関西方言を対象とした形態素解析用辞書の拡張

小木曽 智信 [12]　尹 熙洙 [21]　王 竣磊 [13]　岡田 純子 [1]
[1] 人間文化研究機構 国立国語研究所　[2] 総合研究大学院大学 先端学術院
[3] 東京大学 人文社会系研究科
{togiso, gs20233504, wang-junlei, jun-okada}@ninjal.ac.jp

概要

現代語用の UniDic をベースとして関西方言を対象とした形態素解析用の辞書「関西弁 UniDic」の拡張を行い、方言特有の見出し語の追加、特に機能語の品詞や活用形の整備を行った。また、短単位データとして整備した関西方言用の学習用コーパスを大幅に充実させ、これらのデータを用いて MeCab 用の辞書を作成した。さらに、各種の学習用コーパスを組み合わせて解析精度の検証を行った。これにより従来より高い精度で、広域の関西方言の会話書き起こしテキストを解析することを可能にした。

1　はじめに

自然言語処理タスクの多くが End-to-End の処理で行われるようになった今日でも、形態素解析は言語研究にとって研究対象のデータを分析するための重要な手がかりが得られる手段として重要である。形態素解析済みのコーパスは、言語研究において欠かすことのできないものとなっている。国立国語研究所が公開している各種コーパスは、短単位という統一された基準の下でアノテーションが行われており、現代語の書き言葉、話し言葉から各時代の歴史的な日本語資料までが揃っており、日本語の研究に幅広く使われている。こうしたコーパスの整備のために、各種の日本語のバリエーションのための形態素解析用辞書 UniDic シリーズが整備・公開されてきた[1)]。UniDic はもともと日本語研究を目的に設計されたものであり、斉一な見出し語単位、階層化された見出し語構造や語種情報の付与などの言語研究に適した特徴を持っている。また、『日本語歴史コーパス』（CHJ）の整備に伴って各時代別の UniDic が整備され、概ね各時代の資料の実用的な精度での解析が可能になっている [1]。

一方で、各地の方言のテキストを解析することができる辞書はなく、その整備が望まれている。こうした背景から、発表者らは現代語用の UniDic[2] をベースとして、関西方言を対象とした形態素解析用の辞書の開発を行ってきた（小木曽ほか 2024）[3]。ここで開発した関西弁 UniDic を用いることで、関西方言の書き起こしテキストを、既存の話し言葉用の UniDic で解析した場合よりも高い精度で解析することが可能になった。しかし、学習用コーパスのサイズが小さいこと、見出し語（特に方言特有の機能語）の整備が不十分であることから、標準語を対象とした形態素解析の精度と比べると未だ十分とは言えないものであった。

そこで、本研究では、新たに整備した「短単位版関西弁コーパス」[4] を学習用のコーパスとして用い、関西方言に見られる助詞・助動詞をはじめとする見出し語を追加整備することで、より高い精度の解析を行うことを目指した。

2　利用したコーパス

関西方言を収録したコーパスとして『関西弁コーパス』（KVJ）[5] [2)] がある。このコーパスは関西の大学生が家族や親しい知り合いと行ったインタビューを収集し、書き起こしたテキストである。公開されているデータは形態素解析が行われているが、IPADIC による解析結果であり、修正が施されているものの完全ではない。収録データは、KSJ（大阪・神戸都市圏）、KYT（京都）、TKC（多可町・西脇市）、RGS（関西に住んでいる留学生）の 4 パートに分かれている。

小木曽ほか（2024）ではこの一部を短単位版として整備しなおして関西弁 UniDic ver.0.5 の学習用コーパスとして活用した。この時点で用意できた人手修正済みのコーパスは KSJ 20 ファイル（約 18.5 万語）であったが、今回、新たに「短単位版関西弁コーパス」が整備され、人手修正済みのデータが 74 ファ

1）https://clrd.ninjal.ac.jp/unidic/download_all.html

2）https://sites.google.com/view/kvjcorpus/

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

イル（77.5 万語）整備された。また、KSJ 38 ファイルだけでなく、KYT 22 ファイル、TKC 14 ファイルを含み、地域的なバリエーションも確保された。その詳細は尹ほか（2025）[4] を参照されたい。

