関西方言を対象とした形態素解析用辞書の開発

小木曽 智信 [12]　尹 熙洙 [21]　王 竣磊 [1]　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
{togiso, gs20233504, wang-junlei, jun-okada}@ninjal.ac.jp

概要

現代語用の UniDic をベースに関西方言の会話書き起こしテキストを対象とした形態素解析用辞書を開発・公開した。学習用のコーパスとして、新たに短単位データとして整備した「関西弁コーパス」の一部と「日本語歴史コーパス」に含まれる近世・近代の関西の資料等を用いた。辞書にはコーパスに出現した未登録語等から方言の語彙を追加したほか、方言形の助動詞類を活用表を含めて整備・登録した。これによって現代の関西方言の書き起こしテキストを、「現代話し言葉 UniDic」を上回る精度で解析することが可能となった。

1　はじめに

今日、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り、従来に比べて形態素解析の重要性は低下してきた。しかし、言語研究にとって個々の語に付与された単語情報は極めて重要な手がかりになるものであり、形態素解析済みのコーパスは欠かせないものである。近年では、日本語の歴史研究や方言研究など、多様な日本語のバリエーションを扱うコーパスの整備が課題となっている。

歴史的な資料の形態素解析については、『日本語歴史コーパス』（CHJ）[1)] 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] [2)]。一方、方言資料の形態素解析はほとんど手つかずであり、後述する「日本語諸方言コーパス」では方言の書き起こしテキストについては解析を行わず、それに対応する標準語訳テキストに形態素解析を施すにとどまっている。そのため、日本各地の方言の書き起こしテキストを直接解析することのできる辞書の整備が求められている。

そこで、発表者らは UniDic をベースとして日本語諸方言の形態素解析を行うための辞書の開発を開始した。本発表は、日本語諸方言の形態素解析の最初の試みとして、比較的テキストの入手がしやすい関西方言について、UniDic 短単位での形態素解析を可能にする辞書の構築を行ったものである。

2　関連研究

2.1　関西方言のコーパス

関西方言を収録したコーパスとして「関西弁コーパス」[2] [3)] がある。このコーパス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、CC ライセンスで公開されているもので、現在 200 以上のファイル、約 121 万語のデータが公開されている。内容は大学生が家族や親しい知り合いと行った社会言語的なインタビューを収集し、書き起こしたテキストである。Mecab による形態素解析が行われているが、標準の 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。

また、関西方言を含む日本語の諸方言を収録したコーパスとして、国立国語研究所の「日本語諸方言コーパス」（COJADS）がある。これは、文化庁が 1977〜1985 年に行なった「各地方言収集緊急調査」の方言談話の収録データの一部を再整備して収録したものである。方言談話の書き起こしは表音的なカタカナ表記によって行われており、これに対応する標準語のテキストが付与されている。コーパスは標準語テキストにのみ UniDic による形態素解析を施してオンラインで公開されている[4)]。

このほかに利用可能なコーパスとして、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」[3] [5)] と、同じく「明治・大正編 VI 落語 SP

1）https://clrd.ninjal.ac.jp/chj/
2）https://clrd.ninjal.ac.jp/unidic/download_all.html#unidic_chj
3）https://sites.google.com/view/kvjcorpus/
4）https://www2.ninjal.ac.jp/cojads/
5）https://clrd.ninjal.ac.jp/chj/edo.html#share

盤」[4] [6] 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 [5]。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、東京のものを除くと、大阪の 51 作品（落語家 10 人）の約 8.5 万語を収録している。

2.2 関西方言の形態素解析

関西方言の形態素解析を扱った研究として廣川・深澤・松村・原田 (2016)[6] がある。廣川らは形態素解析器として JUMAN[7] を利用し、語彙を拡充、関西方言用の活用表を整備した JUMAN 辞書を構築した。これを用いて「関西弁コーパス」を解析して評価を行い、標準の JUMAN 辞書を上回る精度で解析が可能になったことを報告している。

