers to convert the text to their dialect in the same meaning. They read the text in recording, and annotators transcribed their pronunciations. We prepared 20 dialect speakers from 4 areas, there are 5 speakers from 1 area, respectively. Thus, the resultant resources are parallel corpora of 1 common language and 4 dialects with their read speeches and transcriptions (texts and pronunciations).

We evaluated the collected data in automatic speech recognition (ASR) and Kana Kanji conversion (KKC) (Mori et al., 1999; Takahashi and Mori, 2015) systems. For the ASR evaluation, we constructed dialect speech recognizers that uses machine translation techniques to work on dialect speeches (Hirayama et al., 2015). The ASR accuracies of dialect speeches are improved by adapting recognizers by using the speech and transcription resources based on basic adaptation methods of ASR system. For the KKC evaluation, we constructed dialect Kana Kanji conversion sys-

[†]This work was done when they were at Kyoto University.

tems. The KKC accuracies of dialect are also improved by adapting converters by using the constructed language resources.

2. Related Works

There are some dictionary based approaches for dialect speech processing (Brinton and Fee, 2001; Thomas, 2004; Ramon, 2006; Woods, 1979). The dictionary based approach requires a specialist who analyzes the dialect, and the cost of dictionary construction is very high. Recently, corpus based statistical approaches are popularly used in speech processing. It realizes the system construction if we have a minimal corpus of a dialect.

2.1. Dialect Language Resource in Japanese

The state of the art dialect language resource is published by (National Institute for Japanese Language and Linguistics, 2001 2008). The corpora consist of 20 volumes, records 48 dialects that are spoken by pure dialect speakers (mainly elder people). Each recording has more than 20 minutes and two speakers. The corpora also have transcription in Katakana character and translation in common language (Tokyo dialect) for each speech. The corpora preserve pure Japanese dialect, however, the recording setting is not suitable for spoken language processing (noisy, not closed talk).

The lack of dialect data is also a problem of NLP. Collecting large scale text data becomes easier by the buildup of Web, however, it is still difficult to collect large scale dialect data from Web. We tackled this problem with a conversion of common language resources on Web to dialects based on machine translation techniques (Hirayama et al., 2012). We trained Weighted Finite State Transducer (WFST) that converts common language sentences to dialect pronunciation sequences by using a small parallel corpus of common language in Japanese and Kansai dialect (dialect spoken in Kansai area in Japan) extracted from the database of (National Institute for Japanese Language and Linguistics, 2001 2008). The trained WFST generates large scale simulated dialect corpus to be used for the training of NLP and SLP applications.

2.2. ASR system in Japanese Language

One of the state-of-the-art ASR system in Japanese is Julius decoder (Lee et al., 2001)[1]. Julius packages deep neural network based acoustic model and language model trained from Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa, 2008)[2]. The language model supports to recognize Japanese common language utterances, however, it does not support any dialect speech. The language model is easily updated by adding a training text data, and we tried to update the language model by adding simulated dialect text generated from the WFST based converter (Hirayama et al., 2012).

3. Parallel Dialect Corpora

We selected one common language (Tokyo dialect) and four dialects (Kansai, Kyushu, Tohoku, and San-yo) to construct parallel corpora. The location of each area is shown in Figure 1. These areas are selected to cover major dialects spoken in Japan. Hokkaido (the northernmost area in Figure 1) and Okinawa (south islands depicted on lower right of Figure 1) are not covered by the following reasons.

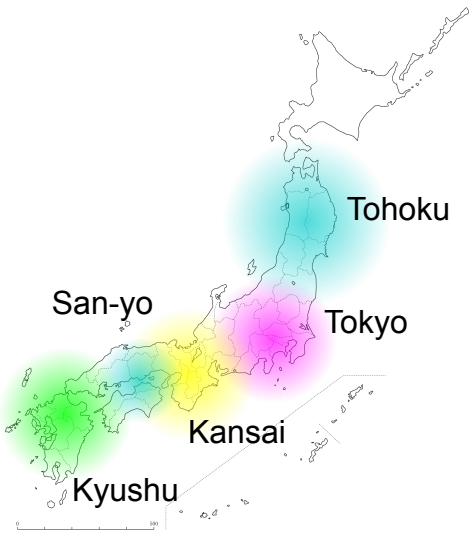

Figure 1: Locations of areas where dialects are spoken.

- Hokkaido is exploited after the Meiji restoration in 1868 by people from various region, thus, the major dialect spoken in Hokkaido is common language.
- Okinawa has different building history from other areas, and the dialect is greatly different from the common language.