このほかに利用できる関西方言の短単位のコーパスとして、やや古い時代の関西の資料を対象とした CHJ「江戸時代編 I 洒落本」[6][3] と、同じく「明治・大正編 VI 落語 SP 盤」[7] [4] がある。前者は江戸時代（18 世紀）に刊行された会話を主体とする戯作小説である「洒落本」をコーパスにしたもので、上方語の資料としては京都 10 作品、大坂 10 作品が収録されている。一方、後者は明治期に録音された SP 盤の落語を書き起こしたもので、大阪の 51 作品（落語家 10 人）の約 4.7 万語が収録されている。

また、関西方言ではないが、日常的な話し言葉を収録した「日本語日常会話コーパス」（CEJC）[8][5] が公開されている。小木曽ほか（2024）では、このコアデータを追加用の学習用コーパスとして効果を上げており、今回もこれを利用した。以上のコーパスをまとめると表 1 のようになる。

コーパス名	略称	語数
「関西弁コーパス」 74 ファイル	KVJ	77.5 万語
「日本語歴史コーパス」洒落本+落語 SP 盤	CHJ	7.0 万語 + 4.7 万語
「日本語日常会話コーパス」コアデータ	CEJC	24.5 万語

表 1　利用したコーパス

3　見出し語の拡充・整備

小木曽ほか（2024）の関西弁 UniDic ver.0.5 のために整備された学習用コーパスは、「関西弁コーパス」のなかでも主として阪神都市圏のデータとなっており、そこに現れる方言の語彙は、既存の現代語用 UniDic にすでにある程度登録されていた。今回、新たに京都方言ならびに播磨方言のコーパスを整備するにあたり、辞書側では、語彙のさらなる拡充や活用の整理などが必要となった。

3.1　語彙の拡充

Ver.0.5 までに追加した関西の地名や固有名詞のほかに、普通名詞や形容詞、助詞などの新規登録も行った。例えば「ひょっとで（柏餅）」「りょんりょん（竜王の舞）」など、平仮名で表記されることの多い、民俗に関連する語彙は、UniDic に新規の語彙素として追加している。「しらこすぎる」に出現している形容詞「しらこい」は「わざとらしい」「明らかに嘘をついている」などの意味を有し、これも UniDic に新たに追加した。播磨方言「助けてくれっこ？」に現れる、疑問や感嘆を表す「こ」は、終助詞として新たに登録した。

動詞の拡充は既存の語彙素を活用した形で行った。例えば「給料なんぼもうとる」の「もう」は既存語彙素「貰う」の連用形ウ音便として、「やってかれへんちゃん？」の「ちゃん」は既存語彙素「違う」の異語形「ちゃう」の終止形撥音便として、新たに登録している。語彙を拡充していく中で、標準語の「それで」「そして」「そしたら」に相当する、関西方言の接続表現はとりわけバリエーションに富んでおり、特に注意が必要であった。これら接続表現の形式や機能について議論を行い、『現代日本語書き言葉コーパス』（BCCWJ）所収の標準語データとの整合性を考えつつ、既存の語彙素の異語形とするか、新規の語彙素とするかを認定した上で整備を行った。下記はその例である。

- ほんで・へで ⇒ 既存語彙素「それ」の異語形＋格助詞「で」
- ほんなら・ほんだら ⇒ 既存語彙素「それ」の異語形＋助動詞「だ」仮定形
- ほて・へて ⇒ 既存語彙素「そして」の異語形
- ほいたら・へたら ⇒ 新規語彙素「そしたら」の諸語形
- ほた・へた ⇒ 既存語彙素「ほな」の異語形
- ほなったら ⇒ 既存語彙素「そう」の異語形＋動詞「なる」連用形＋助動詞「た」仮定形

3.2　活用の整理

語彙の拡充のみならず、関西方言全体にわたって、アスペクト形式や、敬意を表す待遇表現などを整理した。また、Ver.0.5 開発時に課題となっていた点、例えば「へん」上接動詞の活用や省略形、短縮形などについても、方言研究の成果を取り入れつつ、整備を行った。下記はその例である。

アスペクト形式

関西地方全域にみられる形式として下記がある。

3）https://clrd.ninjal.ac.jp/chj/edo.html#share
4）https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#rakugo
5）https://www2.ninjal.ac.jp/conversation/corpus.html

