撥音（の解析）は機械（UniDic）にとっても簡単ではなかったんだ！ : BCCWJを中心に

著者	劉 志偉
雑誌名	言語資源活用ワークショップ発表論文集
巻	3
ページ	368-371
発行年	2018
URL	http://doi.org/10.15084/00001671

撥音（の解析）は機械（UniDic）にとっても簡単ではなかったんだ！ ―BCCWJ を中心に―

劉　志偉（埼玉大学）†

/N/ is not easy for UniDic as well

take BCCWJ as an example

要旨

日本語の撥音は種々雑多であるゆえ、日本語学習者にとっては学習しにくい項目である。本発表では、BCCWJ の非コアデータも視野に入れて、撥音の解析に関しては解析精度が98%に到底及ばないことを提示するとともに、具体的に「一般名詞」「オノマトペ」「漢語副詞」「漢字読み」「慣用句」「近畿方言」「呼称」「古典」「語尾」「固有名詞」「ぞんざい表現」「駄洒落」「同音異語」「動詞連用」「特定」「入力ミス」「話し言葉」「表記仮名」「表記仮名遣い」「表記漢字」「フィラー」「複合語」「（近畿以外）方言」「略語」「若者表記」「若者言葉」等の単純誤解析が多いことを明らかにする。

１．はじめに

劉（2018）では、日本語の特殊拍の一つである撥音が学習者にとって難しいことについて述べられている。現代語に限って考えても、日本語の撥音は実に種々雑多である。例えば、話し言葉には「君んち」「嫌んなる」「そんで」といった、くだけた言い方があるのに対し、書き言葉では「割れんばかりの拍手」「いざ行かん」「触れなば落ちん」等固い表現が挙げられる。また、用言の活用に関しては、いわゆる標準語においてだけでも「わかんない」「謝んなさい」「飛べんの」「かもしんない」のようにラ行音が撥音化する場合がある。さらに近畿方言の「食べんで」「行きまんねん」等も考え合わせると、教科書では「飛ぶ」のテ形「飛んで」またはタ形「飛んだ」しか習わない日本語学習者にとって撥音が極めて難解である。

一方、コーパスを用いてデータを収集する際、解析器による誤解析のうち、撥音に関するものがとりわけ多いことに気づかされる。解析器も言わば日本語を学習する存在と見なすことができる。そこで、本稿では『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）を手がかりに、解析システム（UniDic）にとってどのような撥音の判定が難しいか、また日本語学習者にとっての難点との異同について考察を行う。

２．検索条件

筆者本来の目的は動詞または助動詞に後続する撥音を抽出することにあった。従って、「キー」の箇所に順次に「品詞」→「中分類」→「動詞－一般」（交替で「動詞－非自立可能」）を選択し、「後方共起条件 1」の箇所に順次に「活用形」→「小分類」→「未然形－撥音便」（交替で「連用形－撥音便」「終止形－撥音便」「 終止形－撥音便」）を設定した上で、検索ツール中納言 2.4 を用いてデータバージョン 1.1 のデータを抽出した。「動詞－一般」と「動詞－非自立可能」がそれぞれ後続する 4 種類の撥音便と組み合わせると、計 8 個のファイルのデータを収集した。

† di82zhi@yahoo.co.jp

なお、BCCWJ の解析精度については、山崎（2013）で以下のように述べられている。

> BCCWJ の形態論情報はその大半をプログラムで自動的に付与している。１億語というデータを全部人手でチェックすることは現実的ではないためである。形態論情報の精度は約 98%である（コアデータでは約 99%）。したがって、平均して 100 語に１語の解析エラーがあることになる。エラーの種類は、言語単位の区切りが違っているもの、品詞が違っているもの、読みが違っているもの等である。（115 頁）

３．結果

３.１　各ファイルの誤解析の割合

入手した 8 個のファイルをそれぞれ目視で用例を確認し、「キー」の箇所の情報を「語彙素」及び「語彙素読み」と照らし合わせて、「誤解析」と思われる数及びパーセンテージを表１に示した 。

表 1 BCCWJ における各ファイル誤解析の割合

no.	判別不可用例	近畿方言以外の方言	誤解析（キー）	誤解析（%）	後件誤解析（非撥音）	考察可能な対象例	各ファイル用例数
1	0	45	645	9.02%	4	6458	7152
2	0	0	3	5.45%	0	52	55
3	1	8	207	5.45%	9	3572	3797
4	0	3	81	9.62%	5	753	842
5	2	43	423	9.89%	3	3807	4278
6	0	0	7	36.84%	0	12	19
7	0	5	194	10.91%	22	1558	1779
8	0	2	12	6.73%	1	164	179
合計	3	106	1572	8.64%	44	16376	18101

