首都大学東京都市教養学部人文・社会系
首都大学東京人文科学研究科
「人文学報」第513-7号
（日本語教育学）
2017年３月抜刷

語彙・文型調査を目的とした『幼稚園の配布文書コーパス』の作成について

―特定の目的コーパスの作成例として―

長谷川　守寿　西尾　広美

語彙・文型調査を目的とした『幼稚園の配布文書コーパス』の作成について
—特定の目的コーパスの作成例として—

長谷川　守寿・西尾　広美

1．はじめに

本稿は、『幼稚園の配布文書コーパス』の作成の手順を詳細に記述し、今後のデータ追加作成のための手順書となることを目指したものである。

現在、多くの幼稚園では日本語を母語としない保護者（NonNativeSpeaker保護者。以下、NNS保護者）が見られるようになったが、中には、日本語学習の機会が少なく、日本語が十分理解できないケースが出ている。そのような場合、幼稚園からの配布文書が正しく理解されないため、情報伝達や意思疎通がうまくいかずに保育活動に支障をきたす、という問題も出てきている（西尾2013）。

そこで我々は、地域や運営団体の異なる幼稚園で配布された文書を元に『幼稚園の配布文書コーパス』を作成し、語彙・文型調査を行い、将来的に教師とNNS保護者を結ぶ「保護者に伝わるやさしい日本語」のテキスト化や、NNS保護者が文書を理解する際に役立つ語彙表の作成などを予定している。本稿では、調査を行う前段階として、配布文書をどのようにテキストデータ化したのか報告し、今後のコーパスの規模拡大へ向けた手順書とする。

本稿のコーパスの利用目的は、語彙・文型調査が主であるため、作成の過程では語が正しく認定できることを優先している。より精度の高い語彙調査ができるようにするために、どのような作業をしているのか明らかにする。

2．『幼稚園の配布文書コーパス』の必要性

汎用のコーパスとしては、2011年より国立国語研究所が『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）のDVD版の配布を開始し、さらに少納言・中納言という検索サイトの公開を開始した。また特定目的のコーパスとしては、『日中Skype会話コーパス』(中俣2015）や『児童・生徒作文コーパス』(宮城・今田2015)、『学校お便りコーパス』(李2016）などのように、特定目的のコーパスも多数作成・公開されている。

しかし現在までのところ、我々の関心の対象である幼稚園の配布文書を収集したデータは存在しない。幼児教育の面から、その分野で使用されている用語集などにあたるという方法も考えられるが、そうした用語が実際の配布文書に使用されているのかというデータとしての真正性が保証されないため、採用できない。そこで実際の配布文書を元に、語彙や文型調査に向け『幼稚園の配布文書コーパス』を作成している。本稿ではその手順について述べる。

3.『幼稚園の配布文書コーパス』の設計と基本方針

3.1　対象幼稚園

本稿で説明する文書が実際に配布されたのは、都内にある公立幼稚園であり、3歳児クラスが1クラス、4・5歳児クラスが2クラスずつで、合計5クラスからなる。

3.2　コーパスの構成

対象とする文書は、この幼稚園で平成19年度（平成19年4月9日から平成20年3月13日）に、園児の保護者に向けて配布された文書93種類である（幼稚園内部の文書は対象外）。ページ数はA4用紙相当で228枚である（A3用紙1枚は、A4用紙2枚に換算）。なお、この期間に配布されたと考えられる資料『土と緑の○○幼稚園』(○○には実際の幼稚園名が入る)・『要覧』は、この幼稚園への入園を考えている幼児の保護者に向けた文書であり、入園に向けて準備する物の説明なども含まれ、非常に重要な配布文書と考えられるため、厳密にはその当時だけの園児の保護者向けではないが、対象とする。また後述する李（2016）の『学校お便りコーパス』に含まれるような、いわゆるお便りだけではなく、保護者会などの資料も含めている。これは、保護者が幼稚園で配布される全ての文書を理解するよう求められているからである。

文書はいろいろな観点から幾つかの種類に分類することができる。ここでは、便宜的に文書を書いた人という観点から、「あ. 園長が書いた文書」「い. クラス担任が書いた文書」「う. 保護者が書いた文書」「え. 園医が書いた文書」「お. 誰が書いたか不明な文書」の５つに分けて説明する。

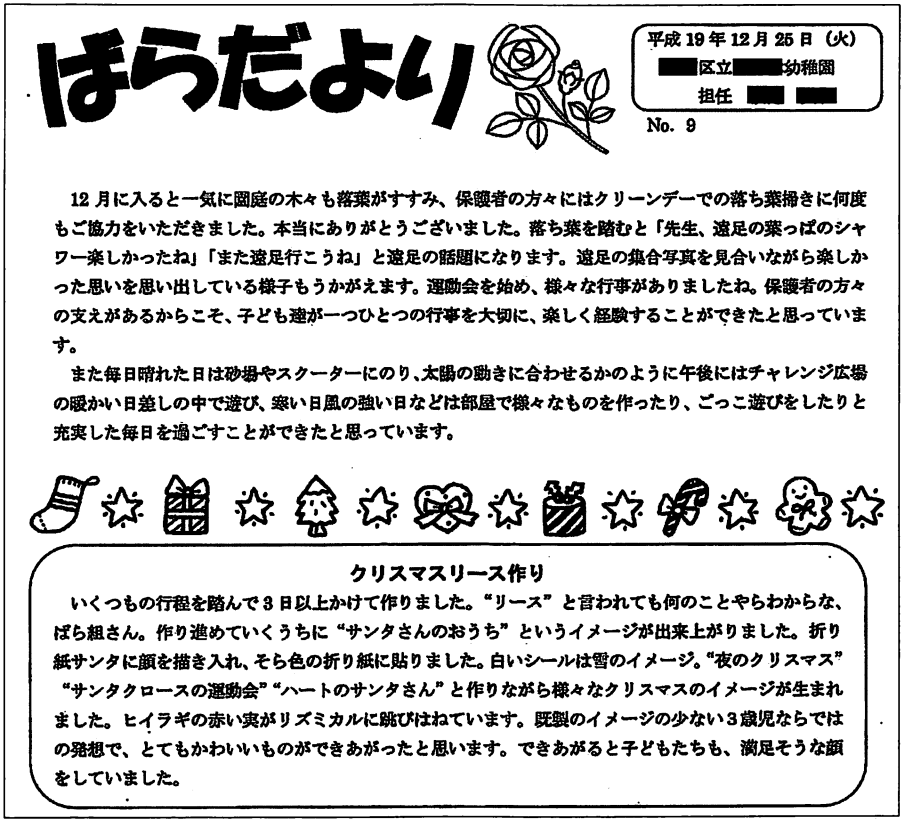
「あ. 園長が書いた文書」は、夏休み期間の8月を除いて毎月1回配布される文書で、園全体に配布される（理由は不明であるが、6月号のみ副園長が執筆）。その前の月に行われた行事の感想やそれに関連した話題が挙げられ、その月の行事予定や前月の園児の様子と当月の教育の目標、当月に誕生日を迎える園児の名前などが記載されている。他に「春休みの過ごし方」「冬休みの過ごし方」、父親と園児が遊ぶ行事、祖父母を招く行事など、園全体に関連する文書が含まれる。

