関西方言 UniDic を用いた西日本諸方言書き起こしテキストの形態素解析と精度評価

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

本発表では、「関西方言 UniDic」を用い、西日本各地の諸方言に対する解析精度と頑健性を評価した結果を報告する。「日本語諸方言コーパス」の西日本各地（大阪、三重、滋賀、奈良、福岡）の談話データを用い、精度評価と解析エラーの分析を行う。エラーの原因調査から、機能語の地域差や融合形が精度に与える影響を精査し、特定方言用辞書の周辺地域への適用可能性と歴史と方言を統合する「時空間統合 UniDic」構築への課題を確認する。

1　はじめに

今日、現代標準語の解析精度は飛躍的に向上したが、方言テキストの直接的な解析は依然として困難な課題である。国立国語研究所が公開している「日本語諸方言コーパス」においても、方言書き起こしテキストそのものの形態素解析は実現しておらず、対応する標準語訳を解析するにとどまっている。

発表者らはこれまで、現代語用 UniDic をベースに関西方言の語彙や活用体系を拡充し、短単位で整備し直した学習者のコーパスを用いて「関西方言 UniDic」の開発を進めてきた。本発表では、小木曽ほか（2025）[1] での報告を大幅に拡張し、表記の書き換えと人手による形態論情報の修正を完了した西日本各地のデータを用いて、関西方言 UniDic の他の地域への適用可能性について確認するために、詳細な精度評価とエラー分析を行う。

2　関西方言 UniDic と学習データ

関西方言 UniDic は、現代標準語用 UniDic の言語単位と階層的な見出し語情報 [2] を継承している。歴史資料を対象とした各種の UniDic[3] と並んで、時代や方言を超えて解析結果の比較が可能である。このような日本語の形態素解析用の辞書は他になく、さまざまな日本語の変異を扱う言語研究にとって重要な言語資源となっている。関西方言 UniDic のコスト学習には、短単位版「関西弁コーパス」(KVJ) 77.6 万語 [4]、「日本語歴史コーパス」(CHJ) 上方語 8.8 万語、「日本語日常会話コーパス」 (CEJC) 25.7 万語を組み合わせて使用している [5]。

3　西日本諸方言のコーパス

3.1　COJADS 漢字仮名交じりデータ

「日本語諸方言コーパス」（COJADS）は、1970-80 年代に文化庁によって実施された方言調査を基盤としている [6]。オリジナルのデータ構成は、談話の録音音声に対応する方言の書き起こしテキストが文節単位の「分かち書きカタカナ表記」で付与され、これに対する漢字仮名交じり表記の「標準語訳」が併記されている。しかし、一般的な形態素解析は漢字仮名交じり表記を前提として設計されているため、カタカナ表記のみの方言テキストを直接解析することは困難であった。

そこで本研究では、方言テキストを UniDic 短単位で直接解析可能にするために、カタカナ書き起こしデータをベースに漢字仮名混じり化した、漢字仮名交じり書き換えテキスト [1] を新たに整備した。以下はこのデータ（サンプル ID : 27_b_004）の一部であり、最下段のデータが新たに整備した漢字仮名交じりテキストである[1)]。

- **カタカナ書き起こしテキスト**：ホッタ　キョーガク　ナッテカラノ　コーカワ　オモシロー　ナイワ。　ムツカシーバッカリデ。〔フン。〕ヤッパリ　ムカシノ〔ウン。〕アノ　コーカノ　ホーガ　ウタイヨーテ　ナツカシーテ　ユーテ

1) 実際のデータは発話（話者）ごとに構造化されているが、ここでは会話相手の発話（合いの手）を〔〕内に示した。

- **標準語訳**：そうしたら　共学［に］　なってからの　校歌は　おもしろく　ないよ。　難しいばかりで。　〔ふん。〕　やはり　昔の　〔うん。〕　あの　校歌の　方が　歌いやすくて　なつかしいと　言って
- **漢字仮名交じりテキスト**：ほった　共学　なってからの　校歌は　おもしろう　ないわ。　難しいばっかりで。　〔ふん。〕　やっぱり　昔の　〔うん。〕　あの　校歌の　方が　歌いようて　なつかしいて　言うて