３.２　誤解析のタイプ

誤解析の内実を明らかにすべく、本稿では BCCWJ で抽出した撥音の誤解析（計 1572 例）に対して下位区分を行った。UniDic を学習者に見立てて、間違った解析をもたらした理由に基づき 、表 2 のようなタグ付けをした。

表 2　誤解析の区分一覧

誤解析区分	用例数	誤解析区分	用例数	誤解析区分	用例数
呼称（人名を含む）	305	若者表記	39	同音異語	14
表記漢字	231	入力ミス	31	動詞連用	13
固有名詞	184	漢語副詞	28	一般名詞	12
表記仮名	138	方言	27	同字異訓	11
近畿方言	125	フィラー	27	語尾	7
漢字読み	106	古典	21	若者言葉	7
オノマトペ	76	複合語	18	表記仮名遣い	7
複合要素	66	特定	16	駄洒落	6
話し言葉	53	同音異語	14	総計	1572

４．考察

４.１　「キー」の誤解析

検索条件については 2 節で述べたように、「キー」の箇所に「動詞」（一般／非自立可能）を置き、「後方共起」（後文脈）に撥音諸形を後続させた。本節では「キー」が誤解析にな

っている場合、「実際の語」 を提示すると同時に、「区分」の箇所に誤解析をもたらした理由も示した。表 3 を参照されたい。

表 3　誤解析の諸タイプの代表例

no.	前文脈	キー	後文脈	実際の語	区分	語彙	語彙素読	サンプル ID
1	十日はソウルに滞在する生活が続いた。#夜遊びは韓国にいても変わらない。# 「やくざは夜ひとりでは	寝ん	(もん)でっせ」# 取り引きの仲間が笑いながら遊びに誘った。#「やくざは金や。#金が力や」#先輩たちから	寝る	異語異訓	休む	ヤスム	PB12_00144
2	したら商標権を侵害されたと訴えたりしないのですか？#そんな事言ったら「かっぷぬーどる」や「カッパ海老	せ	(ん)」そっくりのお菓子で違いはハングルだけと言うのも有ります。#もちろん本家日本企業は全く関係有りません。#もともと	海老せん	一般名詞	為る	スル	OC05_02403
3	ありげに見つめ合った。# メグレはそのつるはしをもってキャビンにもどった。#それから一時間以上のあいだ、憲兵は	どし	(ん)、どしんという鈍い音を聞きつづけた。#「ねえ、きみ…」# ふたたびメグレは甲板の昇降口から顔を	どしん	オノマトペ	度する	ドスル	LBl9_00192
4	冷凍おにぎりを解凍してお茶漬け状にして頂く ちっちゃくて可愛いおにぎりだが、よく出来てる#昼下	たぶ	(ン)ナポリタンを頂いてしまいそう（胃の調子が悪くて気分が悪いのが解消し、復活してきた） ＞	多分	漢語副詞	食べる	タベル	OY14_54020
5	年生まれの人たちは、西暦何年には何%存命である”という資料ってありますか？#平均余命（へ	いき	(ん)よめい）とは、ある年齢の人々が、その後何年生きられるかという期待値のことである。#生命	平均余命	漢字読み	行く	イク	OC09_10652
6	貰います。#すみ# 中村はん、ほんまにもう色々と…#うめ# いややわ。#そんなに言われたらうち居る所がの	うなり	(まん)がな。#ホホホ…そらそうとな会長はん、いつぞやお話した、ホラ、大阪に…#通仁# 布教所を作る	無くなり（ます）	近畿方言	唸る	ウナル	LBo9_00027
7	買うぞ！#TOD2買うぞ！#なんでTODはないんだばか！#がんばれテイルズ超がんばれ！#そういえば	なり	(たん)から聞いたけど坊ちゃん、マンガでてるんだって…？#買うしかないじゃないですかぁっ（ダンッ！）#さっそく	なりたん	呼称（人名を含む）	成る	ナル	OY14_36367
8	兄ちゃんに言うたとばい。#あん時どがんかしとったら」# 「もうよかて。#うちは兄さんに感謝こそ	すれ	(恨ん)どることなんかこれっぽっちもなかとやっけん。#それより兄さんの言うごとそこで働いてみようかね。#仲居さんなら	（こそ）＋する	古典	擦れる	スレル	PB39_00182
9	えらくのんびりしてるんでやすねえ」# 「いやね。#これはちょいとおまえさんには分りにくい楽しみでね」# 「そう	で	(やすか)…」# 目吉は、少し不満そうな色を浮べたものの、いくつかの腑に落ちぬことを整理し	やすか	語尾	出る	デル	LBh9_00140
10	教育の面もありました。# 一説では、遊女は客をだます狐で、それも尾のない狐だから「	尾い	(らん)」だとか。# 傾城は美人の別称で、中国の故事からきています。# 漢の李延年が帝	要らぬ	駄洒落	付く	ツク	LBa3_00020