「い. クラス担任が書いた文書」は、各クラスの担任が書いたもので、運動会や「○○まつり」など行事に関する内容が中心である「＜クラス名＞便り」(図 1参照）と、学期末に行われた保護者会の資料（全体・クラス別)、「保育参観のしおり」等が含まれる。「＜クラス名＞便り」はクラスにより配布にばらつきがあるが、多いクラスで年9回、少ないクラスで年6回配布されている。

「う. 保護者が書いた文書」は、幼稚園が企画したイベント（講演会・保育参観など）後に保護者が提出した感想を幼稚園の教諭がまとめたものである。「え. 園医が書いた文書」は、保健便りであり、「お. 誰が書いたか不明な文書」は、署名文書ではないもので、「園児募集」「幼稚園案内」「夏休みのしおり」「電車に乗るときのマナー指導」などが含まれる。この中には歴代の園長が担当されたものと思われるものを入れた。

内訳は、「あ」が52枚、「い」が130枚、「う」が5枚、「え」が1枚、「お」が40枚、計228枚で

ある。



ばらだより

平成19年12月25日（火）
[redacted]区立[redacted]幼稚園
担任 [redacted]

No. 9

12月に入ると一気に園庭の木々も落葉がすすみ、保護者の方々にはクリーンデーでの落ち葉掃きに何度もご協力をいただきました。本当にありがとうございました。落ち葉を踏むと「先生、遠足の葉っぱのシャワー楽しかったね」「また遠足行こうね」と遠足の話題になります。遠足の集合写真を見合いながら楽しかった思いを思い出している様子もうかがえます。運動会を始め、様々な行事がありましたね。保護者の方々の支えがあるからこそ、子ども達が一つひとつの行事を大切に、楽しく経験することができたと思っています。

また毎日晴れた日は砂場やスクーターにのり、太陽の動きに合わせるかのように午後にはチャレンジ広場の暖かい日差しの中で遊び、寒い日風の強い日などは部屋で様々なものを作ったり、ごっこ遊びをしたりと充実した毎日を過ごすことができたと思っています。

クリスマスリース作り

いくつもの行程を踏んで3日以上かけて作りました。“リース”と言われても何のことやらわからなばら組さん。作り進めていくうちに“サンタさんのおうち”というイメージが出来上がりました。折り紙サンタに顔を描き入れ、そら色の折り紙に貼りました。白いシールは雪のイメージ。“夜のクリスマス”“サンタクロースの運動会”“ハートのサンタさん”と作りながら様々なクリスマスのイメージが生まれました。ヒイラギの赤い実がリズミカルに跳びはねています。既製のイメージの少ない3歳児ならではの発想で、とてもかわいいものができあがったと思います。できあがると子どもたちも、満足そうな顔をしていました。

図1. 実際の配布物の例（一部）

3.3 基本方針

紙の文書をテキスト化する際、レイアウトは無視し、文は文の形式で、箇条書きは箇条書きというように、そのまま入力することを基本とした。イラスト等は入力しない。表がある場合も語句のみ入力し、表形式では入力しない。表記の多様性を調べる目的ではないため、フォント情報等も考慮しない。

個人情報に関わる部分（個人が特定される可能性のある語句や氏名、呼び名など）は、全て“山田太郎”“太郎”で置き換え、幼稚園に関わる語句は全て“南大沢”に置き換える。これは、後に形態素解析を行うので、“○○”などの記号で置き換えた場合、正しく解析できなくなる可能性があるため、正しく人名・地名と形態素解析できるように“山田太郎”“南大沢”としたものである。

また本目的を遂行するために、配布文書をそのままテキスト化したのでは形態素解析で正しく語の境界を認定できない場合には、正しく認定できるようにテキストに修正を加えることとした。これについては4.3.4で詳述する。なおこれ以後、紙の状態のものを"プリント"、電子化されたものを"テキスト"と呼ぶこととする。