もともと COJADS 収録データは、文節単位で音声・カタカナ書き起こしテキスト・標準語訳が文節単位で対応づけられている。今回作った漢字仮名交じりテキストは、これらを対照し、分かち書きを維持したまま、形態素解析に適した漢字仮名交じり表記へと変換した。

このため、漢字仮名交じりテキストもこれら 3 者と対応が取れており、テキストを解析して得られる単語から「カタカナ書き起こしテキスト」「標準語訳テキスト」「音声データ」をたどることが可能となっている。したがって、漢字仮名交じりテキストは形態素解析を容易にし、単語から現情報を得るインデックスとしても位置づけられる。

3.2　対象地域と規模

本研究では、西日本各地の計 5 地域を解析対象とした。各データの規模を表 1 に示す。これらのデータは上述した人手による表記書き換えを行い、関西方言 UniDic で解析した上で、人手による形態論情報の修正が施されており、精度を検証するための正解データとして使用した。

表 1　解析対象方言データの規模（正解数）

地域（当時の市町村名）	サンプル ID	語数
大阪府（大阪市）	27_b_004	5,039
三重県（阿児町）	24_c_099	4,988
滋賀県（甲賀町）	25_e_099	8,202
奈良県（五條市）	29_c_003	3,237
福岡県（北九州市）	40_a_002-1	4,586
合計		26,052

なお、上記地点のデータの整備は完全には終了しておらず、一部に検討中の語や辞書に未登録の語が含まれている。これらの語は、解析精度の評価の対象として含め、正解できなかったものとして扱った。また、言い間違い・言いよどみ・ 解釈不明等の、正解となる形態論情報を付与することができない箇所が含まれる。これらは解析精度の評価の対象外とした。

4　精度評価と結果

4.1　精度評価

精度評価は、小木曽ほか（2013）[3] で設定された 4 つの評価レベルにおいて実施する。各評価レベルは UniDic の階層的設計に対応しており、その定義は以下のとおりである。

- **Lv.1 境界**: 単位境界（分かち書き）の認定
- **Lv.2 品詞**: Lv.1 ＋ 品詞・活用型・活用形の認定
- **Lv.3 語彙素**: Lv.2 ＋ 語彙素（見出し語）の認定
- **Lv.4 発音形**: Lv.3 ＋ 発音形（語形変異）の認定

COJADS 西日本各地点におけるレベル別の精度（F 値）をまとめたグラフを図 1 に示す。

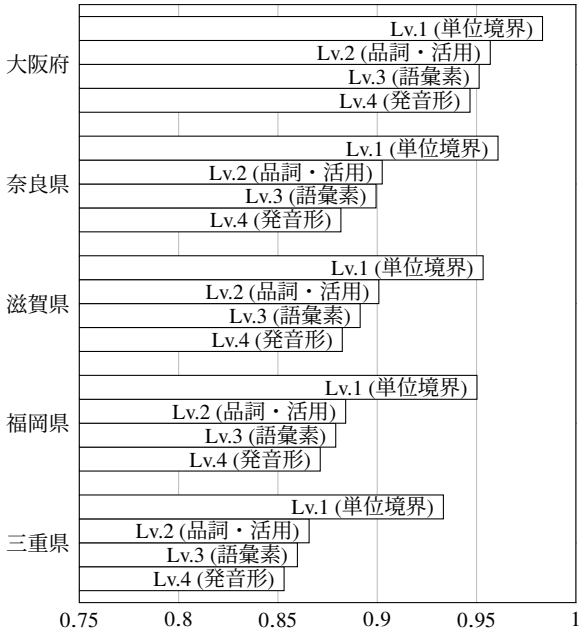

図 1　地域別・レベル別解析精度（F 値）

大阪府では単位境界の精度（Lv.1）が F 値 0.98、もっとも難しい発音形の精度（Lv.4）でも 0.95 と、非常に高い精度を保っている。このことは、関西方言 UniDic の主たる対象が大阪方言であり、学習用コーパスの大半を占める KVJ が当該地域のデータであることを考えれば、自然な結果である。