3.1. Recording Procedure

The recording procedure of proposed dialect corpora consists of following four steps.

1. We randomly selected 100 sentences of common language to be read by common language and dialect speakers. These sentences are selected from BCCWJ blog category. Honorific expressions are normalized to plain expressions, to make expressing speaker's dialect easier.
2. Dialect speakers convert the selected 100 sentences into their own dialect sentences. We have 5 speakers for each dialect, and each speaker of the same dialect converts sentences by themselves. In other words, we have 5 different transcriptions for each dialect, because expression details depend on the backgrounds of speakers even if their dialect categories are the same. This step is skipped if the speaker is common language (Tokyo dialect) speaker.
3. Speakers read their own converted sentences. Common language speakers read the produced sentences. Speeches are recorded by a close talking microphone.
4. Pronunciations are transcribed into phonemes. We used phoneme set defined in Japanese Newspaper Ar-

[1] http://julius.osdn.jp/
[2] http://pj.ninjal.ac.jp/corpus center/bccwj/

Table 1: Age and gender of speakers (M: males and F: females). Nos. 1–5 correspond to indices of speakers of five dialects.

	#1	#2	#3	#4	#5
Tokyo	38, F	36, M	32, M	25, M	21, F
Kansai	30, F	27, M	24, F	23, M	20, F
Kyushu	28, M	24, F	22, M	20, F	40, M
Tohoku	26, M	24, F	21, F	20, F	26, M
San-yo	49, F	24, M	22, M	21, F	21, M

Table 2: Recording time of each speaker. Nos. 1–5 correspond to indices of speakers of five dialects.

	#1	#2	#3	#4	#5
Tokyo	9:21	8:17	9:40	8:39	9:24
Kansai	9:07	8:06	8:09	7:57	8:09
Kyushu	6:29	8:22	6:53	7:42	8:14
Tohoku	7:05	8:35	8:19	10:24	7:59
San-yo	7:54	8:43	8:02	7:55	7:56

ticle Sentences (JNAS) corpus (Acoustical Society of Japan, 1997)[3].

3.2. Corpus Specifications

The age and the gender of each speaker are summarized in Table 1. Every speaker lived in the area of the dialect until 18 year old. Note that, we did not limit the other conditions. For example, we did not care about the living are after 18 year old, thus, subjects are affected by several dialect that they touched. They also must be affected by TV broadcasting in common language. There are so many factors that may affect the speaking style of each speaker. Table 2 shows total recording time of each speaker. We did not control the speaking speed. Table 3 shows number of total phonemes in pronunciations of each speaker.

3.3. Example of Corpora

We show an example text set of the proposed corpora in Table 4. The original texts is Tokyo, which came from BCCWJ corpus. The original BCCWJ includes annotations of word segment and pronunciation. After the recording, phonemes are transcribed. Pronunciation of each dialect is converted from the transcribed phoneme. Translation examples of other 4 areas are shown in other lines (Kansai, Kyusyu, Tohoku and San-yo). The meanings of sentences are the same, however, there are small differences in minor word choices. For example, “i t e” (there is) in Tokyo is changed as “o tt e” in Kansai, Kyusyu, and San-yo. This is one of a typical change of pronunciation in the west-area of Japan.

4. Applications

4.1. Evaluation in ASR System

We evaluated the collected corpus in ASR systems. Figure 5 shows ASR accuracies of each speaker recognized by a common language ASR system. The language model is trained in Yahoo! Q&A corpus (over 3 million sentences)

[3]http://research.nii.ac.jp/src/

Table 3: Number of phonemes.

	#1	#2	#3	#4	#5
Tokyo	5701				
Kansai	5,525	5,582	5,603	5,687	5,486
Kyushu	5,629	5,848	5,555	5,727	5,721
Tohoku	5,580	5,813	5,512	5,566	5,539
San-yo	5,478	5,481	5,624	5,507	5,485

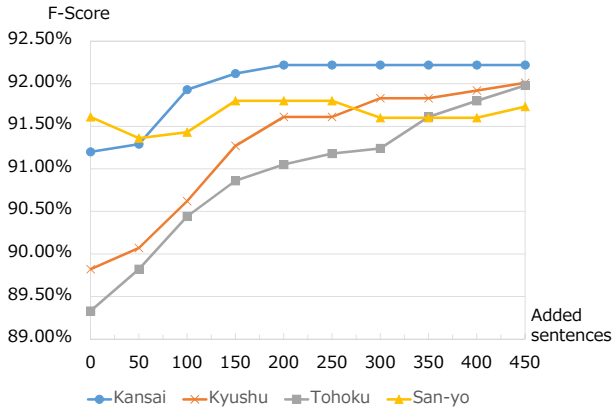

Figure 2: F score of KKC and the added number of dialect sentences to the training set.