4.『幼稚園の配布文書コーパス』の作成法

コーパスの作成法は、以下の示すとおりである。本稿では、図2の流れに沿って『幼稚園の配布文書コーパス』の作成方法について述べる。

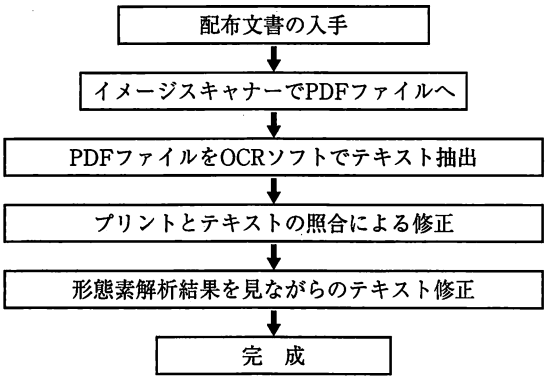


図2.『幼稚園の配布文書コーパス』作成の手順

4.1 配布文書の入手

本稿で扱う幼稚園の配布文書の資料は、当時園長であった人から、平成19年度に園が保護者に配布した一年分の資料全てを提供してもらったものである（提供者の希望により名前は伏せる）。園長が年間の記録として保管しているということからも、真正性・網羅性の観点でも問題がないと考える（研究使用の許諾も得ている）。配布文書は全てコピーし、実物は所有者に返却した。これ以降、配布文書と言及する際は全てコピーしたものを指す。

4.2 イメージスキャナーでPDFファイルへ

イメージスキャナー（以後、スキャナー）を用いて配布文書のイメージを取り込み、PDFファイルにして保存する。イメージから直接文字を抽出することも可能であるが、管理が容易になることと、共同研究者との研究にあたり、実物のコピーを使用するだけでなくイメージファイルで確認する手段を確保することが合理的と考え、PDFファイルにした。本研究ではフォントの色等は研究の対象外となるので、白黒でスキャンこととした。

4.3 PDFファイルをOCRソフトでテキスト抽出

OCRソフト（『読取革命ver15』を使用）でPDFファイルとして保存された画像を文字化する。ファイル形式はテキストファイルにする。ファイル名は現在、スキャナーの設定のままである。今後、配布物の分類も含め、どのようなファイル名が適当か考慮が必要となる。文字コードはS-JISを採用する。これは、今後多くの研究者にデータを使用してほしいためであることと、JISコードでカバーできない漢字は「髙﨑」の「﨑」、「栁田」の「栁」など人名に多く出現するが、前述したように人名に関してはプライバシーの保護の観点から全て「山田太郎」にするため、人名漢字を考慮する必要がなくなることによる。なおこの文字コードを選択したことにより発生した問題については、「4.3.4 表記を変更したもの」で述べる。

手書きされたお知らせについてはOCRソフトを用いても正しく文字認識ができないので、キーボード入力を行う。

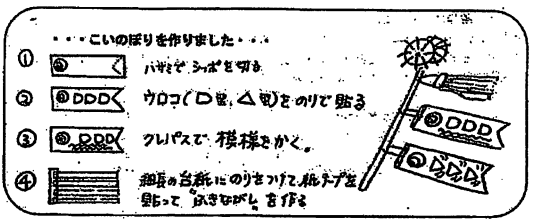


図3．手書き文書の例

例えば図3のような手書き部分は、以下で示した（1）のようにテキストを入力した。なお、例文の出典は、全て幼稚園で配布された文書であるが、文書のタイトルを示すと園が特定される可能性があるため、表示しないこととした。

（1）　・・・こいのぼりを作りました・・・
　①ハサミでシッポを切る
　②ウロコ（D型、△型）をのりで貼る
　③クレパスで模様をかく
　④細長の台紙にのりをつけて紙テープを貼って"ふきながし"を作る

4.3.1 プリントとテキストの照合による修正

OCRソフトにより作成されたテキストの修正を行う。プリントとテキストを照合することで、OCRソフトの誤りを発見する。OCRソフトを用いたコーパスデータの作成法に関しては三井（2011）が詳しい。三井（2011）では、原稿ごとにどのような誤りが見られたか述べられていて、今回作成しているコーパスでも、"「"（2バイト文字）を"｢"（1バイト文字）のように認識する誤りが見られた。ここでは、幼稚園の配布文書を扱う場合に見られた誤りを挙げておき

たい*1。便宜的に以下のAからCのように分けることができる。(2) から (4) に示すような誤り（矢印の左側）が見られたので、正しい形（矢印の右側）に変更した。

A. 清音・濁音・半濁音などを含む語の認識が正しくないもの
B. 小書き文字を含む語の認識が正しくないもの
C. 字体の似ているもの（A・B以外）

Aに含まれるものには、(2) の「キッズ」が「キッス」、「たんぽぽ」が「たんぽぽ」という形で認識されていたので、右側の正しい形に修正した。Bに含まれる誤りには、(3) の「よ」と「ょ」のような小書き文字を含む誤りである。Cに含まれる誤りには、(4) の「せ」と「さ」、「一」と「ー」、「間」と「間」のように、三井 (2011)・田野村 (2014) で指摘されているものも含まれたが、「於」と「1を」のように字体が似ていると考えにくいものまで含まれていて、確認の作業では注意が必要である。

(2) キッス＝＞キッズ、たんぽぽ＝＞たんぽぽ、ひろぱ＝＞ひろば、なるぺく＝＞なるべく、楽しんたり＝＞楽しんだり
(3) でしよう＝＞でしょう、ペインテイング＝＞ペインティング
(4) 年長せん＝＞年長さん、み一つけた＝＞みーつけた、時間＝＞時間、終巣式＝＞終業式、1を＝＞於

このようにプリントとテキストを照合していく作業をしていく中で、OCRソフトは正確に文字が抽出されているのであるが、中には元々プリント自体に誤字・脱字があったり、文法的な誤りがあるものがあることが分かった。

コーパス作成の方針としては、なるべくプリントに使われる語を正確に反映し、そのまま文字化することを優先するが、我々の目的が語彙表・文型リストの作成であるため、使われている意図を反映した語の抽出が必要となる。そこで前後の文脈から明らかな誤りと筆者らが判断したものは、テキストを正しい語・表記に修正することとした。