しかし、KVJ は 2010 年代に収録された幅広い年齢の話者のデータであるのに対し、評価データの COJADS は 1978 年の高齢者の談話である。このように年代差・世代差が大きいことを考えると、この辞書が世代を超えた古い大阪の談話であっても非常に高い精度で解析できる頑健性をもつといえる。

4.2 地域別精度の分析

単位境界（Lv.1）については全地点において高い精度を示しており、最も低い三重でも 0.93 以上と、どの地域でも比較的安定している。一方、品詞認定（Lv.2）の精度は大阪府 0.96 と他地点で大きな差があり、滋賀県・奈良県が 0.90、福岡県 0.88、三重県 0.87 と低下する。語彙素認定（Lv.3）、発音形認定（Lv.4）の精度も低下していくが、品詞認定レベルでの低下幅が最も大きく、広域の方言に対応していくためにはここに課題があることが分かる。

総じて、Lv.4 以外のエラーは方言の機能語に集中していると看取される。助詞・助動詞などの機能語は出現頻度が高いため、誤解析は全体の解析精度を低下させる主な原因となる。

語彙素認定で F 値 0.9 を下回る精度は、そのままで方言の分析に利用できるものとはいいがたい。しかし、「日本語歴史コーパス」の整備ではより低い精度の解析結果をもとに人手修正で精度を引き上げて完成させており [7]、言語研究用コーパスの整備のために、人手による一定の修正を前提とするのであれば、実用的なものであるといえる。

なお、地理的に隔たった福岡県の精度が全レベルで三重県を上回っており、近畿の他県と同等の解析精度を示している。このことは、関西方言 UniDic が関西方言のみならず、より広い西日本広域の方言の解析にも利用できることを示唆している。

5 レベル別のエラーの実態

5.1 Lv.1：単位境界のエラー

単位境界は全体的には解析精度が高く、過分割を抑制する傾向が確認された。しかし、方言に多用される一部の**助詞・助動詞**が過剰に分割され、逆に**融合形・連語**は各要素が分離できていない例がある。

- **助詞・助動詞の過分割**: 近畿地方一般にみられる、原因理由の接続助詞「よって」、逆接の接続助詞・副助詞「かて」「かって」は、接続助詞「て」や副助詞「って」を含んだ形式と誤判定され、滋賀にみられる完了存続の助動詞「たある」は、動詞「有る」を含む形式と誤判定されている。
- **融合形・連語の分割不足**: 三重にみられる、準体助詞「に」と助動詞「ゃ」の融合形「にゃ」は、否定の助動詞「ず」の仮定形と誤判定され、福岡にみられる、完了の助動詞「つ」と推量の助動詞「ろう」からなる連語「つろう」は、動詞「吊る」や形容詞「辛い」と誤判定されている。

5.2 Lv.2：品詞・活用のエラー

同じ出現形で複数の意味・用法を有する場合、品詞および活用形の**取り違え**が起こりやすい。

- **助詞・助動詞の取り違え**: 助詞の出現形「け」は、原因理由の接続助詞「けん」の語形か疑問の終助詞「け」かで誤判定が多くみられる。三重・滋賀・奈良において、終止形に接続する出現形「で」は、原因理由の助動詞「だ」の連用形か単なる終助詞かで誤判定が多い。滋賀のアスペクト形式「たった」に含まれる助動詞「たる」の連用形「たっ」は、敬語形式「たった」にみられる助動詞「てや」の連用形「たっ」と誤判定されている。福岡にみられる比況の助動詞「ごと」は名詞「事」の語形と誤判定されている。
- **活用形の取り違え**: 近畿地方一般にみられる当為表現「〜んならん」「〜んなん」に含まれる最初の「ん」は、「ねば」に相当し、否定の助動詞「ず」の仮定形に認定すべきところ、終止形か連体形と誤判定されている。三重にみられる、「行た」「入っとた」などに含まれる動詞・助動詞の連用形の省略形は、連用形と認識されずに、終止形や未然形と誤判定されている。