(Yahoo Japan Corporation, 2007)[4], and the acoustic model is trained in Corpus of Spontaneous Japanese (CSJ) corpus (National Institute for Japanese Language and Linguistics, 2006)[5] and Japanese Newspaper Article Sentences (JNAS) corpus, respectively. We used Julius decoder.

As shown in the Table 5, the ASR performances are drastically decreased in dialects even if the system works well in common language (= Tokyo dialect), and this result corroborates the importance of maintenance of dialect speech resources. Table 6 shows accuracies of ASR with models adapted by the adaptation technique proposed in our previous work (Hirayama et al., 2015). The accuracies are increased more than 10% in total, and these results verify the importance and usefulness of our parallel dialect speech data.

4.2. Evaluation in KKC System

Kana Kanji conversion (KKC) is a task of converting a given pronunciation sequence to a Kana and Kanji sequence, which is generally used in input method of Japanese. We constructed a statistical Kana Kanji converter (Mori et al., 1999; Takahashi and Mori, 2015) by using BCCWJ corpus as a training set. We picked up 50 utterances from each dialect (50×5) in the constructed dialect corpus as testing sets, and used lest 450×5 sentences as additional training sets. We evaluated the F score (harmonic mean of precision and recall) of KKC of the testing sets. Figure 2 shows that the F scores of KKC depend on the added numbers of dialect sentences to the training set. 0 in horizontal axis means the baseline converter as is, and 450 means fully adapted converter by adding dialect sentences to the

[4]http://www.nii.ac.jp/dsc/idr/yahoo/chiebkr2/Y_chiebukuro.html
[5]http://pj.ninjal.ac.jp/corpus_center/csj/

Table 4: Examples of transcribed corpora.

Area	Speaker	Texts (word/pronunciation)	Transcribed phonemes
Tokyo		アニメ/あにめ に/に は/わ 作画/さくが 監督/かんとく と/と いう/いう 人/ひと が/が い/い て/て 、/NA 毎回/まいかい 話/はなし ごと/ごと に/に 絵/え が/が 違う/ちがう の/の は/わ 、/NA その/その 監督/かんとく さん/さん が/が 前回/ぜんかい と/と 違う/ちがう 人/ひと だ/だ から/から だ/だ そう/そー だ/だ 。/NA	a n i m e n i w a s a k u g a k a N t o k u t o i u h i t o g a i t e sp m a i k a i h a n a sh i g o t o n i e g a ch i g a u n o w a sp s o n o k a N t o k u s a N g a z e N k a i t o ch i g a u h i t o d a k a r a d a s o: d a sp silE
Kansai	#1	アニメ/あにめ に/に は/わ 作画/さくが 監督/かんとく って/って 人/ひと が/が おっ/おっ て/て 、/NA 毎回/まいかい 話/はなし ごと/ごと に/に 絵ー/えー 違う/ちがう ん/ん は/わ 、/NA その/その 監督/かんとく さん/さん が/が 前回/ぜんかい と/と 違う/ちがう 人/ひと や/や から/から や/や ねん/ねん て/て 。/NA	a n i m e n i w a s a k u g a k a N t o k u q t e h i t o g a o q t e sp m a i k a i h a n a sh i g o t o n i e: ch i g a u N w a sp s o n o k a N t o k u s a N g a z e N k a i t o ch i g a u h i t o y a k a r a y a n e N t e sp silE
Kyusyu	#1	アニメ/あにめ に/に は/わ 作画/さくが 監督/かんとく って/って いう/ゆー 人/ひと が/が おっ/おっ て/て 、/NA 毎回/まいかい 話/はなし ごと/ごと に/に 絵/え が/が ちごう/ちごう とっと/とっと は/わ 、/NA その/その 監督/かんとく さん/さん が/が 前回/ぜんかい と/と ちがう/ちがう 人/ひと だ/だ から/から そう/そー だ/だ 。/NA	a n i m e n i w a s a k u g a k a N t o k u q t e y u: h i t o g a o q t e sp m a i k a i h a n a sh i g o t o n i e g a ch i g o u t o q t o w a sp s o n o k a N t o k u s a N g a z e N k a i t o ch i g a u h i t o d a k a r a s o: d a sp silE
Tohoku	#1	アニメ/あにめ に/に は/わ 作画/さくが 監督/かんとく と/と いう/いう 人/ひと が/が い/い で/で 、/NA 毎回/まいかい 話/はなし ごと/ごと に/に 絵/え が/が 違う/ちがう の/の は/わ 、/NA その/その 監督/かんとく さん/さん が/が 前回/ぜんかい と/と 違う/ちがう 人/ひと だ/だ はん/はん で/で だ/だ そう/そー だ/だ 。/NA	a n i m e n i w a s a k u g a k a N t o k u t o i u h i t o g a i d e sp m a i k a i h a n a sh i g o t o n i e g a ch i g a u n o w a sp s o n o k a N t o k u s a N g a z e N k a i t o ch i g a u h i t o d a h a N d e d a s o: d a sp silE
San-yo	#1	アニメ/あにめ に/に は/わ 作画/さくが 監督/かんとく と/と ゆう/ゆー 人/ひと が/が おっ/おっ て/て 、/NA 毎回/まいかい 話/はなし ごと/ごと に/に 絵/え が/が 違う/ちがう ん/ん は/わ 、/NA その/その 監督/かんとく さん/さん が/が 前回/ぜんかい と/と 違う/ちがう 人/ひと だ/だ から/から じゃ/じゃ けえ/けえ と/と 。/NA	a n i m e n i w a s a k u g a k a N t o k u t o y u: h i t o g a o q t e sp m a i k a i h a n a sh i g o t o n i e g a ch i g a u N w a sp s o n o k a N t o k u s a N g a z e N k a i t o ch i g a u h i t o d a k a r a j a k e e t o sp silE
(cf. Translation)		There is a job named animation director who direct drawing of animation, and the reason of each animation episode has different characteristic drawing is that each animation episode has a different animation director.	