例えば、誤字脱字の例には、(5) のような例に仮名漢字変換ではよく見られるものから、前後の文脈を確認しなければ誤りと気づけない (6) のような例が見られた。また (7) のような活用の問題だけでなく、(8) のように判定が難しいものも含まれたが、著者らで相談し、前後の文脈から誤りと考えられるもののみ修正した。

また、対象とする文書は複数の教員によって書かれていることから表記のゆれが生ずる。たとえば、「チーム」と「ティーム」、「貸し出し」と「貸し出」のようにゆれが見られた。そこで誤解析を防ぐために事前に「チーム」「貸し出し」に統一した。

*1 BCCWJに見られた、機械的文字認識や手入力によるテキストの誤りについては、田野村 (2014) が詳しい。

(5)　スッテプ＝＞ステップ、として＝＞とおして、ゲー＝＞ゲーム、うっこつけい＝＞うこっけい、園庭解放＝＞園庭開放、機械＝＞機会、年長時＝＞年長児、あったたかくって＝＞あたたかくって、ジョカー＝＞ジョーカー、ちょとんと＝＞ちょこんと、めいいっぱい＝＞目一杯、絵本舗読み聞かせ＝＞絵本の読み聞かせ

(6)　テレビ付け＝＞テレビ漬け、持ち返す＝＞持ち返らす

(7)　ふれたりかかわりして＝＞ふれたりかかわったりして
楽しくようです＝＞楽しいようです
寄付していただきけるとありがたいです＝＞寄付していただけるとありがたいです

(8)　家族に一員として＝＞家族の一員として
子供に育てたいことを明確に捉え＝＞子供に育てたいことを明確に伝え

(9)　チーム／ティーム＝＞チーム、貸し出し／貸し出＝＞貸し出し

なお、この段階での修正には「ドッチボール」「少しづつ」のように厳密には誤った形であるが、多く使われているため、もはや誤りと考えられないものも含まれている。ここでは、正しい形の方がNNS保護者は辞書などで調べやすいであろうと推測し、「ドッジボール」「少しずつ」に変更した。なお、これらは次の形態素解析の際にも誤りとなるため、修正されることとなる。

4.3.2 形態素解析結果を見ながらのテキストの修正

前述のようなプリントとテキストのチェックを我々で二度行ったが、OCRソフトを使って文字認識を行い作成したテキストの中から“ー”と“一”“カ”と“力”　など、全ての誤りを目視で発見することは困難である。そこで、形態素解析システムにかけて、形態素解析結果を検証する中で、テキストの入力の精度を上げ、表記を修正し、語彙調査に適したコーパスに変えていくこととする。

また長谷川・西尾（2016）の調査では、幼稚園のお知らせは、子供向けの文章ではないのだが、通常の表記に比べて、漢字よりもひらがなで書かれることが多いなど、いくつかの特徴を指摘した。このテキストを形態素解析にかけると、誤った結果が出てしまい、正確な語彙調査ができなくなってしまう。小木曽（2014）では、形態素解析が失敗する例について以下のように述べている。

> （前略）かな書きするだけで未知語になってしまうため、子供向けの簡単な文章も解析エラーが多くなる場合がある。たとえば「めいたんてい（名探偵）」のように、漢字で書かれるのが普通である語が、子供向けにひらがなで書かれている場合である。（略）子供向け表記のような一般的ではないものは辞書にないのが普通であり、そのために簡単に見える文章でかえって失敗してしまうのである。（pp.108-109）

以上のように、正しく文字認識されていないことによって、またひらがなで書かれることによって、形態素解析器では正しく解析できない可能性がある。そのため正確な語彙調査を行うには、入力や表記の修正・変更を行い、正しく解析できるようにする前処理が必要となる。

そこで我々は、形態素解析に使用する形態素解析器にMeCab（mecab-0.996.exe）[*2]、形態素解析用辞書にUniDic-mecab[*3]（ver2.1.1）を使用し、入力したテキストを形態素解析する。その結果を基に、正確に語に区切ることができるか確認する。例えば、形態素解析をすると、(1)の①「ハサミでシッポを切る」は表1のような形で出力される（必要な部分のみ表示する）。

表 1. 形態素解析の実際（正しい解析の例）

書字形	語彙素読み	語彙素	品　詞
ハサミ	ハサミ	鋏	名詞-普通名詞-一般
で	デ	で	助詞-格助詞
シッポ	シッポ	尻尾	名詞-普通名詞-一般
を	ヲ	を	助詞-格助詞
切る	キル	切る	動詞-非自立可能

形態素解析を行い、語の境界・品詞・見出し語（UniDicでは語彙素）といった結果を確認する段階で、修正を行うことが必要となる箇所が多数見られる（ここでは、正しく解析できているとは、少なくても語の境界・語の品詞・見出し語の認定が正しいことを意味し、見出し語の読みについては問わないこととした）。そこで後述の「トマトを食べる」や「ボタンかけを練習する」の解析結果（表2）を用いて、正しく解析できていない場合について説明する。

表 2. 形態素解析の実際（誤解析の例）

書字形	語彙素読み	語彙素	品　詞
トマ	トマ	トマ-Thomas	名詞-固有名詞-人名-一般
ト	ウラナイ	占い	名詞-普通名詞-一般
を	ヲ	を	助詞-格助詞
		＜略＞	
ボタン	ボタン	ボタン-button	名詞-普通名詞-一般
かけ	カケ	欠け	名詞-普通名詞-一般
を	ヲ	を	助詞-格助詞
		＜略＞	

表 2で「トマト」は、見出し語で「トマ-Thomas」「占い」となっており、入力が不正確な場合、正しく語に区切れていないことから、ここに何らかの問題があることが分かる。また、「ボタンかけ」は見出し語では「ボタン-button」「欠け」となっており、語彙素は「掛ける」である

