論　文　ニダバ (*Nidaba*) 43(2014), 31-39

方言に対応した形態素解析辞書の拡張

―『広島大学電話会話コーパス』構築に際して―

廣川純子

1．はじめに

広島大学東広島キャンパス（以下「広大キャンパス」）は、広島市に隣接した東広島市に位置していながら、広島県外各地から流入してきた学生数が非常に多い。広島県出身の学生が全体に占める割合は約3割に過ぎない[1]。平成25年度の入学者数が特に多い都道府県を順に挙げると、広島776名、兵庫141名、福岡140名、岡山・愛媛124名（同数）である[2]。広大キャンパス内では、様々な出身地方言の接触が生じている。学生は入学当初、出身地方言の特徴を比較的よく保った話し方をしているようであるが、一定の期間を経た後には、彼らの話し言葉は、典型的な広島方言とも、各自の出身地方言や全国共通語とも異なる、独自のキャンパス内共通語のようなものに収束していることがうかがえる。

筆者は、広大キャンパス内における言語接触の結果生じる言語収束にはどのような傾向が見られるのか[3]、および、その要因はどのようなものかを明らかにすることを目的として、以下に述べる研究方法を用いて『広島大学電話会話コーパス』を構築することとした。

まず、言語収束の最初期にある広島大学の学部1年生から、概ね言語収束が完了した状態にあると推定できる学部4年生まで、数十名を対象として、電話会話を録音する。話し相手は「出身地の（大学入学以前からの）親しい人」および「広大キャンパスで知り合った（同郷人以外の）友人」とする。録音された音声信号に種々の研究用情報を付すことによって『広島大学電話会話コーパス』を構築する。同時に、被験者に対して紙面によるアンケートも実施し、広島方言および出身地方言に対する意識なども併せて記録する。そして、このコーパスの分析に基づいて、各学生の会話に見られる、接続助詞、終助詞、アスペクト辞、動詞の否定形といった語彙および文法項目における①広島方言、②それぞれの出身地の方言、③全国共通語（いわゆる標準語）の特徴が、経年的にどのように変化してゆくのかを明らかにする。

本稿では、この作業の過程において必要不可欠であり、また、各地域の方言を含む話し言葉の分析においても有効活用が見込まれる、形態素解析辞書の拡張について検討中の内容について報告する。

2．方言に対応した形態素解析辞書の拡張

2．1．概要

発話中の語彙および文法項目の分析を行うための準備作業として、それぞれの会話音声の転記テキストを作成し、その転記テキストに形態素解析を施す。

転記テキストとは、会話全体を文字で書き起こしたものである。まず音声信号を 200ms 以上の物理的なポーズで区切られた区間で分割し、これを「転記基本単位」とする。この単位内の発話を一貫した基準で書き起こす。その際、極力、筆者の耳に聞こえた通りの表記を行う。文字化された情報には、笑い、咳払い、息などの非言語的なイベントなどの情報も付与する。

転記テキストは、音声解析ソフト Praat（Boersma & Weenink 2012）および動画解析ソフト ELAN[4]（Max Planck Institute for Psycholinguistics 2012）を利用して作成する。作成された転記テキストを、形態素解析エンジン MeCab（Kudo & Nippon Telegraph and Telephone Corporation 2008）および形態素解析辞書 UniDic（Den, Yamada, Ogura, Koiso and Ogiso 2010～）を用いて解析する。これにより、形態素単位での品詞、活用形、語彙素表記などのさまざまな情報が出力される。

これらのソフトウェアは一括解析処理により、形態素単位での品詞、活用形、語彙素表記などのさまざまな情報を出力する。しかしながら、現在のUniDicは共通語を対象としており、共通語に存在しない各方言特有の表現等については正確な形態素解析を行わないことがある。このため、この手順において正しく解析できなかった形態素を手作業で解析し直す必要がある。筆者は、この作業に用いる拡張版辞書を作成するための準備作業として、作業記録をまとめたMS Excelブックを随時更新（追記・修正）している。

現在『広島大学電話会話コーパス』には、54人の話者による約22時間（約11万語）の発話が格納されている。参考までに、現在公開されている主なタグ付日本語音声コーパスの規模を以下の表に示す。形態素情報を付した音声コーパスとしては、現在構築中の『広島大学電話会話コーパス』の規模は決して小さくないということができる。