11	泣いて、学校から帰ってきた途端力が抜けて泣いてで・・・・。#俺はどんだけ泣いたら気が	すめ	(ん)orz でもこうやって毎日泣いていってるとさ、日に日に涙は少なくなっていってるような気	済む	同音異語	住む	スム	OY14_28469
12	良かったのだと思います。#楽器屋さんによってはいろいろ吹き比べも出来るので好みの音色とお値段を照らし合わせて	選ら	(ん)でみてはいかがでしょう？#もしかしたらクランポン以外にも素敵な楽器と出会えるかもしれませんよ。	選ぶ	同字異訓	選る	エル	OC01_02482
13	。#食わせる物がなくって屋根ェ上げて風ェ食らわしとこうとな、餓鬼に着せるものがなくってアンベラへ	くる	(ン)どこうと俺の勝手だい…何を言ってやん。#他人の財政ィ立ち入りやがらァ…。#嫌なら俺ァひとり	包む	動詞連用	来る	クル	PB29_00172
14	ま」と出る状態です。#直しかたを教えてください#ALTキーを押しながら、カタカナひらがなキーを押す。#くらすちみちらかかち	しい	(とん)ら#ほらなおったでしょ	くらすちみちらかかちしいとんら	特定	為る	スル	OC02_07788
15	こちらを見ている。#それでも僕にはかすかな震えが伝わってくるんだ。#ほら、池に小石を投げ込んだら	さ	(ざ)彼が立つだろう？# あんな感じでね。#僕は歩調を落として彼女の顔に自分の顔を近づけた。#距離	さざ波	入力ミス	為る	スル	LBi9_00023
16	偉大なマナのイメージが崩壊しちゃいそうだから。#だって自我消えるかも知んないジャン！#一生懸命育てたのに！#	つか	(どん)だけ不運なの！？#（アレンの自我が消えるなんて誰も言ってないから#本業に戻ってるラビ様#キャー――	つか	話し言葉	付く	ツク	OY14_12410
17	セリエAは「せりえあー」と呼ぶのに、なぜACミランは「えーしー	み	(らん)」と言うのですか？#セリエAの正式名称は Campionato Italiano del Calcio Serie-A で、Serie	ACミラン	表記仮名	見る	ミル	OC06_03990
18	そのわけを聞いたところが、軍艦に乗り甲板に起つてゐる時の練習なのださうであつた。#休憩時間に	し	(やがん)だり、凭れたりした者は罰せられる。	しゃがむ	表記仮名遣い	為る	スル	LBa9_00077
19	をやっているわけですね。#いわゆる志布志湾波見港の公有水面埋め立てに関連しての東串良町漁協総会の有効性	いか	(ん)、こういうことでありますが、この県議会等での議論、県当局のとっている態度、これらを含め	如何	表記漢字	行く	イク	OM21_00010
20	昌史。#彼はステーションキッズという事務所で大江千里のマネージメントを担当している強者だ。#いたって冷静にうけ流す。# 「	あら	(ん)、かわいいお店」# 「わああ、いい。#いいわあ」# 普段は男まさりにサブでマネージメントしているヒロミちゃんで	あらん	フィラー	有る	アル	LBg7_00053
21	の安らぎが破れる。#昔の飲食は空腹をみたせば足りた。#それを今では林を焼き池をさらえ、生物を	切りこま	(ざ)いているではないか。# 抱朴子が言う、# 物事は現在行き過ぎがあるからといって、すべて止めて	切り細裂く	複合語	切り込む	キリコム	LBe1_00021
22	NOVAキッズのCMですが、我が家では大うけです。#「I am エーと student」#「えーと	はいら	(ん)よ。」#というのです。#お宅ではうけているCM何かありますか。#杉田かおるさんが出ている「	要らぬ	複合要素	入る	ハイル	OC01_07191
23	はらちがあがないべ。#困ったごどだなあ～。#一関の観光にとっては大打撃だじゃな。#追い討ちを	かげ	(でん)のがこのガソリン高と光熱水費の値上がりだじゃ。#な～んか写真の内容ど載せでいる話がかみあわないがら、	かける	方言	嗅ぐ	カグ	OY11_01706
24	の荒しさんが 死ねると信じていた あふぉなアタシに 教えてくれたんで、（けんか腰で	すまそ	(ん))#いびきを かいてた 彼を起こした。#「ごめん」と 言った。# 「切ったか？」	すむ	若者言葉	澄ます	スマス	OY07_00095
25		こ	(ン)ばンわン♪つー#今日は雨だったから#珍しく1日家にいた～笑#まあタ方はぶらぶらしたけどねw	こんばんは	若者表記	来る	クル	OY14_50842
26	授業。#めんどくさい。#気分のらない。#まあ、授業に気分がのる日なんて無いけどね・・・。#とりあえず、今見てる	はがれ	(ん)1期全部見終わってから学校行きたい。#見る時期間違ったかな・・・・。#夏休みまで待って・・・	ハガレン	固有名詞	剥がれる	ハガレル	OY14_52453