5.3 Lv.3：語彙素のエラー

品詞・活用が同じでも、引き続き語の**取り違え**が確認される。それに加え、内容語の面で、現在は**一般的でない語形**の判定に誤りがみられる。

- **取り違え**: 「町い行く」などに出現する助詞「い」は、格助詞「に」か「へ」かで、全体にわたって誤判定が多くみられる。近畿地方一般にみられる、「〜てもうて」「〜てもうた」に含まれる連用形「もう」は、動詞「貰う」か「仕舞う」かで誤判定が多発している。
- **一般的でない語形**: 滋賀に出現する「女子（オナゴ、女の人）」「苞（ツト、食品を包むもの）」、福岡に出現する「大作（オオザク、大規模な農業）」などは正しく認識されていない。

5.4 Lv.4：発音形のエラー

Lv.4 では主に内容語の判定に誤解析が生じており、地域にかかわらず、**予測しにくい発音の転訛**が

主な問題となる。例えば「難しい」に対する「むつかしい」、「石臼」に対する「いひうす」、「汁」に対する「し」などの発音形は正しく認識されていない。

このレベルでは、「先生（センセ）」と「学校（ガッコ）」のように、漢語の発音形が標準的な発音形「センセー」「ガッコー」とずれるものが正しく認識されていない。書き起こしテキストでこれらは「先生」「学校」のように漢字で表記されているため、語彙素レベルでは容易に解析できる一方、非標準的な発音形であるため正しく解析されていない。

6 考えられるエラーの原因

6.1 機能語の地域差

学習に用いた「関西弁コーパス」（KVJ）では近畿地方中心部である京阪神都市圏のデータが主であり、同地方周辺部や北九州に現れる機能語は当然カバーしきれていない。例えば、福岡には「言わるりゃ（言わるれば）」といった下二段活用の融合形が出現し、上に挙げた連語「つろう」に含まれる「つ」や比況の「ごと」は評価用データで文語助動詞とされている。こういった近畿地方のデータからほぼ予測できない機能語は、エラーになる可能性が非常に高く、解析精度を大きく左右する。

また、学習用コーパスに既に含まれる出現形であっても、地域の拡張につれ、意味・用法に差が観察される。例えば、出現形「け」に対し、近畿地方では主に疑問の終助詞として使われており、原因理由の接続助詞としては、KVJ に収録している播磨方言のデータに少数みられる程度であるが、北九州では接続助詞が基本となっている。また、出現形「たった」は学習用コーパスにおいて、ほぼ上述の播磨方言のデータに敬語形式としてしか出現しないが、滋賀ではアスペクト形式として使われている。今後は地域ごとに独立に辞書を整備して、地域差による誤解析を改善する必要がある。

6.2 対象資料との年代差

同じ近畿地方でも、評価用データに散見されるものの、学習用コーパスにほとんど出現しない形式が存在する。例えば接続助詞「よって」が挙げられる。興味深いことに、当該形式は学習コーパスのうち「日本語歴史コーパス」（CHJ）上方語資料に数例確認できるが、KVJ にはほぼみられない。これは、評価用データである COJADS の話者がほとんど明治・大正生まれであるのに対し、KVJ の話者は 10 代から 70 代までバリエーションに富んでおり、20 世紀 30 年代後半を生年の上限としていることに起因するであろう。同じ現代方言の談話資料とはいえ、世代差にさらに目を配る必要があると考えられる。

6.3 歴史コーパスとの不整合

学習用コーパス内部にみられる不均一性も解析精度に影響を及ぼす重要なポイントである。例えば CHJ の上方語にも「たある」形式が確認されるが、CHJ の規程によりこの形式は、接続助詞「た」と動詞「有る」に分割されており、滋賀における短単位の認定と齟齬している。今後は、歴史文献資料を扱う見地から整備された辞書と現代談話資料を扱う立場から開発した辞書との統合を図ることで、解析精度を高めるとともに、より高度な利用が可能になると期待される。

