短単位版「関西弁コーパス」の構築と予備的分析

尹 熙洙 [12]　王 竣磊 [23]　岡田 純子 [2]　小木曽 智信 [21]
[1] 総合研究大学院大学 先端学術院　[2] 人間文化研究機構 国立国語研究所
[3] 東京大学 人文社会系研究科
{gs20233504, wang-junlei, jun-okada, togiso}@ninjal.ac.jp

概要

関西方言を研究するための形態論情報付きコーパスとして，ケビン・ヘファナン教授によって構築された「関西弁コーパス」のテキストの一部を UniDic 短単位に分割し，171 万語に形態論情報を付与した短単位版「関西弁コーパス」を構築した．うち 77.5 万語は人手による検証・修正を行い，残りの 93.5 万語は形態素解析器 MeCab と人手検証済みデータで学習した関西方言用 UniDic を利用して解析した．

本稿では，短単位版「関西弁コーパス」の構築について論じ，その価値を示す活用例として予備的な分析結果を示す．短単位版「関西弁コーパス」は，関西方言の研究に有効な資料として利用されることが期待される．

1　はじめに

証拠に基づく言語現象の研究のために，形態論情報付きコーパスは極めて有用なデータである．しかし，日本語の非標準的変種を対象とした形態論情報付きコーパスは少ない．例えば，「日本語諸方言コーパス」[1] には，方言テキストに加えて方言の音声データと標準語訳テキスト（形態論情報付き）が収録されているが，方言テキストには形態論情報が付与されていない．

「関西弁コーパス」は，形態論情報が付与されている関西方言の公開コーパスであるが，その形態論情報は IPA 辞書（mecab-ipadic）による解析結果をベースとして修正を施したものである．IPA 辞書は，関西方言用に開発された辞書ではないため，関西方言の解析精度は完全ではない．

本研究では，「関西弁コーパス」のテキストを利用し，新たに UniDic 短単位 [1] に分割して形態論情報を付与する．単位の斉一性・見出しの同一性の問題を解決するために提案された UniDic の階層的単位設計を導入することによって，より言語学的な研究に適したコーパスとする効果が期待される．また，同じ UniDic 短単位を採用している他の言語研究用コーパスとの互換性を確保することができる．

1） https://www2.ninjal.ac.jp/cojads/index.html

2　構築

2.1　関西弁コーパス

「関西弁コーパス」[2][2] は，関西学院大学のケビン・ヘファナン教授によって構築され，クリエイティブ・コモンズ・ライセンス 表示-非営利-継承 4.0（CC BY-NC-SA 4.0）で公開されているコーパスで，学生が家族や親密な知り合いと行った 200 件以上の社会言語的なインタビューが収録されている．

サブコーパス　「関西弁コーパス」は，以下のサブコーパスで構成される．

- **KSJ** — 大阪・神戸都市圏
- **KYT** — 京都府（主に京都市出身）
- **TKC** — 兵庫県多可町・西脇市
- **RGS** — 関西在住の留学生

ただ，テキストとともに公開されている話者情報によると，サブコーパス KSJ にも京都府・兵庫県（阪神圏域以外）出身の話者のインタビューが含まれていることに注意が必要である（4 節で後述）．

2.2　収録データの選定

本研究では，本研究が開始した 2021 年 10 月の時点で公開されていた 168 ファイル[3] を，74 ファイルのコアデータと 93 ファイルの非コアデータに分けた（表 1）．サブコーパス KSJ のうち，ファイル KSJ024F6 と KSJ025F6 は同じインタビューを収録しているため，本研究では KSJ025F6 を除外した．

2） https://sites.google.com/view/kvjcorpus
3） ヘファナン教授の「関西弁コーパス」は，テキストファイルで公開されており，1 ファイルには 1 件のインタビューが収録されている．

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

表 1　コアデータと非コアデータ

サブコーパス	コア	非コア	除外
KSJ	38	93	1
KYT	22	0	0
TKC	14	0	0
合計	74	93	1

2.3　形態論情報の付与

UniDic 短単位　UniDic 短単位は，原則として意味を持つ最小の単位（最小単位）2 個の 1 回結合と定義される短い単位である [1]．短単位認定の例を図 1 に示す．