[*2] http://download.unidic.org/?page_id=20
[*3] http://pj.ninjal.ac.jp/corpus_center/unidic/

べきであり、正しく語彙素が確定できていない（問題のある部分はセルに色づけした）。このような方法で誤りを発見し、その箇所を修正・変更する作業を現在までに三回行った。その結果どのような誤りがあり修正・変更したかを、OCRソフトの問題と表記等に由来する問題に分けて述べる。

4.3.3 OCRソフトによる誤認定の修正

まずはOCRソフトの認識の問題である。上記作業の人手による確認作業では見逃してしまったが、形態素解析の結果を確認して明らかになったものには、以下のような例がある。例えば、上でも触れた（10）の「トマト」と「トマト」は非常に字体が似ている。しかし"トマ"はカタカナで、"ト"は漢字である。このような例として「コート」があった。また"り"と"り"は前者がひらがなで後者がカタカナである。"口"はくにがまえ、"口"はクチである。目視ではフォントの微妙の違いしかない誤りは確認できなかったが、形態素解析にかけ、結果を確認し、品詞や見出し語が想定しているものと異なっているものを発見することで、（11）のような目視では見逃してしまった誤りにも気づき、修正を行うことができた。この作業によりテキストの精度向上が期待できる。

(10)　トマト＝＞トマト、べりー＝＞ベリー、口＝＞口
(11)　教青＝＞教育、雲悌＝＞雲梯

この作業は、使用した『読取革命ver15』の設定を変更したり、使用するOCRソフトそのものを変更することでも正しく認識できる可能性もあるが、入力の質を高める作業はいずれにせよ必須となるであろう。

4.3.4 表記を変更したもの

形態素解析を行い、表1表2のように、語の境界・品詞・見出し語といった結果を確認することで、変更が必要となる箇所が数多く見られた。

そこで形態素解析の結果を検討し、実際に別表記で茶まめ[*4]に入力して確認しながら、テキストを変更した。ここでは変更を便宜的に以下の3タイプに分け、説明する。

M１：文字種の変更
1．ひらがな表記だったものを漢字表記に変えたもの
2．ひらがな表記だったものをカタカナ表記に変えたもの
3．カタカナ表記だったものをひらがな表記に変えたもの
4．カタカナ表記だったものを漢字表記に変えたもの

[*4] 形態素解析用辞書UniDicを使って形態素解析を行う作業を容易にするソフトウェアで、UniDicのWindows用パッケージに含まれている。

5．漢字表記を別の漢字表記に変えたもの

M2：音引きの変更

M3：出現環境の変更

M1：文字種の変更

まず、M1－1に該当するひらがな表記を漢字に変更したものは、(12)(13) のように多数存在する。左側の括弧内には、そのまま解析した場合にどのように解析されるかを示した。

(12) おかずはいりません（そのままでは「入りません」と解析） ＝＞ おかずは要りません
ボタンかけ（「欠け」）＝＞ボタン掛け
テーブルふき（「吹き」）＝＞テーブル拭き
コップについで（「継いで」）＝＞コップに注いで
こげる（「漕げる」）＝＞焦げる
○○だより（「だ／より」）＝＞○○便り
友だちとかかわり（「とか／かわり」）＝＞友だちと関わり
くらいよみち（「くらい［助詞-副助詞］／よ［助詞-終助詞／道］」＝＞暗い夜道
うこっけい（う［感動詞］／滑稽）＝＞烏骨鶏
しっぽとり（しっ［感動詞］／ぽとり［副詞］）＝＞しっぽ取り
○○ぐみ（○○グミ）＝＞○○組、
(13) おわん＝＞お椀

なお、(13) の「おわん」は文頭では「お（感動詞）／わん（副詞）」と解析されるが、「左手でおわんをもつ」のように文の形では「御」「椀」と正しく解析できる[*5]。(13) は幼稚園のあるイベントを説明する文書において、必要となる持ち物リストの中にあり、文頭に出現しているものであるため変更した。

次に、「M1-2ひらがな表記だったものをカタカナ表記に変えたもの」について説明する。これには (14) が該当する。例えば、“どろけい”は、ひらがなのままでは（泥＋ケイ［人名］）と解析されるが、カタカナ表記に変えると「泥警」[泥棒と警察] と正しく解析できる。“すだしい”は、スダシイとカタカナ表記にすることによって「すだ椎」(ブナ科の常緑高木でシイの一種）と一語に解析できる。ただし文中での「なす」は正しく解析できるが、括弧の後の「なす」は動詞と解析されるため、その出現位置に現れる“なす”のみ変更した。

(14) どろけい＝＞ドロケイ、すだじい＝＞スダジイ、なす＝＞ナス

[*5] UniDicでは、どこに位置するのか（文頭か文中か）によって、連接コストが異なるため、出現位置を考慮した対応が必要となる。

同様に、「M1－3カタカナ表記だったものをひらがな表記に変えたもの」(15)、「M1－4カタカナ表記だったものを漢字表記に変えたもの」(16)、「M1－5漢字表記を別の漢字表記に変えたもの」(17) を挙げることができる。(17) は、文脈では「たくさん休んで下さい」という意味で用いており、「十分」でも「じゅうぶん」という読み方は存在するが、形態素解析では数字としてしか解析されないため、「充分」に変更した。

(15) シトシト降る＝＞しとしと降る
(16) ヨウシュヤマゴボウ＝＞洋種山牛蒡、ダンボール＝＞段ボール
(17) 十分休む＝＞充分休む

この他に、テキストの変更において、文字コードの問題が1件発生した。S-JISを採用し、正しく語が認定できるように変更する作業を行ってきたが、どうしても変更できない問題が生じた。「鼻をかむ」という表現は、語彙素のレベルでは「鼻」「を」「嚙む」と解析される。正しく「擤む」と判定されるには、テキストを「擤む」に変えなければならないが、S-JISでは保存できず、UTF-8などで保存するしかない。テキストがUTF-8を採用していれば正しく語彙素を判定できるのであるが、現在は変更できていない。現在1件のみであるが、複数出現した場合は文字コードの再考も必要となろう。