training set. This result shows that the constructed dialect resource increases the accuracy of Kana Kanji conversion, and it is also useful for applications of NLP.

4.3. Other Possible Applications

There are some possible applications improved by the proposed dialect data in this paper. Text to speech (TTS) includes some modules that require dataset for the adaptation (Nagano et al., 2005), pronunciation estimation or F0 estimation, and it is one of the most major area that requires parallel speech data. Statistical machine translation (Brown et al., 1993) is also an application that requires parallel data of texts and speeches. The machine translation between the common language and dialect intermediates in conversion of systems for common language to systems for dialect.

Parallel corpora of several dialects benefit who constructs a system that assumes mixed dialect as an input, and it also realizes to estimate the proportion of affected dialect of the user (Hirayama et al., 2015). This technology can be applied for other types properties that affect the speaking style of people, for example, jobs or hobbies. This work indicates the importance of the construction of such parallel

Table 5: Accuracies of ASR system trained in common language.

	#1	#2	#3	#4	#5
Tokyo	84.7	78.1	84.7	82.4	80.0
Kansai	51.6	49.4	61.2	50.9	50.1
Kyushu	44.6	46.0	41.2	57.5	50.4
Tohoku	44.5	33.0	28.9	33.3	58.8
San-yo	66.1	65.5	51.7	54.4	66.3

Table 6: Accuracies of ASR system adapted to the target dialect.

	#1	#2	#3	#4	#5
Kansai	61.4	60.1	67.3	60.3	60.0
Kyushu	49.4	57.5	47.2	66.6	59.9
Tohoku	49.7	42.7	37.9	42.8	67.9
San-yo	81.8	76.1	65.2	66.0	76.1

corpora.

5. Discussion for Annotation

There are some discussions for annotation of the collected corpora. Word segments, accents, and the phoneme set are the points it should be updated in future.

5.1. Word Segment

The original texts from BCCWJ (= common language) have word segmentation annotations. The translated dialect texts are automatically segmented by KyTea (Neubig et al., 2011)[6]. We plan to annotate the word segments of dialects that can be aligned to the original text of common language, however, we still need discussion for annotation standard of word segmentation of dialects.

5.2. Accent

Accents are characteristics of dialect speech in Japanese. This information will benefit to use the corpora in speech synthesis area (Nagano et al., 2005).

5.3. Phoneme Set

In the first version of the corpora, we annotated phonemes with in the phoneme set defined in JNAS. However, JNAS is a common language speech corpus and the phoneme set has some mismatches to real dialect speeches. We need to extend the phoneme set in future version.