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

- 運転してんねん／手袋してる人 ⇒ 標準語と同様に助動詞「てる」
- 整形しとるやん／輸入しとる水 ⇒ 連用形接続の助動詞「とる」

また兵庫県を中心に次の形式が見られる。

- 後悔しとうけど／結婚しとう人 ⇒ 上記「とる」の特殊活用形とする
- 仕事しよるから／運転しよる人 ⇒ 連用形接続の助動詞「よる」

待遇表現

- 言ってくれやる（大阪） ⇒ 連用形接続の助動詞「やる」
- 言いはる（阪神）／言わはる（京都） ⇒ 連用形または未然形接続の助動詞「はる」
- 言うてや（播磨） ⇒ 接続助詞「て」+助動詞「や」
- 言うたった（播磨） ⇒ 連用形接続の助動詞「たる」の連用形＋助動詞「た」とする。

否定の助動詞「へん」上接の動詞

- 行かへん ⇒ 五段動詞「行く」の未然形
- 行けへん（可能の意でない場合） ⇒ 五段動詞「行く」の未然形
- 行けへん（可能の意の場合） ⇒ 下一段動詞「行ける」の未然形
- 行かれへん ⇒ 五段動詞「行く」未然形＋助動詞「れる」未然形

上記の認定基準により、上接語が同じエ段の形として出現しても意味によってその活用型（五段か下一段か）を区別する必要が生じている。

活用語の省略形・短縮形

- 食べたあかん ⇒ 助動詞「た」の仮定形-省略
- 行きたない ⇒ 助動詞「たい」の連用形-省略
- かなんがな ⇒ 動詞「敵う」の未然形-省略
- 電気代がたこつく ⇒ 形容詞「高い」の連用形ウ音便「たこう」の短縮形とする
- 新幹線つこて行って ⇒ 動詞「使う」の連用形ウ音便「つこう」の短縮形とする

このほかに、例えば「何しょん」「あんにゃ」のような融合形や、「入れやへん」「居やはる」などの「や」挿入形などが課題となる。

4 解析精度

整備した見出し語データをもとに、MeCab[6] を用いて形態素解析用の UniDic 短単位辞書を作成した。

学習用のコーパスは、上述の人手修正済みの短単位版関西弁コーパス（KVJ）から評価データ 8 ファイルを除いた 66 ファイルを用意し、さらに CHJ の京都・大阪データ、CEJC のコアデータと組み合わせて用いて精度評価を行った。各コーパスの語数は表 2 の通りである。

	KVJ	CHJ	CEJC
語数	685625	117177	244059

表 2 学習用コーパス語数

評価データは、KVJ から、KSJ 4、KYT 2、TKC 2 の 8 ファイルを選んだ。

精度評価は小木曽ほか（2024）と同様に、UniDic の階層構造に対応した 4 レベルで行った。Lv.1 は語（短単位）の境界の認定が正しく行われているかを見るもの、Lv.2 はこれに加えて UniDic の語形の階層の品詞・活用型・活用形の認定が正しく行われているかを見るもの、Lv.3 はこれらに加えて UniDic の語彙素の認定が正しく行われているかを見るものである。Lv.3 は例えば「金(カネ)」を語種が異なる見出し語「金(キン)」ではなく「金(カネ)」と正しく認定できているかを評価するものである。Lv.4 は上記に加えて UniDic の発音形の認定が正しく行われているかを見るものである。例えば見出し語「何(ナニ)」を「ナニ」ではなく「ナン」に正しく認定できているかまでを評価する。

評価結果を表 3 に示す。一般公開されている話し言葉用 UniDic（UniDic-CSJ v202302）をベースラインとし、関西弁 UniDic ver.0.5 を比較対象として、新たに作成した辞書との精度比較を行った。新しい辞書は、学習用コーパスの組み合わせで、KVJ のみ、KVJ に CHJ 京都・大阪を加えたもの（KVJ+CHJ）、KVJ に CEJC コアデータを加えたもの（KVJ+CEJC）、KVJ に CEJC と CHJ を加えたもの（ALL=KVJ+CEJC+CHJ）の 4 種類を作成した。なお、関西弁 UniDic ver.0.5 は、もっとも精度の高かったもの（ALL）のみを比較対象として挙げた。

新たに作成した関西弁 UniDic はどの点でも従来の精度を上回っており、今回の見出し語整備と学習用コーパスの追加が効果的であったことを示している。95%を上回った段階でさらに 1%ポイント以上

