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

著者	長谷川 守寿, 西尾 広美
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	377-384
発行年	2017
URL	http://doi.org/10.15084/00001493

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

長谷川 守寿（首都大学東京）†
西尾 広美（国立国語研究所）

Building a 'Corpus of Documents Distributed in Kindergarten' for the Investigation of Vocabularies and Sentence Structures

Hasegawa Morihisa (Tokyo Metropolitan University)
Hiromi Nishio (National Institute for Japanese Language and Linguistics)

要旨

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）が見られるが、日本語学習の機会が少なく日本語が十分に理解できない場合、幼稚園の配布文書が正しく理解されず、情報伝達がうまくいかずに保育活動に支障をきたすこともある。そのため、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」のテキスト化をめざし、『幼稚園の配布文書コーパス』を作成している。

コーパスの作成では、より精度の高い語彙・文型調査が行えるよう、OCR ソフトの認識誤りを人手だけで修正するのではなく、形態素解析システム（unidic-mecab2.1.2）も活用して誤りを発見して修正し、さらに正確に語に区切れない場合は表記の変更・記号の追加を行っている。本発表では、そのコーパス作成法について報告する。

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、幼稚園児の保護者には日本語を母語としない人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、日本語が理解できないケースも出ている。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ず、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を報告し、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)

† hasegawa-morihisa@tmu.ac.jp

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に配布文書で使われているのかというデータ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た『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』の構成と基本方針

３.１　コーパスの構成

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Ｓ幼稚園で、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。

対象とする文書は、Ｓ幼稚園で平成 19 年度（4 月 9 日から翌年 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Ｓ幼稚園』・『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、保護者は幼稚園で配布される全ての文書を理解することが求められるからである。

３.２　コーパスの基本方針

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する。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、“○○”などの記号で置き換えた場合、正しく形態素解析できなくなる可能性があるため、正しく人名・地名と解析されるように“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。

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とおりである。図 1 の流れに沿って作成方法について述べる。

４.１　配布文書の入手

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たということからも、真正性・網羅性の観点で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。

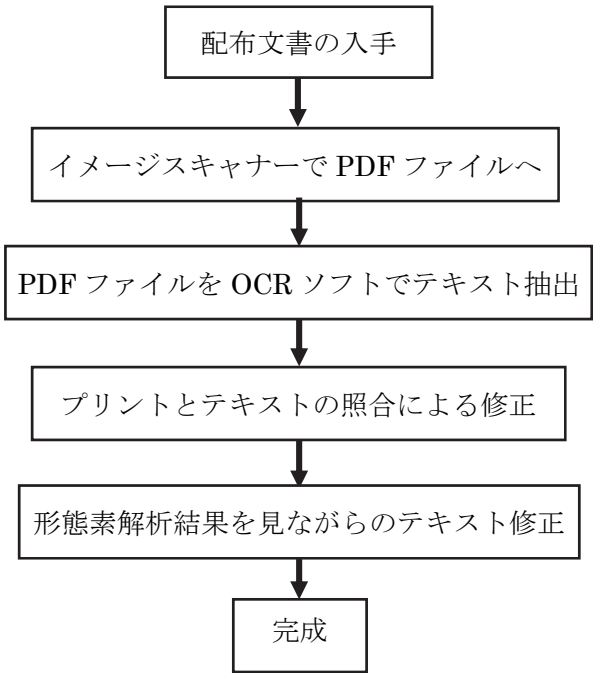

図 1. 配布文書コーパス作成の手順

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。本研究ではフォントの色等は研究の対象外となるので、白黒でスキャンする。

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』）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、イメージスキャナーの設定のままである。文字コードは S-JIS を採用する。手書きされたお知らせは 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。

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、様々な原稿ごとにどのような誤りが見られたか述べられているが、ここでは、幼稚園の配布文書の場合に見られた誤りを挙げておきたい 。誤りは、便宜的に以下のＡからＣのように分けることができる。(1)から(3)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、(1)のように「キッズ」が「キッス」、「たんぽぽ」が「たんぽぼ」と認識されていたので、右側の正しい形に修正した。Ｂは(2)の「よ」と「ょ」のような小書き文字を含む誤りである。Ｃには、(3)の「せ」と「さ」、「一」と「ー」、「間」と「問」のように、三井(2011)などで指摘されているものも含まれるが、「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、確認の作業では注意が必要である。