M2．音引きの変更

一部の語の音引きについては誤った形態素解析の結果になるので、(18) のような変更を加えた。ただし、UniDicには「だーいすき」「できなーい」「ずーっと」「はーい」「よーい」など音引きされた形が辞書の書字形に登録されている語もあり、正しく解析できる語は変更していない。

(18) いただきまーす＝＞いただきます、おいしーい＝＞おいしい、はいりたーい＝＞はいりたい、てんきにな～れ＝＞てんきになれ、楽し～い＝＞楽しい、すご～い＝＞すごい

M3．出現環境の変更

UniDicは、単語（短単位）に区切られたものの組み合わせの中からコストが最小の組み合わせを正解として出力するという特徴がある（小木曽2014)。本調査の例で説明すると、「①節分」は、「①／節分」よりも、コストが小さい「①／節／分」が解析結果として選ばれる。2月の豆まきの予定で出てきた表現なので、このように語を認定されるのは正しくなく、「節分」で正しく語に区切ることができるように工夫する。この場合は“①”と“節分”の間に読点“、”を入れると、「①／節分」と正しく語に区切れることが確認できたので、読点を挿入し、正しく語に区切ることができるようにしておく。読点“、”や空白“□[*6]”ならば、記号として解析され、数

[*6] ここでは見やすさを考慮し、2バイト空白“　”を表すのに“□”を用いる。

える際に対象から外せばよいので、正しく解析でき、語数に影響しないからである。

例えば（19）は、そのままでは「違い（名詞-普通名詞-一般）」と解析されるが、読点を入れることによって、「違う（動詞-一般）」と正しく解析することができる。また、（20）は、語彙素レベルでは「水揚げ／良い／ね」と解析されてしまうが、読点を挿入することによって、「水／上げる／ね」と正しく解析することができる。なお、（21）のような例の場合、読点の挿入では「学びや」は「学舎（まなびや）」と解析され結果は変わらない。そこで、このような場合は空白"□"を挿入した。この方法により、専門性が高く独特な語でかつ臨時一語的な語も、語の境界を正しく認定できると思われる。例えば（22）は、そのままでは「新年／中」となるが、正解は新しい年中児という意味なので、「新／年中」となるのが正しい。そのため「新□年中」とした。

(19)　製作とは違い大掛かりな作業です＝＞製作とは違い、大掛かりな作業です
(20)　水あげようね＝＞水、あげようね
(21)　その後の学びや創造性が＝＞その後の学び□や創造性が
(22)　新年中＝＞新□年中、弁当時＝＞弁当□時、再任用主事＝＞再□任用□主事

UniDicの開発がさらに進み、より精度の高い解析結果が出せるようになったとき、上記のような前処理は必要なくなるかもしれないが、当面UniDicを用いて調査を行うには、必須となる処理であろう。特に幼稚園の配布文書のような、かなり特殊なテキスト化したものを対象とする際には、辞書の書字形に含まれている表記も考慮する必要が出てくる。

5．既存コーパスとの比較

ここでは、代表的なコーパスで明らかになっている問題点を参考に、本稿で作成している『幼稚園の配布文書コーパス』の特徴を考えてみたい。

まず現在日本語の代表的なコーパスとして『現代日本語書き言葉均衡コーパス』(略称BCCWJ）を挙げる。田野村（2014）では、BCCWJの特徴として「コーパス・サブコーパスのサイズ」「サンプルのサイズ」「サブコーパスの性質の差」「データの重複」「数表現の問題」「テキストの誤り」「付与情報」のような点を挙げている。

サイズについては、元々限定された枚数のものであり、またサブコーパスも存在しない。数表現はそのまま入力しており、用例数も少ない。テキストの誤りは、上述の通り修正は済んでおり、付与情報はない。残る問題として「データの重複」を取り上げる。

目視による調査であるが、『幼稚園の配布文書コーパス』にもデータの重複の問題が存在する。93文書中6文書であるが、これは全て年長クラスの学期末保護者会資料であり、以下に示すように重複が見られる（網掛けの部分のみ異なる）。例えば（23）と（24）の違いは、句点が1つと「そこで、今年は」の有無だけである。なお、重複が多く見られるのは3学期の保護者資料だけで、1・2学期の資料ではそれほど著しくない。

(23) ○○組・●●組の学級の枠を超えて、保護者が協力し、子どもたちを楽しませてくれたこと、教師に対しても、様々な形で感謝の気持ちを表してくれたことに、感謝。(略)子ども会を終えてからの子どもたちは、次への遊びに意欲的になった。今まで挑戦していたこま回しや縄跳び、一輪車に、さらに上を目指して磨きをかけている。お別れ会の中で得意技大会をした。

(24) ○○組・●●組の学級の枠を超えて、保護者が協力し、子どもたちを楽しませてくれたこと、教師に対しても、様々な形で感謝の気持ちを表してくれたことに感謝。(略)子ども会を終えてからの子どもたちは、次への遊びに意欲的になった。今まで挑戦していたこま回しや縄跳び、一輪車に、さらに上を目指して磨きをかけている。そこで、今年はお別れ会の中で得意技大会をした。

これは教育内容が同一であったとか、合同授業が行われたことも考えられるため、他のお知らせから理由を推測する必要がある。また、内容の重複が語彙調査にどのような影響を与えるのか不明であるが、調査を行う際には充分留意するべき点であると思われる。