3 関西方言用の UniDic

言語研究を目的とした関西方言の形態素解析用の辞書としては、UniDic 短単位をベースとすることが望ましい。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長を持っている [7]。また、歴史的資料を対象とした UniDic シリーズが整備されており、見出し語に互換性があるため、ジャンルや時代を超えた解析結果の比較が可能となる。ここに日本語諸方言用の UniDic が加われば、時間と空間を超えた日本語のバリエーション比較が可能となり、言語研究の上で大きなメリットがある。

歴史的資料を対象とした UniDic の開発時の調査から、日本語のバリエーションに対応した辞書を構築するためには、単に語彙を追加するだけではなく、教師データを整備して再学習を行うことが有効であることが判明している [1]。そこで、関西方言用の UniDic では、主たるターゲットである現代の関西方言を収録する「関西弁コーパス」の一部を短単位で整備し直して学習用コーパスとし、同時に UniDic 辞書に語彙を追加し、活用表を整備して辞書の再学習を行うこととした。形態素解析器には従来の UniDic と同様、MeCab[8] を用いた[8]。

4 学習用コーパス

学習用のコーパスとして、「関西弁コーパス」を「現代話し言葉 UniDic」[9] で解析したテキストを国立国語研究所の形態論情報データベース [9] 上で人で修正を施し、正解データとして 20 ファイルを整備した。下記の通り、このうち 18 ファイルを学習用とし、2 ファイルを評価用とした。

- 「関西弁コーパス」学習用（約 16.8 万語）：
  - KSJ005F3, KSJ006F7, KSJ010F3, KSJ011F3, KSJ012M4, KSJ014F6, KSJ016M6, KSJ017F8, KSJ018M7, KSJ019F6, KSJ020F5, KSJ021M6, KSJ022M9, KSJ027F5, KSJ029F6, KSJ031F4, KSJ033F5, KSJ037F7
- 「関西弁コーパス」評価用（約 1.7 万語）：
  - KSJ028M9, KSJ030F7

なお「関西弁コーパス」にはアラビア数字で書き起こされた語が含まれるが、UniDic による解析では数字変換処理が前提となっていること、もとの正確な語形が不明であることと等から今回は評価対象外として、アラビア数字のみからなる語はコーパスの整備も行わなかった。

学習用のコーパスとしてこれだけでは十分でないと考えられるため、ターゲットに比較的近いドメインのコーパスである上述した CHJ 洒落本の上方語データ、落語 SP 盤の大坂のデータ、さらに「日本語日常会話コーパス」（CEJC）[10][10] のコアデータを追加用の学習用コーパスとして用意した（表 1）。

コーパス名	略称	語数
「関西弁コーパス」 18 ファイル	KVJ	16.8 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	6.3 万語

表 1 学習用のコーパス

5 UniDic の整備

関西方言の解析のためには辞書の側でも語彙の追加が必要になるが、現代語用の UniDic は、もともと「現代日本語書き言葉コーパス」[11] 構築のために

6） https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#rakugo
7） https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN
8） https://taku910.github.io/mecab/
9） https://clrd.ninjal.ac.jp/unidic/download.html#unidic_csj
10） https://www2.ninjal.ac.jp/conversation/corpus.html

開発されており、この中に含まれる関西方言の語彙がある程度登録されている。そのため、これらを既存の登録語を活用しつつ不足する語を追加する形で行った。

5.1 語彙の追加

「関西弁コーパス」の学習用コーパス整備に当たって追加した語としては、現段階では関西の地名など、固有名詞が中心である。方言の特徴を反映したものとしては、例えば「うせやん」として出現する「嘘（ウソ）」の異語形「うせ」などがある。この語形は UniDic の階層構造を活かして、語彙素「嘘」の語形として新たに「ウセ」を追加している。

活用語では、助動詞とした「よる」の可能動詞形「よれる」（言いよれへん）、「たる」（「てやる」の意）の可能動詞形「たれる」（行かしたれへん）、動詞「しまう」の語形「まう」の可能動詞形「まえる」（消化してまえる）等があった。

