日本語方言における音声対訳コーパスの構築

吉野 幸一郎 [1], 平山 直樹 [2,†], 森 信介 [3], 高橋 文彦 [4,†], 糸山 克寿 [3], 奥乃 博 [3,5]
[1] 奈良先端科学技術大学院 情報科学研究科, 630-0192 生駒市高山町
[2] 株式会社東芝 インダストリアル ICT ソリューション社, 183-8512 府中市片町 3-22
[3] 京都大学 情報学研究科, 606-8501 京都市左京区吉田本町
[4] ヤフー株式会社, 107-6211 港区赤坂 9-7-1 ミッドタウンタワー
[5] 早稲田大学 創造理工学研究科, 169-0072, 新宿区大久保
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

1. はじめに

テキストとその読み上げ音声は、統計的手法に基づく自然言語処理、音声言語処理に不可欠である。自然言語処理・音声言語処理研究の初期から今日に至るまで、こうした言語資源の収集が続けられており、その結果これらの研究領域における様々なタスクの精度が向上している。しかし、使用者が少ない言語における言語資源は未だに十分ではなく、そうした言語における自然言語処理・音声言語処理の精度向上が難しい (Kominek and Black, 2006)。

特に方言に関する音声言語資源は十分ではなく、方言話者が自然言語処理・音声言語処理アプリケーションを利用する際の問題となっている。例えば、標準語音声で学習された音声認識モデルを利用して高精度で方言音声を認識することは非常に難しい。音声認識は、すでに国会や地方議会における議事録作成 (Akita et al., 2009) などの公共サービスとして利用されている。方言を利用できないことは、こうしたサービスが求められる状況を考えると十分ではなく、また裁判所における議事録作成など、より市民が利用する場面において方言への対応は不可欠である。

標準語のために開発された音声言語処理技術を方言に適応するためには、方言と標準語の対訳コーパスを用意することが効果的である。日本語方言音声は言語学の分野で収集・書き起こしが行われている (National Institute for Japanese Language and Linguistics, 2001 2008) ものの、音声言語処理のために収集されたものではなく、接話マイクでデジタル録音を行ったものは存在しない。また、対応する標準語音声まで含めると、そうした対訳音声は存在しない。

さらに、人・社会のグローバル化が方言の境界を曖昧にしている。例えば、成長の過程で複数の地域に居住することは珍しいことではなく、そうした人の発話は複数の方言の影響を受けることになる。こうした混合方言を処理するためには、単一方言と標準語の対訳だけではなく、複数の方言同士の対訳も必要となる。このような標準語と方言のような異なる話し方の対訳を用いる枠組みは、方言以外にも個人の属性、例えば職業や生育環境などが話し方に与える影響をモデル化することができ、またこうした個人の属性の逆推定に用いることができることが期待される。

そこで、本論文では日本語における標準語と複数の方言音声対訳を収集した。具体的には、語彙の観点から均衡化された日本語文 100 文を用意し、5 人の標準語話者（東京方言話者）の読み上げ音声を収録した。この読み上げ文は、単語境界および読みの情報を含む。また複数の方言話者に、これらの 100 文を同じ意味の方言文に訳してもらった。その後、訳した方言の音声を収録し、その音素についてアノテーションを行った。この方言は、4 地域各話者、合計 20 話者から収録した。つまり、1 種類の標準語と、4 種類の方言の音声対訳と、そのテキスト・音素の書き起こしが存在する。

これらの収録データを、音声認識およびかな漢字変換システムによって評価した。まず音声認識の評価のために、機械翻訳の技術を用いて方言音声認識器を構築した (Hirayama et al., 2015)。音声認識実験を行った結果、収録した方言音声データによって適応した音声認識器は、未適応の音声認識器よりも高い精度を示した。また、かな漢字変換の評価のために、方言のかな漢字変換システムを構築した。かな漢字変換も音声認識と同様に、収集した方言コーパスで適応することにより精度が向上した。

2. 関連研究

方言音声言語処理の研究では、辞書を用いた手法がいくつか試みられてきた (Brinton and Fee, 2001; Thomas, 2004; Ramon, 2006; Woods, 1979)。辞書ベースの手法は各方言の分析を行う専門家を必要とし、また構築のコストも高い。そこで近年、音声言語処理ではコーパスを用いた統計的手法が一般的になっており、少量でも方言のコーパスを用意することによってこうした手法を適用することができる。