6） https://taku910.github.io/mecab/

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

		UniDic-CSJ	関西弁 UniDic	Proposed 関西弁 UniDic			
		v202302	0.5 ALL	KVJ	KVJ+CHJ	KVJ+CEJC	ALL
Lv.1 (境界)	Precision	98.11%	98.65%	99.18%	**99.19%**	99.18%	99.18%
	Recall	98.02%	98.69%	99.08%	99.08%	99.10%	**99.11%**
	F1	98.07%	98.67%	99.13%	**99.14%**	99.14%	99.14%
Lv.2 (品詞)	Precision	94.80%	96.09%	97.28%	97.29%	**97.31%**	97.29%
	Recall	94.71%	96.14%	97.17%	97.18%	**97.23%**	97.21%
	F1	94.75%	96.11%	97.22%	97.23%	**97.27%**	97.25%
Lv.3 (語彙素)	Precision	94.41%	95.81%	97.00%	97.02%	**97.06%**	97.04%
	Recall	94.32%	95.85%	96.90%	96.91%	**96.99%**	96.97%
	F1	94.36%	95.83%	96.95%	96.97%	**97.03%**	97.01%
Lv.4 (発音形)	Precision	93.52%	95.54%	96.77%	96.80%	96.82%	**96.83%**
	Recall	93.43%	95.58%	96.66%	96.70%	96.74%	**96.75%**
	F1	93.47%	95.56%	96.72%	96.75%	96.78%	**96.79%**

表 3　解析精度

の差をつけており、大きな精度向上があったと言えよう。解析精度は標準語の話し言葉向けの UniDic と遜色ないものになってきている。

しかし評価データと同じ KVJ のみで学習するよりも、やや異質なコーパスを加えてサイズを大きくした方が高精度となることから、まだ KVJ の学習用コーパスのサイズが十分ではないと考えられる。なお、ALL が全てにおいて最善であった Ver.0.5 とは異なり、語彙素認定の精度では僅差だが KVJ+CEJC が ALL の精度を上回った。これは CHJ に伝統的な（コテコテの）関西方言が多く、若者の発話や大阪以外の関西圏の方言を含む今回の評価データとは差があることが影響していると考えられる。

5　おわりに

関西方言用の見出し語と学習用コーパスの整備によって、従来を上回る精度で関西方言の形態素解析が可能になった。また、大阪のみならずその周辺の方言についても解析が行えるようになった。

UniDic 短単位での解析結果は、古文用の UniDic とも互換性があるため、各地の方言のテキストを短単位解析することが可能になれば、時間と空間を超えて各種のコーパスを比較することができ、様々な言語現象の研究に活かすことができることが期待できる。今後、各地の方言向けの短単位辞書の整備を進めていきたい。

謝辞

本研究は、国立国語研究所基幹研究プロジェクト「多様な語彙資源を統合した研究活用基盤の共創」による成果の一部であり、JSPS 科研費 23H00007 の助成を受けたものです。

参考文献

[1] 小木曽智信, 小町守, 松本裕治. 歴史的日本語資料を対象とした形態素解析. 自然言語処理, Vol. 20, No. 5, pp. 727–748, 2013.

[2] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源 : 形態素解析用電子化辞書の開発とその応用. 日本語科学, No. 22, pp. 101–123, 2007.

[3] 小木曽智信, 尹煕洙, 王竣磊, 岡田純子. 関西方言を対象とした形態素解析用辞書の開発. 言語処理学会第 30 回年次大会発表論文集, 2024.

[4] 尹煕洙, 王竣磊, 岡田純子, 小木曽智信. 短単位版「関西弁コーパス」の構築と予備的分析. 言語処理学会第 31 回年次大会発表論文集, to appear.

[5] ケビン・ヘファナン. 関西弁コーパスの紹介. 総合政策研究, No. 41, pp. 157–163, 10 2012.

[6] 国立国語研究所（村山実和子ほか）. 『日本語歴史コーパス 江戸時代編 I 洒落本』, 2019.

[7] 国立国語研究所（服部紀子・松崎安子ほか）. 『日本語歴史コーパス　明治・大正編 VI 落語 SP 盤』, 2022.

[8] 小磯花絵, 天谷晴香, 居關友里子, 臼田泰如, 柏野和佳子, 川端良子, 田中弥生, 伝康晴, 西川賢哉, 渡邊友香. 『日本語日常会話コーパス』設計と特徴. 国立国語研究所論集, Vol. 24, pp. 153–168, 1 2023.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).