４.２　名大会話コーパスとの比較

コーパスの解説では「機械的に形態素解析を行い，一部手修正を行った後，結果をタグ付けしてい」ると記されている。すべてがコアデータではないということになるが、撥音に限ってみると、表４の通りほぼコアデータに匹敵する解析精度に達している。

表４　名大会話コーパスにおける各ファイル誤解析の割合

no.	判別不可用例	近畿方言以外の方言	誤解析（キー）	誤解析（%）	後件誤解析（非撥音）	考察可能な対象例	各ファイル用例数
1	1	3	1	0.08%	0	1225	1230
2	0	0	0	0.00%	0	13	13
3	0	0	0	0.00%	0	241	241
4	0	0	0	0.00%	0	13	13
5	2	1	8	1.60%	0	490	501
6	0	0	0	0.00%	0	11	11
7	0	0	0	0.00%	0	135	135
8	0	0	0	0.00%	0	2	2
合計	3	4	9	0.42%	0	2130	2146

５．結びにかえて

解析システムは人間ではないが、言語を学ぶという意味では人間と同じく日本語学習者と見なすことができる。劉（2018）で示した、学習者にとって撥音に関する学習が難しいとされる箇所と比較すると、いわゆる標準語における話し言葉（話し言葉／若者表記）、書き言葉（古典）、準標準語（近畿方言／方言）等が共通して難しいということが言えよう。また、劉（2018）では考察対象としていなかったが、「駄洒落」と「特定」は日本語学習者にとっても判定が難しいタイプであると思われる。ただし、全体的に言えば、日本語学習者に比べ、解析システムが難しいと感じる種類の方が圧倒的に多いと見なすことができる。

謝　辞

本研究は基盤研究（C)「中国語話者から見たニア・ネイティブレベルを目指すための語彙に関する総合的研究」（16K02818）の助成を受けた成果の一部である。また、調査ではBCCWJ と名大会話コーパスを利用させて頂いた。開発関係者の皆様に謝意を申し上げる。

文　献

『現代日本語書き言葉均衡コーパス』利用の手引 第 1.1 版（第５章　形態論情報）
（http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html）2018 年 7 月 23 日最終確認
名大会話コーパス（全文検索システム「ひまわり」）
（https://mmsrv.ninjal.ac.jp/nucc/）2018 年 7 月 23 日最終確認
山崎誠(2013)「コーパスでできること２－BCCWJ を例に－」『日本語学』32-14、pp.104-116、明治書院
劉志偉（2016）「学習者の視点から見た「準標準語」文法項目について」『武蔵野大学日本文学研究所紀要』3、pp.53-69、武蔵野大学日本文学研究所
劉志偉(2018)「日本語教育の立場から垣間見たラ行音撥音化―日本語学習者の視点から―」『埼玉大学紀要（教養学部）』54-1、頁数未定、埼玉大学教養学部