[表 1] 現在公開されている主なタグ付日本語会話音声コーパスの規模

コーパスの名称	対象話者数（延べ数）	収録時間	語数（短単位）
国立国語研究所 『日本語話し言葉コーパス』[5]	1751 人	661 時間	752 万語
『理研日本語母子会話コーパス』[6]	22 人	14 時間	3 万語
『千葉大学 3 人会話コーパス』[7]	36 人	5 時間 40 分	―― (不明)
『広大電話会話コーパス』	54 人	22 時間	11 万語*

(* ：作業途中のため推計値)

２．２．UniDic のバージョンについて

形態素解析辞書 UniDic は、公開以来、改良（バージョンアップ）が繰り返され、2013 年 11 月 24 日現在の最新版は 2013 年 3 月 14 日リリースの version 2.1.2[8]である。UniDic には、Windows 用パッケージ版と同時に、ソースファイルやバイナリ辞書等も配布されている。この中の語彙定義ファイル lex.csv に、拡張部分のテキストデータを追記することにより、形態素解析辞書の拡張が可能となる。

但し、正式公開準備中の新規ウェブサイト[9]において「今後このサイトでは、XML 版などの新しい形式で UniDic[10]を公開する予定です」との記載が見られる。さらに、現在の UniDic 付属ユーザーズマニュアルの中にも「形態素解析辞書の生成および拡張については UniDic Tools マニュアル[11]を参照してください」との記述がある。したがって、近々、まったく新しい形式（XML 版）の辞書と、その拡張等に利用できるツール（UniDic Tools）のリリースが期待される。この付属ツールにより、それ以降の拡張部分のデータを入力すれば XML 形式に整形して、拡張辞書が生成できるようになるものと推察される。

本稿においては、UniDic 2.1.2 を用いて作業を進めているが、作業課程で作成した拡張部分のテキストデータは、勿論そのまま「新しい形式」の UniDic 辞書拡張にも適用可能である。

２．３．修正項目のカテゴリー分け

MeCab と UniDic の組み合わせによる一括解析処理を経て、意図した通りの解析結果が得られなかった場合は、以下のいずれかに該当する[12]。まず、これらのいずれに該当するかを、MS Excel ブックとして保存されている解析結果の各行左端（ファイル名が出力されている）の背景色を塗り分けることで直感的にわかりやすい状態にしておく。

(1) 誤解析のタイプ分け

a. 分節誤り［赤］
b. 品詞誤り［黄緑］
c. 同音異義語あるいは同字異義語［紫］
d. 人名以外の固有名詞［水色］
e. その他独特な語彙［青］
f. 人名・あだ名［橙］

左端のセルがこれらのいずれかの色になっている行の各形態素が、一括処理の後、手作業による修正が必要なものとなる。

２．４．手動置換内容の実例

紙面の都合もあるので、前節で記した (1 abc) の三つのパターンに限定して、一件ずつ例を挙げる。ここでは、出現形、語彙素表記および品詞（必要に応じて「品詞の下位区分」を併記）という最小限の項目のみを抜粋記述する。なお、参考のためにこれらの形態素を含む発話（部分）を、それぞれ最初に付した。下線部分が該当例である。

(2) 分節誤りの例

発話：なるほどな、おまえ、そんなこと言いよったらおまえ（以下略）

修正前の解析結果：

出現形	語彙素表記	品詞
言いよっ	言い寄る	動詞
たら	た	助動詞

修正後の解析結果：

出現形	語彙素表記	品詞
言い	言う	動詞
よっ	よる	助動詞
たら	た	助動詞

ここでは、そんなことを「言っていたら」という意味を持つ方言形である「言いよったら」が正しく解析されていない。その代りに「言い寄る」という動詞に条件をあらわす助動詞「たら」がついた形と誤って解析されている。