6.4 漢語の問題

5.4 節で述べた漢語の発音形の解析エラーは、標準的な漢字表記に非標準的な発音を当てていることが原因である。このエラーを低減するためには、漢字仮名交じりテキストの表記を「せんせ」のように発音形に寄せておくことが考えられる。

しかし、3.1 節で述べたとおり、形態論情報から COJADS の表音的表記や音声を参照できることを考慮すると、過度に発音に寄せた本文にするより、発音形は原データに任せ、形態論情報はそのインデックスと考えることが適切であるかもしれない。

7 おわりに

本研究により、関西方言 UniDic が、大阪方言において世代を超えて高い精度での解析が可能であること、西日本諸方言に対しても一定の有効性を有することが明らかになった。

今後は、エラー分析で明らかになった点を踏まえて全国の COJADS データの整備を進め、各方言圏専用の UniDic を整備することで、方言コーパスと各地の方言を統合できる辞書の整備を進めていく。また、「日本語歴史コーパス」や各時代別の UniDic の整備と連携し、方言と歴史とを統合して扱うことのできる「時空間統合 UniDic」の構築を目指す。

謝辞

本研究は JSPS 科研費 23H00007 の助成を受けたものである。

参考文献

[1] 小木曽智信, 王竣磊, 尹熙洙, 岡田純子. 短単位版関西弁コーパスと関西方言 UniDic の構築—関西方言書き起こしテキストの形態素解析—. 日本言語学会第 171 回大会 予稿集, pp. 51–52, 2025.

[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, 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.

[4] 尹熙洙, 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, 2025.

[5] 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, 2025.

[6] 国立国語研究所. 日本語諸方言コーパス (COJADS). https://cojads.ninjal.ac.jp/, 2023.

[7] 小木曽智信. 『日本語歴史コーパス 江戸時代編』の設計と構築, コーパスによる日本語史研究　近世編. ひつじ書房, 2023.

A　COJADS 西日本各地点における形態素解析精度の詳細評価結果

表 2　COJADS 西日本各地点における形態素解析精度の詳細評価結果（F 値・適合率・再現率）*

地域	Lv.1（単位境界）			Lv.2（品詞・活用）			Lv.3（語彙素）			Lv.4（発音形）		
	P	R	F	P	R	F	P	R	F	P	R	F
大阪府	**0.9800**	**0.9865**	**0.9832**	**0.9538**	**0.9601**	**0.9569**	**0.9482**	**0.9545**	**0.9513**	**0.9436**	**0.9499**	**0.9467**
直前のレベルとの差：				0.0262	0.0264	0.0263	0.0056	0.0056	0.0056	0.0046	0.0046	0.0046
奈良県	0.9584	0.9632	0.9608	0.9003	0.9048	0.9025	0.8972	0.9017	0.8994	0.8795	0.8839	0.8817
直前のレベルとの差：				0.0581	0.0584	0.0582	0.0031	0.0031	0.0031	**0.0177**	**0.0178**	**0.0178**
滋賀県	0.9435	0.9633	0.9533	0.8916	0.9103	0.9008	0.8822	0.9007	0.8914	0.8733	0.8917	0.8824
直前のレベルとの差：				0.0519	0.0530	0.0524	**0.0094**	**0.0096**	**0.0095**	0.0089	0.0091	0.0090
福岡県	0.9350	0.9659	0.9502	0.8700	0.8987	0.8841	0.8651	0.8936	0.8791	0.8574	0.8857	0.8713
直前のレベルとの差：				0.0650	0.0671	0.0660	0.0049	0.0051	0.0050	0.0077	0.0080	0.0078
三重県	0.9174	0.9497	0.9333	0.8510	0.8809	0.8657	0.8452	0.8749	0.8598	0.8387	0.8681	0.8531
直前のレベルとの差：				**0.0665**	**0.0688**	**0.0676**	0.0058	0.0060	0.0059	0.0065	0.0068	0.0067

* P: Precision (適合率), R: Recall (再現率), F: F-score (F 値)。数値は評価用データに基づき算出。