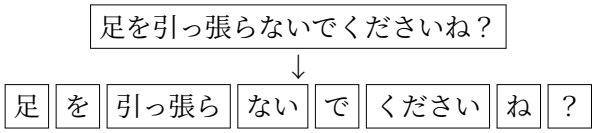

図 1　短単位認定の例

関西方言の特徴をより忠実に反映するために，見出し語の新規登録や活用型・活用形の整備などを行った [3][4]．

コアデータの形態論情報　コアデータは，形態素解析器 MeCab [5][4] と「日本語歴史コーパス」の洒落本（京都） [6]・落語（大阪） [7] 及び「日本語日常会話コーパス」[8][5] を用いて学習した辞書によって解析した後，人手による形態論情報の検証・修正を行った．

形態論情報の修正は，「現代日本語書き言葉均衡コーパス」の形態論情報アノテーション支援システム [9] として開発された「形態論情報データベース」上で行われた．コーパスの修正には「大納言」，見出し語の管理には「UniDic Explorer」を利用した．

また，修正作業の途中に，修正が完了したファイルのデータを用いて辞書の学習を行い，その辞書によって残りのファイルを再解析することにより，人手による修正のコストを削減した．最終的に人手による修正箇所は 40,553 箇所[6] で，全体の 5.23%であるが，最初に修正作業が行われた KSJ に比べて，その次に修正が行われた KYT や，最後に修正が行われた TKC は修正箇所が少なくなっている（表 2）．

非コアデータの形態論情報　非コアデータは，形態素解析器 MeCab とコアデータを用いて学習した辞書 [4] によって解析した結果をそのまま形態論情報とした．

表 2　コアデータの修正箇所

サブコーパス	修正箇所	%
KSJ	22,277	5.93
KYT	10,496	4.79
TKC	7,780	4.31
合計	40,553	5.23

2.4　語数

総語数（句読点などの記号などを含む）を表 3 に示す．

表 3　短単位版「関西弁コーパス」の総語数

サブコーパス	コア	非コア	(コア + 非コア)
KSJ	375,684	935,460	1,311,144
KYT	219,287	0	219,287
TKC	180,412	0	180,412
合計	775,383	935,460	1,710,843

3　形態論情報の精度

2 節で述べた通り，コアデータは人力で検証・修正を行っているが，非コアデータは検証・修正を行っていないため，方言研究の資料として使うためには，精度の評価が必要となる．

本研究では，非コアデータからランダムに抽出した約 3,000 語（短単位）を対象に，[10] で設定された 5 つの評価レベルにおいて人力で精度評価を行った（表 4）．各評価レベルは，Lv.0 を除き UniDic の階層的設計における単位・見出しの階層に対応するものであり，以下のように定義される．

Lv.0 テキスト　Lv.0 は，書き起こしテキストが正しいかを確認するための評価レベルである．正しい形態論情報を付与するためにテキストの修正が必要とされる場合は，Lv.0 の誤りと判定する．例えば，「黄色」が「ヒーロ」になっている場合りである．

Lv.1 境界　Lv.1 は，短単位境界の認定が正しいかを確認するための評価レベルである．例えば，副助詞「なんか」を「なん」（何）と「か」に分けている場合が Lv.1 の誤りとなる．

Lv.2 品詞　Lv.2 は，品詞・活用型・活用形が正しいかを確認するための評価レベルで，UniDic の階層的見出しにおける語形の階層に対応する．品詞の誤りの例は，「うち」（代名詞）を「家」（名詞）としている場合である．活用型の誤りの例は，「繋がらない」という意味の「繋がれへん」の「繋がれ」の部

4）https://taku910.github.io/mecab/
5）https://www2.ninjal.ac.jp/conversation/corpus.html
6）データベース上で修正履歴がある短単位の数．

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

分を「繋がれ」（動詞・五段・未然形）ではなく「繋がれ」（動詞・下一段・未然形）としている場合である．最後に，活用形の誤りは，例えば連体形を終止形としているような場合である．

Lv.3 語彙素　Lv.3 は，語彙素読み・語彙素・語種が正しいかを確認するための評価レベルである．例えば，「教えてもらって」を意味する「教えてもうて」の「もう」を，語彙素「貰う」ではなく「仕舞う」としている場合が Lv.3 の誤りである．