また、現時点では未処理となっている語として、「好っきゃねん」のような融合形があり、こうしたものは辞書未登録のままとなっている。

5.2 活用の整備

単に語彙を追加するだけでなく、新たに助詞・助動詞の語形と接続について議論を行い、活用形を整備し直した語も多い。下記はその例である。

言わへん 未然形接続として上接の一段活用動詞の活用表を再整備

言っててん 助動詞「てる」の連用形とする

言わなあかん 助動詞「ず」の仮定形とする

言ってん 連用形に接続する終助詞とする

言いもって 連用形に接続する接続助詞とする（「言いながら」の意。）

言わんと 「ん」は助動詞「ず」の終止形、「と」は接続助詞とする

否定の助動詞「へん」「ず」については、「へんなんだ」「んなんだ」「へなんだ」のように多様な形式が現れ、今後さらに活用表を整備する必要がある。

今回整備しきれず、今後の対応が必要な課題として、「何しょん」「やってもたな」、「食べたあかん」「言っとたわ」のような融合形や省略形の問題がある。また、「入れやへん」「入れやなあかん」「見やん」「来(こ)やん」「居(い)やはる」などの「や」挿入形の扱いも課題である。

6 精度評価

上述の「関西弁コーパス」評価用データ約 1.7 万語に対して、既存の形態素解析用辞書である「現代話し言葉 UniDic」と、新たに整備したコーパスを用いて学習した辞書を用いて形態素解析を行い、その解析精度の評価を行った。解析は mecab-0.996 で行った。

6.1 評価方法

精度評価は小木曽・小町・松本（2013）[1] と同様に、UniDic の階層構造に対応した 4 つのレベルに分けて行った。Lv.1 は語の単位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 の語彙素の階層の語彙素読み・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 の発音形の階層に相当する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」という形に正しく認定できているかまでを評価する。

6.2 解析精度

新たに関西方言用に構築した UniDic 辞書は、適した学習用のデータの組み合わせを探るため、表 1 のコーパスを組み合わせて学習を行って 4 種類用意した。したがって、既存の「現代話し言葉 UniDic」と合わせた下記の 5 つの辞書による解析精度を比較することになる。

- UniDic-CSJ : 現代話し言葉 UniDic
- KVJ : KVJ（18 ファイル）のみで学習した関西方言用 UniDic
- KVJ+CEJC : KVJ+CEJC コアデータで学習した関西方言用 UniDic
- KVJ+CHJ : KVJ+CHJ 関西データで学習した関西方言用 UniDic
- All (KVJ+CEJC+CHJ) : KVJ+CHJ 関西データ+CEJC コアデータで学習した関西方言用 UniDic

その結果を 表 2 に示す。評価値として、適合率

表 2 解析精度

評価レベル	評価項目	UniDic-CSJ	KVJ	KVJ + CEJC	KVJ + CHJ	All
Lv.1 (境界)	Precision	0.9859	0.9889	0.9884	0.9895	**0.9896**
	Recall	0.9866	0.9902	0.9905	0.9900	**0.9907**
	F_1 値	0.9862	0.9896	0.9895	0.9897	**0.9902**
Lv.2 (品詞)	Precision	0.9565	0.9642	0.9666	0.9658	**0.9689**
	Recall	0.9571	0.9655	0.9687	0.9663	**0.9700**
	F_1 値	0.9568	0.9648	0.9676	0.9660	**0.9694**
Lv.3 (語彙素)	Precision	0.9528	0.9617	0.9645	0.9633	**0.9671**
	Recall	0.9534	0.9630	0.9666	0.9638	**0.9682**
	F_1 値	0.9531	0.9624	0.9656	0.9636	**0.9677**
Lv.4 (発音形)	Precision	0.9433	0.9589	0.9616	0.9607	**0.9647**
	Recall	0.9439	0.9602	0.9637	0.9612	**0.9658**
	F_1 値	0.9436	0.9596	0.9627	0.9610	**0.9652**