また、「2.『幼稚園の配布文書コーパス』の必要性」で述べたように『幼稚園の配布文書コーパス』と類似するコーパスは、今のところ存在していない。そこで、「学校の配布文書」という観点から共通するコーパスとして、森(2014)と李(2016)を取り上げる。森(2014)が対象としたデータは兵庫県神戸市と石川県金沢市の小学校2校で配布された配布文書で、李(2016)の『学校お便りコーパス』は、平成24年度から26年度にかけて「兵庫県神戸市，大阪府大阪市，福岡県福岡市，福井県福井市の4つの自治体から延べ810枚(総文字数880,869字)の配布物を収集し、構築したもの」[*7]である。共に、小学校で配布された配布文書を扱っており、兵庫県神戸市は共通しているが、どのようにして収集したのか、どのようにして文字化したのかは明らかではない。ここでは、特定の目的のコーパスとして、語数の比較を行う。

森(2014)と李(2016)は、バージョンは異なるが形態素解析器にMeCab、形態素解析用辞書にUniDicを使用して、延べ語数と異なり語数を算出しており、森(2014)が使用したデータは、延べ語数421,560語・異なり語数9,299語(MeCab0.98及びUniDic1.3.12)、李(2016)が使用したデータは延べ語数471,212語(MeCab0.96及びUniDic2.1.2)で、異なり語数については不明である。

比較のために、暫定値ではあるが、本稿で作成している『幼稚園の配布文書コーパス』をMeCab0.996とUniDic2.1.1を用いて形態素解析し、延べ語数・異なり語数を求めた。空白、補助記号は除外し、未知語は全て修正した値であるが、延べ語数で102,519語、異なり語数で5,330

*7 http://lixiaoyan.jp/database/

*8 なおこの数値はさらに修正が必要である。たとえば、「巧技台」「マルチパネ」「温飯器」など、幼稚園でだけ使われる語については、短単位として一語なのか二語以上なのか、辞書作成の基準を参考に考える必要がある。また歌の歌詞などでも見直しが必要となる箇所がある。

語である*8。

延べ語数では、『学校お便りコーパス』の4分の1程度と少ないが、それに比べて異なり語数はかなり多い。これは、園児の様子を書いた毎月のお便りから、保護者会資料・保健に関する資料まで多種多様な文書を対象としてテキスト化していることによるものと思われる。また、枚数も4分の1強で、『学校お知らせコーパス』を基準にするならば、さらに3園くらい追加が必要となるであろう。

ただ『学校お便りコーパス』は公開されているため、我々もダウンロードして一部を確認してみたが、数値はどのくらいの精度なのか疑問に思う部分もある。実際のコーパスを見ると、"／ロウイルス（ノロウイルス）"、"ペーパータオル（ペーパータオル）"など、入力の問題が見つかる。また「低温やけど」のような場合、ひらがな表記を漢字表記に変更しないと、MeCabとUniDicを用いて形態素解析を行ったとき、誤った結果が出てしまう場合があると思われ、その対処法について明記されていない。そのためどこまでこの数値を信じていいのか、判断に迷う部分がある。

この他に、「幼稚園」という観点から共通する『連絡帳コーパス』も存在する。山形県山形市などで収集した外国出身の母親と保育園・幼稚園との連絡帳を対象としたもので、延べ語数8,456語、異なり語数1,016語である（森2014）。配布物とは性格が異なるものであることもあるが、コーパスは公開されておらず、今回は入手することができなかったため、比較の対象からは外した。

このように類似のコーパスと比較すると、本稿で述べている『幼稚園の配布文書コーパス』は、語数・ページ数は4分の1と少ないが、形態素解析結果を検証し、正しく形態素解析されるように、コーパスに前処理を加えていることが特徴と言える。またUniDicの連接コスト・生起コストの学習データになったのはBCCWJであり（小木曽2014）、幼稚園の文書や学校のお便りが対象になったことはない。新しく形態素解析を用いて語に区切る調査を行うデータが、UniDicの学習データになったことがない場合は、語の認定を行う際に上記のような手間のかかる処理が必要となり、そのような処理を経ないと語の認定が正しくできないのではないかと考える。

6．まとめと今後の課題

本稿では、作成中の『幼稚園の配布文書コーパス』について、その作成法を述べてきた。現在存在するコーパスの中の『YNU書き言葉コーパス』には、「データベースでの検索の利便性を主な目的として」一文を一行に変え、不要な改行・空欄を削除し、誤漢字や送り仮名を適宜修正したデータが含まれている（金庭・金澤2014、p.16）。また中納言でアクセスできる『現代日本語書き言葉均衡コーパス　通常版』は、「「１９９９年」のように数字を含んだテキストを形態素解析するために、事前に「千九百九十九年」のように形態素解析しやすい形にテキストを加工」したものである*9（加工されていないデータを使うには『現代日本語書き言葉均衡

*9 http://pj.ninjal.ac.jp/corpus_center/ot.html

コーパス非 numTrans 版』を使う必要がある)。

またコーパスについて、石井・杉本 (2014) では以下のように述べている。

> コーパスは、本来、ある研究課題 (リサーチクエスチョン) を設定した者が、そのために必要なデータとして自ら設計・構築して利用するものである。言い換えれば、コーパスはつねに特定の研究課題と結びついている、あるいは、結びついているべきであり、その限りにおいて、つまり、McEnery & Hardie (2012) も言うように、「そのコーパスの構築時に意図されていたリサーチクエスチョンを扱う限りにおいて、コーパスは最大限に活用しうる」のである。(石井・杉本2014,p.1)

著者らは、このような考え方に強く同意するのであるが、このようにコーパスを捉えた場合、著者らの目的は「どのような語・文型が使われるか」であり、「配布文書はどのような表記で書かれるか」ではないため、特定目的のコーパスはその目的に応じて表記の変更などテキストを加工するのは適切な処理であると考えるが、修正したデータのみで構成されるのがこのコーパスの特徴ともいえる。