Lv.4 発音形　Lv.4 は，発音形を確認するための評価レベルである．例えば，「どうしてだろう」を意味する「何でやろ」の「何」の発音形を「ナン」ではなく「ナニ」としている場合が Lv.4 の誤りである．

表 4　各評価レベルにおける形態論情報の精度

評価レベル	評価項目	評価値	
Lv.0 (テキスト)	Precision	0.9990	(3093/3096)
	Recall	0.9987	(3093/3097)
	F_1 値	0.9989	
Lv.1 (境界)	Precision	0.9955	(3082/3096)
	Recall	0.9952	(3082/3097)
	F_1 値	0.9953	
Lv.2 (品詞)	Precision	0.9816	(3039/3096)
	Recall	0.9813	(3039/3097)
	F_1 値	0.9814	
Lv.3 (語彙素)	Precision	0.9780	(3028/3096)
	Recall	0.9777	(3028/3097)
	F_1 値	0.9779	
Lv.4 (発音形)	Precision	0.9767	(3024/3096)
	Recall	0.9764	(3024/3097)
	F_1 値	0.9766	

4　予備的分析

本節では，短単位版「関西弁コーパス」の価値を示す活用例として，短単位版「関西弁コーパス」のコアデータを対象とした予備的分析の結果について述べる．

4.1　地域による言語使用の違い

「関西弁コーパス」のサブコーパスは，概ね地域によって分けられているが，サブコーパス KSJ（大阪・神戸都市圏）には，京都府・兵庫県（阪神圏域以外）出身の話者のインタビューも一部含まれているため，注意を要する．

コアデータに属するファイルの中では，以下の 3 ファイルがその例である．

- `KSJ028M9` — 兵庫県加古川市出身・同市在住
- `KSJ029F6` — 兵庫県加古川市出身・同市在住
- `KSJ044F9` — 兵庫県加西市出身・同県多可町在住

これらのファイルは，言語的にはサブコーパス TKC（兵庫県多可町・西脇市）に近い特徴を持つため，本研究の予備的分析では，以上の 3 ファイルと TKC の 14 ファイルを「兵庫」とし，KYT の 22 ファイルを「京都」，KSJ の 38 ファイルから以上の 3 ファイルを除いた 35 ファイルを「大阪・神戸」とした．

4.1.1　助動詞「てる」「とる」「はる」「よる」

表 5 は，助動詞「てる」「とる」「はる」「よる」の各地域における出現頻度を示したものである．これらの助動詞は，必ずしもその意味や機能が一致するわけではないが，各地域の方言の特徴をわかりやすく表す例として挙げる．

表 5　助動詞「てる」「とる」「はる」「よる」

($\times 10^6$)	「てる」	「とる」	「はる」	「よる」
大阪・神戸	13236.2	2166.0	375.9	75.2
京都	15080.7	1199.3	2321.2	9.1
兵庫	5691.8	4545.9	180.7	1321.9

4.1.2　助動詞「へん」に前接する動詞の未然形

表 6 は，否定を表す助動詞「へん」に前接する五段活用動詞の未然形がア段である場合とエ段である場合の頻度を地域別に示したものである．

「へん」に前接する五段活用動詞において，「書けへん」（「書かない」の意）のようなエ段の未然形が「書かへん」のようなア段の未然形より極めて優勢であることは，伝統的な大阪方言の特徴の一つであるとされる [11]．実際，コアデータの「大阪・神戸」地域で唯一 70 代以上であるファイル `KSJ022M9` の話者（大阪府大阪市出身）は，エ段の未然形を 7 回，ア段の未然形を 1 回使用した．しかし，それ以外の話者が話す大阪方言では，伝統的な大阪方言ほどエ段の未然形が優勢ではないことがわかる．

表 6　助動詞「へん」に前接する動詞の未然形

($\times 10^6$)	ア段	%	エ段	%
大阪・神戸	999.3	66.9	494.5	33.1
京都	1955.7	96.5	71.8	3.5
兵庫	1553.0	84.3	288.3	15.7