(3) 品詞誤りの例

発話：確かにね、いや、じ、実際、百均にあるかなーとか、ちょっと思いよったけんさ

修正前の解析結果：

出現形	語彙素表記	品詞	品詞の下位区分
けん	ケン	名詞	固有名詞（人名）

修正後の解析結果：

出現形	語彙素表記	品詞	品詞の下位区分
けん	けん	助詞	接続助詞

ここでの「けん」は、共通語の「から」に該当する順接の接続助詞であるが、「ケン」という人名として解析されてしまっていたので、これを接続助詞として修正した。

(4) 独特な語彙の例品詞誤りの例

発話：「情活」終わった？　できた？きのう

修正前の解析結果：

出現形	語彙素表記	品詞
情	情	名詞
活	活	接頭辞

修正後の解析結果：

出現形	語彙素表記	品詞
情活	情活	名詞

この場合の「情活」は「情報活用（演習）」という授業科目名すなわち複合名詞の略称である。「情活」という語は、一般に普及しているものと考えられるので、一つの名詞として扱うものとした。類似例として「教社」（教育学部内における社会科教育関連の学科もしくはコース名の略称）や「教採」（教員採用試験の略称）などがある。

以上のような、修正前と修正後の組み合わせを記録してゆく。同じパターンのものに対して、修正がほぼ一対一対応していると判断できれば、拡張項目リスト（次節参照）に追記するものとする。

２．５．拡張項目リストの作成

２．３．で塗り分け処理した項目を抜き出し、修正前後の分析単位それぞれを掲載したリスト（MS Excel ブック）を作成する。修正前後の分析単位の配列（1 つ以上の分析単位の並び）が、ほぼ確定できるものは LIST という名称のシートに追加し、それ以外のものや、配列要素が同じでも何通りかの解釈が可能なものは LIST2 という名称のシートに追加してゆく。また、何通りかの解釈が可能であっても、そのうちの一種類に置換できる可能性が極めて高いと判断した場合には、その一種類の組み合わせを LIST シートに記し、その他の組み合わせ候補を LIST2 シートに記すものとする。また、LIST2 シートには繰り返し用いられる人名・あだ名等も記録してゆく。

なお、手作業で修正する際にも対象箇所が見つけやすいように、当面は五十音順になるように追加挿入してゆく。参考までに、作業中の「修正項目リスト」LIST シートおよび LIST2 シート（MS Excel 画面）を図 1 および図 2 に示す。

ファイル名	解析単位	解析単位(修正後)	品詞1	品詞2	品詞3	品詞4	活用型	活用形	語彙素読み	語彙素表記	書字形出現	発音形出現	カナ形出現
H 20510_004_ch1	くっ	くっそ	感動詞	一般	*	*	*	*	クソ	くそ	くっそ	クッソ	クッソ
H 20510_004_ch2	そ												
H 20510_012_ch1	けー	けー	助詞	接続助詞	*	*	*	*	ケエ	けえ	けー	ケー	ケー
L1 20508_023_ch1	けえ	けえ	助詞	接続助詞	*	*	*	*	ケエ	けえ	けえ	ケー	ケエ
H 20510_004_ch1	月	月末	名詞	普通名詞	副詞可能	*	*	*	ゲンマツ	月末	月末	ゲンマツ	ゲンマツ
H 20510_004_ch2	末												
H 20525_021_ch1	けん	けん	助詞	接続助詞	*	*	*	*	ケン	けん	けん	ケン	ケン
H 20508_024_ch1	こい	こい	助詞	一般	*	*	五段-ガ行	連用形-イ	コグ	漕ぐ	こい	コイ	コイ
H 20508_024_ch1	と	どん	助動詞	*	*	*	五段-ラ行-	連体形-撥	トル	とる	どん	ドン	ドン
H 20508_024_ch2	ん												
H 20726_001_ch1	コウ	コウハン	名詞	普通名詞	一般	*	*	*	コウハン	合板	合板	コーハン	コウハン
H 20726_001_ch2	ハン												
H 20508_023_ch1	来れ	来	動詞	非自立可能	*	*	カ行変格	未然形-一	クル	来る	来	こ	コ
H 20508_023_ch1	ん	れ	助動詞	*	*	*	下一段-ラ行	未然形-一	ラレ	られ	れ	れ	レ
H 20508_023_ch1		ん	助動詞	*	*	*	助動詞-ン	終止形	ン	ん	ん	ン	ン
L1 20508_023_ch1	来	来	動詞	非自立可能	*	*	カ行変格	未然形-一	クル	来る	来	こ	コ
L1 20508_023_ch1	れん	れ	助動詞	*	*	*	下一段-ラ行	未然形-一	ラレ	られ	れ	れ	レ
L1 20508_023_ch1		ん	助動詞	*	*	*	助動詞-ン	終止形	ン	ん	ん	ン	ン
H 20510_004_ch1	今	今セメ	名詞	普通名詞	副詞可能	*	*	*	コンセメ	今セメ	今セメ	コンセメ	コンセメ
H 20510_004_ch1	セ												
H 20510_004_ch2	メ												
H 20510_013_ch1	さし	さ	動詞	非自立可能	*	*	サ行変格	未然形-サ	スル	為る	さ	サ	サ
H 20510_013_ch1		し	助動詞	*	*	*	下一段-サ	連用形-一	セル	せる	し	シ	シ
L1 20725_021_ch1	シケ	シケバラ	名詞	固有名詞	*	*	*	*	シケバラ	シケバラ	シケバラ	シケバラ	シケバラ
L1 20525_021_ch1	バラ												
H 20508_024_ch1	七	七	名詞	数詞	*	*	*	*	シチ	七	七	シチ	シチ
H 20508_024_ch1	時	時	接尾辞	名詞的	助数詞	*	*	*	ジ	時	時	ジ	ジ
H 20508_024_ch1	し	し	動詞	非自立可能	*	*	サ行変格	連用形-一	スル	為る	し	シ	シ
H 20508_024_ch1	とっ	とっ	助動詞	*	*	*	五段-ラ行-	連用形-促	トル	とる	とっ	トッ	トッ
H 20510_012_ch1	じゃ	じゃ	助動詞	*	*	*	助動詞-ジ	連用形-一	ジャ	じゃ	じゃ	ジャ	ジャ