方言データの不足は、自然言語処理の分野でも問題になっている。Web の発展に伴い大規模なテキストデータを収集することは容易になっているが、Web から特定の方言データを大規模に収集することは難しい。そこで我々は、機械翻訳の手法を使って Web 上の標準語言語資源を方言に変換することによって、疑似的に方言のテキストコーパスを生成している (Hirayama et al., 2012)。この研究では、重み付き有限状態トランスデューサを用いて標準語文を方言発音列に変換する。この際に、標準語と関西弁（日本の関西地方で話される方言）の対訳コーパスを利用している。この重み付き有限状態トランスデューサによって生成された大規模疑似コーパスは、様々な言語処理・音声言語処理の研究に用いることができる。

† 本研究に対する主要な貢献は京都大学在学中になされた。

Table 1: 話者の年齢、性別、18 歳までの居住地（番号は各方言に対する話者番号の対応を示す）

	#1	#2	#3	#4	#5
標準語	38 歳女性 東京都多摩市	36 歳男性 東京都葛飾区	32 歳男性 東京都小平市	25 歳男性 埼玉県さいたま市	21 歳女性 東京都大田区
関西	30 歳女性 大阪府四条畷市	27 歳男性 兵庫県神戸市	24 歳女性 大阪府大阪市	23 歳男性 大阪府泉佐野市	20 歳女性 兵庫県姫路市
九州	28 歳男性 熊本県熊本市	24 歳女性 熊本県山鹿市	22 歳男性 福岡県糟屋郡	20 歳女性 福岡県糟屋郡	40 歳男性 福岡県福岡市
東北	26 歳男性 青森県黒石市	24 歳女性 青森県弘前市	21 歳女性 山形県鶴岡市	20 歳女性 青森県青森市	26 歳男性 青森県黒石市
東山陽	49 歳女性 広島県福山市	24 歳男性 広島県福山市	22 歳男性 広島県福山市	21 歳女性 岡山県岡山市	21 歳男性 広島県東広島市

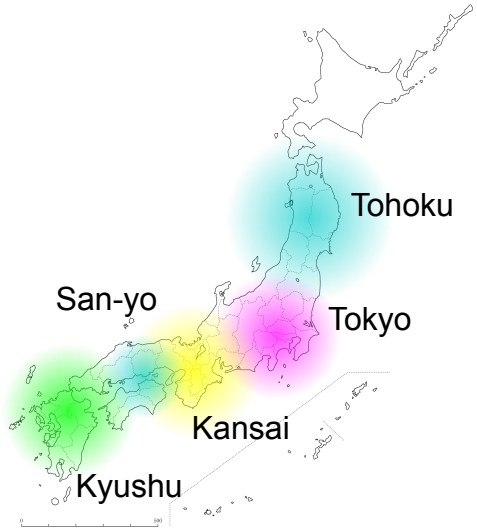

Figure 1: 各方言が話されている地域

Table 2: 各話者の発話収録時間（分:秒, 番号は各方言に対する話者番号の対応を示す）

	#1	#2	#3	#4	#5
標準語	9:21	8:17	9:40	8:39	9:24
関西	9:07	8:06	8:09	7:57	8:09
九州	6:29	8:22	6:53	7:42	8:14
東北	7:05	8:35	8:19	10:24	7:59
東山陽	7:54	8:43	8:02	7:55	7:56

Table 3: 各話者の収録音素数（番号は各方言に対する話者番号の対応を示す）

	#1	#2	#3	#4	#5
標準語			5701		
関西	5,525	5,582	5,603	5,687	5,486
九州	5,629	5,848	5,555	5,727	5,721
東北	5,580	5,813	5,512	5,566	5,539
東山陽	5,478	5,481	5,624	5,507	5,485

3. 方言対訳コーパス

本研究では標準語として広く用いられる東京方言を基準とし、それ以外に話者の多い関西、九州、東北、東山陽の 4 方言を対訳コーパスの構築対象として選定した。これらの方言が話される地域を図 1 に示す。北海道と沖縄は、以下の理由により今回は対象から除外した。