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

4.2 年齢による言語使用の違い

本研究の予備的分析では，「関西弁コーパス」話者情報の年齢グループを使用する．

4.2.1 動詞「言う」の前の助詞の省略

表 7 は，語彙素「言う」の前の助詞を省略する場合と省略しない場合の頻度を年齢別に示したものである．

「言う」の直前の語の品詞が名詞・代名詞・感動詞・動詞・形容詞・形状詞・助動詞である場合（ただし，活用語は活用形が意志推量形・終止形・命令形である場合のみ），または「助詞-終助詞」「接尾辞-名詞的」である場合は，助詞を省略しているとみなした（副詞可能を除く）．「言う」の直前の語の品詞が連体詞・副詞・接続詞・助詞（終助詞を除く）である場合や，副詞可能である場合，活用語の連用形である場合は，助詞を省略していないとみなした．

年齢が高い話者ほど，助詞を省略する傾向があることがわかる．

表 7 動詞「言う」の前の助詞の省略

($\times 10^6$)	省略する	%	省略しない	%
16〜18 歳	562.4	4.6	11589.5	95.4
19〜23 歳	831.2	6.4	12166.9	93.6
24〜29 歳	620.1	4.9	12072.9	95.1
30〜39 歳	1132.9	9.0	11438.6	91.0
40〜49 歳	1392.5	7.9	16245.7	92.1
50〜59 歳	1430.3	8.7	14997.1	91.3
60〜69 歳	5368.2	31.0	11975.3	69.0
70〜79 歳	6389.4	45.0	7799.4	55.0

4.2.2 ワア行五段活用動詞の促音便とウ音便

表 8 は，ワア行五段活用動詞に助詞「て」や助動詞「た」などがつく場合の連用形が促音便形である場合とウ音便形である場合の頻度を年齢別に示したものである．

促音便形は「言って」のような形，ウ音便形は「言うて」の形である．年齢が高い話者ほど，伝統的な関西方言の形式であるウ音便形を使用していることがわかる．

表 8 ワア行五段活用動詞の促音便とウ音便

($\times 10^6$)	促音便	%	ウ音便	%
16〜18 歳	5061.2	96.3	195.6	3.7
19〜23 歳	5455.3	88.5	707.0	11.5
24〜29 歳	5949.3	95.6	271.3	4.4
30〜39 歳	3712.8	78.2	1037.4	21.8
40〜49 歳	4468.9	75.1	1478.8	24.9
50〜59 歳	5790.5	73.7	2069.0	26.3
60〜69 歳	2498.3	31.1	5533.4	68.9
70〜79 歳	1891.8	20.0	7567.4	80.0

5 おわりに

本研究では，大阪・神戸都市圏，京都府（南部），兵庫県（南部）の方言の特徴を反映する形態論情報付きコーパスとして短単位版「関西弁コーパス」を構築し，精度評価と予備的分析を行った．本研究で構築したコーパスは，関西方言の研究に有効な資料として活用されることが期待できる．

発表者らは今後，日本語の他の非標準的変種の形態論情報付きコーパスも構築していく予定である．

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

謝辞

本研究は，国立国語研究所共同研究プロジェクト「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 JP23H00007 の助成を受けたものです.

「関西弁コーパス」をクリエイティブ・コモンズ・ライセンスで公開してくださったケビン・ヘファナン教授，そして「関西弁コーパス」の構築に参加されたインタビュアー・インタビュイーの方々に深く感謝いたします.

参考文献

[1] 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.

[2] 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, 2012.

[3] 小木曽智信, 尹熙洙, 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] 小木曽智信, 尹熙洙, 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拡張. 言語処理学会第 31 回年次大会発表論文集, to appear.

[5] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[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, 2023.

[9] 小木曽智信, 中村壮範. 『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用. 自然言語処理, Vol. 21, No. 2, pp. 301–332, 2014.

[10] 小木曽智信, 近藤明日子, 雄太, 間淵洋子. 『昭和・平成書き言葉コーパス』の設計・構築・公開. 情報処理学会誌, Vol. 65, No. 2, pp. 278–291, 2024.

[11] 国立国語研究所編. 全国方言談話データベース 目本のふるさとことば集成 第 13 巻 大阪・兵庫. 国書刊行会, 1988.

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