[図 1] 作業中の「修正項目リスト」LIST シート (部分)

ファイル名	解析単位	解析単位(修正後)	品詞1	品詞2	品詞3	品詞4	活用型	活用形	語彙素読み	語彙素表記	書字形出現	発音形出現	カナ形出現
H 20508_024_ch1	なん	なん	動詞	非自立可能	*	*	五段-ラ行-	連用形-撥	ナル	成る	なん	ナン	ナン
H 20508_023_ch1	に	に	名詞	数詞	*	*	*	*	ニ	ニ	ニ	ニ	ニ
H 20508_023_ch1	も	も	副詞	*	*	*	*	*	モウ	もう	もう	モー	モウ
H 20426_005_ch1	もん	もん	助詞	終助詞	*	*	*	*	モノ	もの	もん	モン	モン
H 20426_005_ch1	やー	やー	感動詞	フィラー	*	*	*	*	ヤ	や	やー	ヤー	ヤア
H 20508_023_ch1	やっ	やっ	助動詞	*	*	*	助動詞-ヤ	連用形-一	ヤ	や	やっ	ヤッ	ヤッ
H 20509_023_ch1	て	て	助詞	接続助詞	*	*	*	*	テ	て	て	テ	テ
L1 20525_021_ch1	や	や	助動詞	*	*	*	助動詞-ヤ	終止形-一	ヤ	や	や	ヤ	ヤ
L1 20525_021_ch1	ね	ね	助詞	終助詞	*	*	*	*	ネ	ね	ね	ネ	ネ
H 20508_023_ch1	やん	やん	助詞	終助詞	*	*	*	*	ヤン	やん	やん	ヤン	ヤン
H 20508_023_ch1	やん	やん	助詞	終助詞	*	*	*	*	ヤン	やん	やん	ヤン	ヤン
H 20509_023_ch1	か	か	助詞	終助詞	*	*	*	*	カ	か	か	カ	カ
H 20508_024_ch1	やん	やん	助詞	終助詞	*	*	*	*	ヤン	やん	やん	ヤン	ヤン
H 20508_024_ch1	ねえ	ねえ	助詞	終助詞	*	*	*	*	ネ	ね	ねえ	ネー	ネエ
H 20525_021_ch1	ん	ん	助動詞	*	*	*	助動詞-ダ	連用形-撥	ダ	だ	に	ン	ン
L1 20525_021_ch1	ん	ん	助詞	準体助詞	*	*	*	*	ノ	の	ん	ン	ン
L1 20525_021_ch1	ん	ん	助詞	終助詞	*	*	*	*	ノ	の	ん	ン	ン
H 20508_024_ch1	ん	ん	感動詞	フィラー	*	*	*	*	ン	ん	ん	ン	ン
H 20508_024_ch1	ん	ん	助詞	格助詞	*	*	*	*	ノ	の	ん	ン	ン
H 20525_021_ch1	ん	ん	助詞	格助詞	*	*	*	*	ノ	の	ん	ン	ン
H 20525_021_ch1	ち	ち	名詞	普通名詞	一般	*	*	*	ウチ	内	ち	チ	チ
H 20426_005_ch1	アイ	アイ	名詞	固有名詞	人名	名	*	*	アイ	アイ	アイ	アイ	アイ
H 20426_005_ch1	アキ	アキ	名詞	固有名詞	人名	名	*	*	アキ	アキ	アキ	アキ	アキ
H 20426_005_ch1	ウラ	ウララ	名詞	固有名詞	*	*	*	*	ウララ	ウララ	ウララ	ウララ	ウララ
H 20426_005_ch1	ラ												