また本稿では、正しさについて語の境界と品詞のみとし、読みについては不問とした。今後語彙調査を実施するにあたり、語彙素読みの修正を行う必要がある。例えば、「年中」について、文書の中では (25) のように、"ねんちゅう"と読ませるもののみであり、"ねんじゅう"と読ませる例は一例もない。しかし形態素解析の結果、語彙素読みは"ねんじゅう"であるため修正が必要となる。同様の例が「お母様」(26) であり、読みは"おははさま"である。これはUniDicの問題でもあるが、修正が必要となる。また (27) の"お家"は (28) のように「おうち」と読ませる意図と思われるが、解析結果は"おいえ"である。語彙素の読みに誤りを含むものは他にも多数存在するため、修正が必要となる。

(25) 1学期は年長組を中心に行い、徐々に年中組にも広げていく予定です。
(26) 先日は、お母様方の協力のもと、楽しい夕涼み会ができました。
(27) お家でも遊んでいるようなおもちゃを用意して安心して過ごせるようにしている。
(28) 自信をもって取り組んだ姿におうちでも「がんばったね。すてきだった」と誉めてあげてください。

今後は、読みも含めた短単位の語彙表を完成させ、長単位での語彙表を作成したいと考えている。これは語彙表作成を見すえた場合、単語表は長単位を基に作成した方が望ましいと考えるからである。例えば、短単位で「連絡」という語が104回使われていたという結果よりも、長単位で「連絡網」が21回使われていたというほうが、配布文書での実態を反映しており、今後有益な情報になると思われるからである。これには、短単位を元に長単位を認定する解析器 Comainu (https：//ja.osdn.net/projects/comainu/) を用いる予定であるが、元となる短単位

が正確に区切られていなければ、正確な長単位も認定できないため、短単位の認定の精度を上げる必要がある。

また、今回提供してもらった幼稚園の配布文書は、年少・年中・年長全てのクラスのお便りが含まれていている。園児の保護者は自分の子供の学年のお知らせしか配られないことを考えると、配布側に立ったデータで、BCCWJでいうところの「生産実態」を反映したデータといえる。しかし、実際の保護者は自分の子供の属するクラスからのお便りのみで、別の学年やクラスからのお便りを入手することはない（実際、受け取る側には「5．既存コーパスとの比較」で言及した重複の問題は発生しない）。今回の配布コーパスは、1年間の文書のみで、保護者の配布文書の受け取り実態（3年間）とは乖離していて、配布期間に特徴的な語（例えば、気候や出来事）の影響が出る可能性もある。そこで、保護者の側の立場に立った配布文書のデータが必要となる（これはBCCWJの流通実態を反映したコーパスに該当する）。これについては、保護者として子供を3年間幼稚園に通わせた際に配布された文書を所持している。本稿の対象幼稚園と同じ都内にある幼稚園ではあるが、運営は公立ではなく私立の幼稚園であるため、使用されている語彙が異なる可能性もある。園によってどのように語彙が異なるのか明らかにするためにも、コーパス化を試みてみたいと考える。

参考文献

石井正彦・杉本武（2014）「第1章　コーパスを用いた日本語研究の特徴－語彙・文法を中心に－」前川喜久雄監修・田野村忠温編『講座日本語コーパス6　コーパスと日本語学』朝倉書店、pp.1-20

小木曽智信（2014）「第5章　形態素解析」前川喜久雄監修・山崎誠編『講座日本語コーパス2書き言葉コーパス―設計と構築―』、朝倉書店、pp.89-115

金庭久美子・金澤裕之（2014）「第1部　「YNU書き言葉コーパス」について」金澤裕之編『日本語教育のためのタスク別書き言葉コーパス』、ひつじ書房、pp.1-50

田野村忠温（2014）「第6章　BCCWJの資料的特性―コーパス理解の重要性―」前川喜久雄監修・田野村忠温編『講座日本語コーパス6　コーパスと日本語学』、朝倉書店、pp.119-152

中俣尚己（2015）「『日中 Skype 会話コーパス』について」
（http：//nakamata.info/about_skype_corpus.pdf、最終確認2016年12月20日）

西尾広美（2013）「幼稚園における『やさしい日本語』使用の必要性―教師と非母語話者の保護者のコミュニケーションの現状調査から―」『日本語研究』33、首都大学東京・都立大学・日本語・日本語教育研究会、pp.99-102

長谷川守寿・西尾広美（2016）「『幼稚園の配布文書コーパス』の作成と試行調査」『言語処理学会 第22回年次大会 発表論文集』、言語処理学会、pp.246-249

三井正孝（2011）「第1章 コーパスデータの作成―OCRソフトを利用して―」荻野綱男・田野村忠温編『講座ITと日本語研究5 コーパスの作成と活用』、明治書院、pp.7-45

宮城信・今田水穂（2015）「『児童・生徒作文コーパス』の設計」『第7回コーパス日本語学ワーク

ショップ予稿集』、国立国語研究所、pp.223-228
森篤嗣（2014)「子どもを持つ外国人のための語彙シラバス」『公開シンポジウム　シラバス作成を科学にする —日本語教育に役立つ多面的な語彙シラバスの作成—』平成27年2月22日配付資料pp.49-60（https：//gobunken.files.wordpress.com/2015/02/20150222_goi.pdf、最終確認2016年12月20日）
李暁燕（2016)「『学校お便りコーパス』について」(http：//lixiaoyan.jp/database/)
McEnery T. & Hardie, A.（2012）. Corpus Linguistics：Method, Theory and Practice, Cambridge University Press.（石川慎一郎訳『概説コーパス言語学－手法・理論・実践－』、ひつじ書房、2014）