- 北海道は明治維新以降開拓された結果、多様な地域からの入植があり北海道で主に話される言語は標準語もしくは標準語に近い表現である。
- 沖縄は本土と異なる歴史的成り立ちを持ち、方言の表現も標準語と比較して著しく異なる。

3.1. 収録条件

方言の収録は以下の 4 つの手順に従って行った。

1. 100 文の標準語からなる読み上げ文を用意した。この文は、日本書き言葉均衡コーパス（BCCWJ）[1] (Maekawa, 2008) のブログエントリからランダムに選択した。このうち敬語表現については、方言の表現に変換しやすいよう普通表現へと変換した。
2. 方言話者にこれらの 100 文を、彼ら自身の方言へと変換してもらった。各方言カテゴリには 5 話者がいるが、それぞれの話者は個別に自身の方言に対する変換を行った。つまり、各方言に対してバリエーションが異なる 5 種類の変換結果が存在する。これは、たとえ方言の大分類が同じであっても、その表現の詳細は話者の育った地域、生育環境、などに大きく依存するからである。標準語（東京方言）に関しては、この手順は行わない。
3. 話者はそれぞれ自身が変換した文を読み上げる。標準語話者（東京方言話者）は、変換前の標準語文をそのまま読み上げる。音声は接話マイクによって録音する。
4. 方言の発音を音素へと書き起こす。今回は日本音響学会 新聞記事読み上げ音声コーパス（JNAS）で定義されている音素セットを用いて書き起こしを行った。[2]

3.2. コーパススペック

各話者の年齢、性別、および詳細な出身地を表 1 に示す。各話者は、それぞれの方言カテゴリの地域に、生まれてから 18 歳まで住んでいる。表 2 に話者ごとの合計収録時間を示す。収録の際、話速の制御は行っていない。表 3 に話者ごとに収録された合計音素数をまとめる。

[1] http://pj.ninjal.ac.jp/corpus_center/bccwj/

[2] http://research.nii.ac.jp/src/JNAS.html

Table 4: 標準語音声認識システムを用いた場合の各話者音声の認識精度

	#1	#2	#3	#4	#5
標準語	84.7%	78.1%	84.7%	82.4%	80.0%
関西	51.6%	49.4%	61.2%	50.9%	50.1%
九州	44.6%	46.0%	41.2%	57.5%	50.4%
東北	44.5%	33.0%	28.9%	33.3%	58.8%
東山陽	66.1%	65.5%	51.7%	54.4%	66.3%

Table 5: 目標方言に適応した音声認識システムを用いた場合の各話者音声の認識精度

	#1	#2	#3	#4	#5
関西	61.4%	60.1%	67.3%	60.3%	60.0%
九州	49.4%	57.5%	47.2%	66.6%	59.9%
東北	49.7%	42.7%	37.9%	42.8%	67.9%
東山陽	81.8%	76.1%	65.2%	66.0%	76.1%

4. アプリケーション応用

4.1. 方言音声認識

収録された方言音声データを、音声認識によって評価した。図 4 に、標準語のために構築された音声認識システムを用いた場合の各方言音声に対する認識精度を示す。音声認識精度は以下の式で求める。

$$\frac{H - I}{N} \tag{1}$$

ここで、H は正解した単語数、I は誤挿入された単語数、N は正解中の単語の総数である。音声認識のための言語モデルは Yahoo!知恵袋コーパス [3] から、音響モデルは日本語話し言葉コーパス（CSJ）[4] および日本音響学会 新聞記事読み上げ音声コーパス（JNAS）を用いて学習した。音声認識デコーダには Julius[5] を用いた。表 4 に、標準語モデルを用いた場合の各方言音声の音声認識精度を示す。この結果から、標準語（東京方言）に対して適切なモデルであっても、方言に対しては音声認識精度が著しく低下することがわかり、方言などの言語表現が異なる発話に対する音声言語資源を整備することの重要性が見て取れる。続いて、表 5 に (Hirayama et al., 2015) の手法でモデルを方言に対して適応した場合の音声認識精度を示す。これらのの結果から音声認識精度が、全体として 10% 以上向上していることがわかり、今回提案する方言の音声対訳コーパスの有用性が示されていると言える。