[図 2] 作業中の「修正項目リスト」LIST2 シート (部分)

2．6．Excelマクロによる半自動置換処理

前述の修正項目リストに基づき、以下の手順で半自動置換処理を実行する Excel マクロ・プログラムを作成した。

(5) 処理手順

a. 修正項目リストの LIST シートに掲載されている、上下の空行で区切られた解析単位（図の B 列）配列をすべて読み込む。
b. 新たな解析対象ファイルの解析単位の中から、一連の配列要素が一致するものを見つけ出す。
c. 一致した場合、これを（LIST シートに記載されている通りの解析結果に）置き換えてもよいかどうかを問う Yes/No ダイアログを表示。
d. Y ボタン押下（もしくは Yes ボタン上でのクリック）により、置き換えが実行され、N ボタン押下（もしくは No ボタン上でのクリック）により、置き換えはキャンセルされる。

この際、修正候補として検出された行および実際に置き換えが実行された行の左端の列（A 列）の背景色を、それぞれ青緑色・黄色（他で使っていない色）に塗り替えて目立たせるなど、後の確認作業に役立つような処理も行っている[13]。この半自動処理が一通り終了した後、A 列の背景色が塗り替えられた行を重点的に確認していけば、最初から完全に手作業で解析誤りを探してゆくよりも格段に効率的である。しかし、LIST シートに掲載されているもの以外[14]の修正は、やはり手作業で行う必要がある。

2．7．後処理

前節の Excel マクロで置換できなかった部分を確認してゆく。特に、左端列の背景色がデフォルト以外の色になっている行に留意する。2．5．で言及した LIST2 シートに複数の修正候補が記載されているものについては、その中から最も適していると考えられるものを選べばよい。例えば「ん」という形態素に対しては、否定の意味を含む断定の助動詞／準体助詞／終助詞／感動詞（フィラー）／格助詞のうちのいずれかに、ほぼ該当する。

LIST シートおよび LIST2 シートいずれにも存在しない、新たな置換候補が出現した際には、それぞれに応じたシートに追記してゆく。

3．まとめ

以上の一連の作業を繰り返すことにより、解析修正置換候補の配列も増え、精度が向上してゆく。なお、現在はエクセルシート上の一覧となっているが、最終的には「辞書ファイルの拡張」に適したカンマ区切りあるいは TAB 区切りのデータ形式（CSV 形式）に整

える必要がある。これをUniDicの辞書ファイルに統合することにより、形態素解析辞書の拡張が実現される。

なお、2013年12月時点での修正項目リスト記載件数（形態素数）[15]は、LISTシートで修正前280・修正後253、LIST2シートで修正前69・修正後61となっている。