(1)　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
(2)　でしよう＝＞でしょう、ペインテイング＝＞ペインティング
(3)　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、仮名漢字変換ではよく見られるもの(4)から、前後の文脈を確認しなければ誤りと気づけないような例(5)も見られた。また活用の問題(6)だけでなく、(7)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することとした。

また、対象とする文書は複数の教諭によって書かれていることから表記のゆれが生じる。例えば、「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。

(4)　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
(5)　テレビ付け＝＞テレビ漬け、持ち返す＝＞持ち返らす
(6)　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
(7)　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
(8)　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誤りと考えられないものも含まれている。我々は正しい形の方が 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二度行ったが、OCR ソフトで文字認識を行い作成したテキストから、全ての誤りを目視で発見することは困難である。そこで、形態素解析にかけて、その結果を検証する中で、テキストの入力精度を上げ、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。そのため正確な語彙調査を行うには、入力や表記の修正を行い、正しく解析できるように変更する前処理が必要となる。

そこで形態素解析器に MeCab(mecab-0.996.exe) 、形態素解析用辞書に UniDic-mecab (ver2.1.1)を使用し、入力したテキストの形態素解析を行う。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「ハサミでシッポを切る」のような文を形態素解析すると表 1 のような結果になる（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)。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、正しく語に区切ら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」となるべきであり、正しく語彙素が確定できていない（網掛けは問題のある箇所)。このような方法で誤りを発見し、その箇所を修正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。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、(9)(10)のような例がある。例えば(9)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。また“り”はひらがなで、“リ”はカタカナである。“口”はくにがまえ、“ロ”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、(10)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。

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

この作業は、使用した『読取革命』の設定を変更したり、使用する OCR ソフトを変更することで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、修正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に入力して確認しながら、テキストを修正した。ここでは修正を便宜的に以下の 3 タイプに分け、説明する。

Ｍ１：文字種を変更したもの
　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
　５．漢字表記を、別の漢字表記に変えたもの
Ｍ２：音引きを修正したもの
Ｍ３：語のつながりを修正したもの

Ｍ１：文字種を変更したもの

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、(11)(12)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。

(11)　　おかずはいりません（現状では「入りません」と解析）＝＞おかずは要りません
　　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
　　〇〇だより（「だ／より」）＝＞〇〇便り
　　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
　　〇〇ぐみ（〇〇「グミ」）＝＞〇〇組
(12)　　おわん＝＞お椀

なお、(12)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。(12)はイベントを説明する文書において、持ち物リストの中にあり、文頭に出現しているため、修正した。

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(13)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」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。

(13)　　どろけい＝＞ドロケイ、すだじい＝＞スダジイ、なす＝＞ナス

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」(14)、「Ｍ１－４カタカナ表記だったものを漢字表記に変えたもの」(15)、「Ｍ１－５　漢字表記を別の漢字表記に変えたもの」(16)を挙げることができる。(16)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。

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

この他に、テキストの修正において文字コードの問題が 1 件発生した。上記のように正しく語が認定できるよう修正する作業を行ってきたが、修正できない問題も生じた。それは「鼻をかむ」という表現で、語彙素のレベルでは「鼻」「を」「噛む」と解析される点である。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。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 S-JIS を採用しているため、修正できていない。

Ｍ２．音引きの修正

一部の語の音引きについては誤った解析結果になるので、(17)のような修正を加えた。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そのままである。

(17)　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．出現環境の修正

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）。例えば「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“□ ”ならば、記号として解析され、数える際に外せばよいので、正しく解析でき、語数に影響しないからである。

具体的にどのような修正を加えたか、例を挙げて説明する。(18)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(19)は、語彙素レベルでは「水揚げ／良い／ね」と解析されるが、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(20)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。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ことになる。例えば(21)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。

(18)　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
(19)　水あげようね＝＞水、あげようね
(20)　その後の学びや創造性が＝＞その後の学び□や創造性が

(21)　新年中＝＞新□年中、弁当時＝＞弁当□時、再任用主事＝＞再□任用□主事

UniDic の開発がさらに進んで、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．今後の課題

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(22)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」(23)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(24)の“お家”は(25)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、修正が必要となると考える。

(22)　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。
(23)　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。
(24)　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。
(25)　自信をもって取り組んだ姿におうちでも（略）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ためである。これには、短単位を元に長単位を認定する解析器 Comainu 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と考える。

文　献

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

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 2017 年 2 月 3 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

李暁燕(2016)「『学校お便りコーパス』について」(http://lixiaoyan.jp/database/、最終確認 2017 年 2 月 3 日）