(Precision)、再現率 (Recall)、F_1 値を示した。なお、上述したとおりアラビア数字のみの語は正確な評価ができないので全体から外してある。

表 2 から分かるとおり、すべてのレベル・すべての項目で、All (KVJ+CEJC+CHJ) が最高の数値となった。Lv.3 語彙素認定の F_1 値で 0.9677 に達しており、ベースラインとなる「現代話し言葉 UniDic」の 0.9531 と比較しても十分に高い精度である。現代標準語テキストの形態素解析精度にはやや劣るものの、歴史的資料を対象とした各種 UniDic の解析精度に匹敵するものであり、人手による一定の修正を前提とした言語研究用のコーパスの整備を目的とした場合には十分な精度となっていると言える。

CEJC は同じ現代語の会話を書き起こしたデータではあるものの、主に東京の話者によるものであって、関西方言の会話には必ずしも適していないと思われたが、これを学習用に追加することで、ベースラインのみならず Lv.1 以外では KVJ を上回った。また、CHJ は関西の話し言葉資料のデータではあるものの、明治期以前の資料であって、現代関西方言の会話には必ずしも適していないと思われたが、これを学習用に追加することですべてのレベルで KVJ を上回った。

KVJ が 16.8 万語に対して、CEJC は 6.3 万語、CHJ は 11.7 万語であって、KVJ 以外のものを足した場合には異質なコーパスがかなりの割合を占め、All では半数を超える。それでも精度向上につながっているのは、これらのコーパスに一定の共通性があることと、学習用のコーパスサイズが未だ小さいためサイズを大きくすること自体が効果的であることによるものと思われる。

今後、KVJ の学習用コーパスを増やすことによってさらに高い精度での解析を実現することが期待できる。

7 おわりに

新たに UniDic 短単位版の「関西弁コーパス」を整備し、既存の国語研のコーパスとともに再学習することで、関西方言の書き起こしテキストを高い精度で解析することが可能になった。この辞書は、他の UniDic と同様に公開するとともにオンラインの解析ツール「Web 茶まめ」[12][11)]でも利用可能にする予定である。

しかし、学習用のコーパスのサイズは未だ十分とは言えない。また、現時点では融合形・省略形などの見出し語の整備が不十分であり、それに伴って学習・評価用のコーパスも完全な修正を行うことができていない。言語研究に活用できるようにするために、今後さまざまな語形バリエーションへの対応を進めていきたい。また、関西方言用の UniDic は日本語諸方言の形態素解析実現のための第一歩である。今後、各地の方言用の UniDic を整備する必要があるが、その際には相互の比較が可能になるように配慮した見出し語・活用表の設計を進めていきたい。

最後に、本研究の多くは「関西弁コーパス」を用いて行われたが、このコーパスは CC BY-NC-SA 4.0 ライセンスで公開されている。今後、さらに整備を進めた後に、短単位版の「関西弁コーパス」を同ライセンスの下で公開する予定である。

11) https://chamame.ninjal.ac.jp/

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 23H00007 の助成を受けたものです。

「関西弁コーパス」を CC ライセンスで公開してくださったケビン・ヘファナン氏に感謝します。

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] 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.

[3] 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.

[4] 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.

[5] 岡部嘉幸, 橋本行洋, 小木曽智信. コーパスによる日本語史研究　近世編. ひつじ書房, 2023.

[6] 廣川純也, 深澤拓海, 松村冬子, 原田実. 形態素解析における関西弁の自動認識. Vol. 2016-NL-225, No. 3, pp. 1 – 7, jan 2016.

[7] 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.

[8] 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.

[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.

[11] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. **Language Resources and Evaluation**, Vol. 48, No. 2, pp. 345–371, June 2014.

[12] 堤智昭, 小木曽智信. 複数の unidic 辞書による形態素解析支援ツール『web 茶まめ』の実装と運用. 情報処理学会論文誌, Vol. 64, No. 3, pp. 749–757, 2023.