4．今後の課題

現在のExcelマクロでは、リストを一旦全部読み込んで保持した状態で、修正の必要がありうる形態素の組み合わせを検索しているので、作業中に相応のメモリを消費している状態である。また、修正前と修正後とで、対象の形態素数が一致しない組み合わせが多数存在することにより、LISTシートに不備があると、マクロ実行最中に置換する配列のインデックスがずれてしまう可能性もある。より簡潔で扱いやすいリスト管理の方法も検討したい。

また、現在はLISTシートに記載されている形態素の組み合わせに対してのみ、一対一対応の置換候補を示し、置換してよいかどうかという二者択一式Yes/Noダイアログを表示している。複数の置換候補が存在するものに関しては別シート（LIST2シート）に記録し、後から再度、手作業で確認・修正という手順を踏んでいる。今後は複数の修正候補を示して、その場で最適なものを選択できるように改良できればと考えている。

注

[1] 広島大学公式ウェブサイト内
http://www.hiroshima-u.ac.jp/top/houjin/siryo/ukeire/index.html で公開されているPDFファイルの「地域別志願者・合格者割合」によると、平成24年度の合格者のうち、広島県出身者の占める割合は28.1%、すなわち県外出身者の割合は71.9%である。（2013/11/24閲覧）

[2] 広島大学公式ウェブサイト内「入学状況（２０１３年）」
http://www.hiroshima-u.ac.jp/top/intro/gaiyou/nyugakujyokyo/p_nzlbq7.html （2013/12/12閲覧）

[3] たとえば、共通語化（あるいは標準語化）の状況や、自分が本来使用している方言以外の、他の方言からの借用語の使用実態など。

[4] 「ELAN（エラン）は、動画と音声資源に注釈を作成するための専門的ツールである」（ELAN Descriptionより）が、本研究では映像は用いず、音声のみを対象としている。Praatと併用することで、より効率よく転記を行うことが可能となる。

[5] 日本語話し言葉コーパス：収録データ詳細（第３刷）ウェブページ
http://www.ninjal.ac.jp/corpus_center/csj/data/3rd/ （2013/12/29閲覧）

[6] Den and Enomoto (2007) 参照。

[7] Mazuka et al (2006) 参照。

[8] UniDic プロジェクト日本語トップページ http://sourceforge.jp/projects/unidic/ (2013/12/29 閲覧)

[9] UniDic ウェブサイト http://download.unidic.org/ (2013/12/29 閲覧)

[10] 「UniDic2」との表記もあり。(小木曽・伝 2013)

[11] 現在ダウンロードできる UniDic のパッケージ内に当該のマニュアルは見当たらない。

[12] 基準を 2 つ以上満たす場合には、最も重要と思われる 1 つを選択するものとする。

[13] 実際に Yes/No のいずれを選択したかという証拠にもなるので、もし間違えて正しくない選択をした際にも後から検出することが容易である。

[14] 特定の解析単位配列に関して、複数の置き換え候補が存在する場合など。

[15] 形態素単位の数。2．4．で例示した「言いよった」の場合、これを 1 件とするのではなく、修正前は「言いよっ」＋「たら」で、形態素数は 2；修正後は「言い」＋「よっ」＋「たら」で、形態素数は 3 となる。

参考文献等

Boersma, Paul, and Weenink, David (2012) Praat version 5.3.59 (http://www.praat.org)

Den, Yasuharu, and Mika Enomoto (2007) A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. In T. Nishida (ed.) Conversational informatics: An engineering approach, 307-330, John Wiley & Sons: Hoboken, NJ.

Den, Yasuharu, Atsushi Yamada, Hideki Ogura, Hanae Koiso, and Toshinobu Ogiso (2010～) UniDic 1.3.12～2.1.2 (http://sourceforge.jp/projects/unidic/)

Kudo, Taku & Nippon Telegraph and Telephone Corporation (2008) MeCab: Yet Another Part-of-Speech and Morphological Analyzer 0.994 (http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html)

Mazuka, Reiko, Yosuke Igarashi, and Ken'ya Nishikawa (2006) "Input for learning Japanese: RIKEN Japanese Mother-Infant Conversation Corpus" 『電子情報通信学会技術研究報告 106:165』pp.11-15.

小木曽智信・伝康晴(2013)「UniDic2: 拡張性と応用可能性にとんだ電子化辞書」『言語処理学第 19 回年次大会発表論文集』pp.912-915