5.4. Superposition with Other Annotations

BCCWJ is one of the most popular Japanese language resource, and some previous works tries to annotate several types of annotations: dependency trees (Mori et al., 2014) and predicate argument structures (Komachi and Iida, 2011). We can superpose these annotations with the dialect corpora, to use more higher NLP layers.

6. Conclusion

We constructed parallel speech corpora of common language and several dialects in Japanese. The parallel corpus is used for adapting speech recognition and Kana Kanji conversion, and these applications are improved by the language resource. The constructed parallel speech corpora consist of common language and several dialects, thus, it can be used to model the difference between not only common language and a dialect but also between a dialect and another dialect. It will benefit a variety of tasks in spoken language processing and natural language processing.

7. Acknowledgment

The corpus is created in the project of JSPS KAKENHI Grant No.24220006. This work is also supported by JSPS KAKENHI Grant No.15660505.

8. Use of Corpora

There is no restriction of the usage of the corpora. We request you to cite this paper and the project of JSPS KAKENHI Grant No.24220006 when you publish anything that uses this resource. The resource is provided on the project web page[7].

9. Bibliographical References

Akita, Y., Mimura, M., and Kawahara, T. (2009). Automatic transcription system for meetings of the Japanese national congress. In *10th Annual Conference of ISCA (Interspeech)*, pages 84–87.

Brinton, L. J. and Fee, M. (2001). English in north America. *The Cambridge history of the English language., Cambridge, U.K.: The Press Syndicate of the Univ. of Cambridge*, 6.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Hirayama, N., Mori, S., and Okuno, H. G. (2012). Statistical method of building dialect language models for ASR systems. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 1179–1194.

Hirayama, N., Yoshino, K., Itoyama, K., Mori, S., and Okuno, H. G. (2015). Automatic speech recognition for mixed dialect utterances by mixing dialect language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):373–382.

Kominek, J. and Black, A. W. (2006). Learning pronunciation dictionaries: Language complexity and word selection strategies. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 232–239.

Mori, S., Masatoshi, T., Yamaji, O., and Nagao, M. (1999). Kana-kanji conversion by a stochastic model. *Transactions of Information Processing Society of Japan (in Japanese)*, 7(40):2946–2953.

Nagano, T., Mori, S., and Nishimura, M. (2005). A stochastic approach to phoneme and accent estimation. In *INTERSPEECH*, pages 3293–3296.

Ramon, D. (2006). We are one people separated by a common language. *Viagra, Prozac, and Leeches*, pages 203–206.

[6] http://www.phontron.com/kytea/index-ja.html

[7] http://plata.ar.media.kyoto-u.ac.jp/data/speech/index.html

Takahashi, F. and Mori, S. (2015). Keyboard logs as natural annotations for word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1186–1196.

Thomas, E. (2004). Rural Southern white accents. *A handbook of varieties of English*, 1:300–324.

Woods, H. B. (1979). *A Socio-dialectology Survey of the English Spoken in Ottawa: A study of sociological and stylistic variation in Canadian English*. National Library of Canada.

10. Language Resource References

Acoustical Society of Japan, editor. (1997). *Japanese Newspaper Article Sentences*. Speech Resources Consortium.

Komachi, M. and Iida, R. (2011). Annotating predicate-argument structure and anaphoric relations to bccwj (in japanese). In *Japanese Corpus Workshop*, pages 325–330.

Lee, A., Kawahara, T., and Shikano, K. (2001). Julius—an open source real-time large vocabulary recognition engine. In *Proceedings of Eurospeech*.

Maekawa, K. (2008). Balanced corpus of contemporary written Japanese. In *Proceedings of the 6th Workshop on Asian Language Resources*, pages 101–102.

Mori, S., Ogura, H., and Sasada, T. (2014). A japanese word dependency corpus. In *LREC*, pages 753–758.

National Institute for Japanese Language and Linguistics, editor. (2001–2008). *Database of Spoken Dialects all over Japan: Collection of Japanese Dialects Vol.1-20(In Japanese)*. Kokushokankokai.

National Institute for Japanese Language and Linguistics, editor. (2006). *Corpus of Spontaneous Japanese*. National Institute for Japanese Language and Linguistics.

Neubig, G., Nakata, Y., and Mori, S. (2011). Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Yahoo Japan Corporation, editor. (2007). *Yahoo! QA Corpus*. National Institute of Informatics.