4.2. かな漢字変換

かな漢字変換とはかな文字系列をかな漢字混じり系列に変換するタスクで、日本語入力一般に使われる。今回は統計的かな漢字変換器 (Mori et al., 1999; Takahasi and Mori, 2015) を構築し、BCCWJ コーパスを学習データとした。このかな漢字変換に対し、方言の読み・書き起こしを適応データとして用いることで、方言に対する適応の効果を見る。各方言について 50 文（50×5 方言）をテストデータとし、適応データとして残りの 450×5 文

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

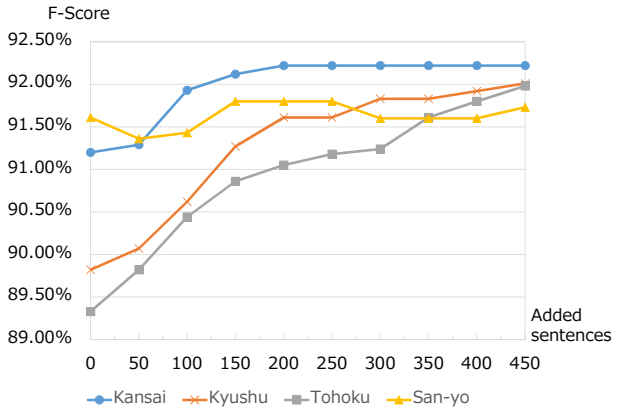

Figure 2: かな漢字変換の F 値と方言適応のため学習データに加えた適応データの文数

を用いた。評価としては、テストデータに対するかな漢字変換の F 値（適合率・再現率の調和平均）を用いた。図 2 に、適応データ数とかな漢字変換の変換精度（F 値）のグラフを示す。横軸の 0 は適応しないベースラインのかな漢字変換器であり、450 は適応データを全て利用してかな漢字変換の適応を行ったものである。この結果から、構築した方言言語資源が、かな漢字変換の精度を向上させることがわかり、自然言語処理アプリケーションに対しても有用であることがわかる。

4.3. その他の可能なアプリケーション

これまでに試した音声認識やかな漢字変換の他にも、本稿で提案する言語資源が有効である場合はいくつか考えられる。特に音声合成（text-to-speech: TTS）は、発音推定や F0 の推定など、方言に対する適応を必要とする様々なモジュールがあり、こうした音声対訳コーパスを必要とする最も大きな分野の 1 つである。機械翻訳においても同様に、こうしたテキストと音声が揃った対訳データの存在は重要である。また、標準語と方言に対して機械翻訳の技術を適用することで、既存の標準語向けに開発された自然言語・音声言語アプリケーションを活用することができる。

複数方言の対訳コーパスは、多様な混合方言が入力として想定されるようなシステムの構築に有用であり、また逆向けに発話者の属性比率推定をすることもできる (Hirayama et al., 2015)。こうした属性推定の技術は、方言に限らず人の様々な属性、例えば仕事や趣味などから現れる言語表現に対しても利用することができ、本研究は今後もこうした対訳コーパスを構築していくことの有用性を示している。

5. まとめ

本稿では標準語と複数の方言の音声対訳コーパスを構築し、それらの有用性をいくつかのアプリケーションで示した。具体的には、音声認識とかな漢字変換の方言適応を試し、提案する言語資源によってこれらのアプリケーションの精度が向上することが確認された。構築された音声対訳コーパスは単一の標準語と複数の方言から成り立っており、標準語–方言間の対比だけでなく、方言同士の比較を扱うことができる。この性質は、多様な音声言語処理・自然言語処理のタスクに対して寄与することが期待される。

6. コーパスの利用に際して

本コーパスの Web サイト [6] に詳細を掲載する。本コーパスの利用に際しては制限はないが、本論文または文献 (Hirayama et al., 2015)、および JSPS 科研費 No.24220006 に対する引用を条件とする。

7. 謝辞

本コーパスは JSPS 科研費 No.24220006 の費用によって構築された。また、JSPS 科研費 No.15660505 の援助を受けた。

8. 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.

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.

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

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.

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.

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

Takahasi, 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.

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