『幼稚園の配布文書コーパス』の作成と試行調査

長谷川守寿（首都大学東京）・西尾広美（国立国語研究所）

１．目的

現在、多くの幼稚園に日本語を母語としない保護者（以下、NNS 保護者）が見られるようになった。NNS 保護者の中には日本語学習の機会がなく、日本語が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、各種調査を行い、将来的に幼稚園の教師と NNS 保護者を結ぶ「幼稚園におけるやさしい日本語」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集の作成を考えている。

本発表では、一部の配布文書を元に語彙調査を行い、『現代日本語書き言葉均衡コーパス』（以後 BCCWJ）と比較することで、文書の特徴を明らかにし、調査をする際にどのような問題点があるのか試行調査を行った結果を報告し、今後の大規模調査へ向けた準備とする。

２．先行研究

小学校での配布物と保育園・幼稚園の連絡帳を対象とした研究として森(2015)が挙げられ、これらに特徴的に現れる語を抽出しているが、現在までのところ幼稚園の配布文書を集めてコーパスを作成したものは見当たらず、新しい試みである。

３．方法

３．１．『幼稚園の配布文書コーパス』

『幼稚園の配布文書コーパス』は、都内外の公立幼稚園の協力を得て、ある 1 年間に配布したお知らせを集め、現在作成中である（今後は都内の私立幼稚園も対象予定）。配付資料の多い幼稚園では、A4 用紙で 228 枚、20 万字程度である。

配布文書とは保護者に配布されたもの全般を指すものとし、配布される頻度が異なるものも同様に扱う。具体的にはほぼ毎月配られるクラス毎の文書、学年毎に配られる文書、園全体に配られる文書、保護者懇談会の資料、入園時の文書等も全て含める。

３．２．対象

本発表は『幼稚園の配布文書コーパス』から、静岡県の市立幼稚園の年少クラスで平成１９年度に配布されたお知らせ「にじ組だより」を対象とする（以後幼稚園サンプル、またはサンプル）。年少クラスを対象としたのは、年中や年長クラスと比べた場合、読み手である保護者にとって初めてのことが多いため、内容がより詳細に説明されているのではないかと考えたからである。また、どこの幼稚園でも、学年あるいはクラス毎のお知らせは作成されることから、一般的な文書だと思われ、特徴を探るには適当と判断した。

３．３．手順

解析エンジン MeCab（0.996）と解析辞書 UniDic（2.1.2）を使用して、形態素解析を行い、結果を ChaKi.NET（3.01.513）で分析する。

まず、形態素解析の結果を確認しながら、本文の修正、表記の統一等を行い、語彙調査にふさわしい文に整えていく。

日本語の誤りとして(1)の下線部には、矢印右の二重下線を加えた部分のような修正を加えた。

(1) いただきける => いただける

また、形態素解析辞書 UniDic には音引きの形を書字形に持つ語もあるが（例「だーい好き」）、持たない語もあるため、配布文書を修正した。これは、出現形の調査のためではなく、語を正しく数えることが目的であるからであり、(2)のような部分を矢印右のように変更した。

(2) 待ってまーす　＝＞　待ってます。

また、同様に表記が原因で誤解析を生じさせる(3)のような表現は修正した。

(3) しっぽとり　＝＞　しっぽ取り

ただし辞書に含まれない語(4)は正しく解析することができないので、手作業で品詞・語種を修正し、未知語には全て何らかの品詞を付与した。

(4) バイキンのおやびん

また、品詞・語種の判定に誤りがある場合のみ修正した。(5)の園は「ソノ・和語」と解析しているが、「エン・漢語」に修正した。なお、語の区切り、読みの修正はしていない。

(5) 園でお預かりします。

以上のように修正した文書の品詞構成、動詞・名詞・形容詞の上位語、語種の構成比、N-gramの比較を行う。比較対象は、BCCWJ の語彙調査「『現代日本語書き言葉均衡コーパス』語彙表」内の短単位語彙表データ・品詞構成表等である。

４．結果

４．１．品詞構成

表 1. サンプルデータと BCCWJ の品詞構成

	サンプル		BCCWJ	
品詞	頻度	比率	頻度	比率
助詞	2438	31.0	31,428,580	30.0
名詞	2163	27.5	36,651,588	35.0
動詞	1431	18.2	14,148,216	13.5
助動詞	854	10.9	10,279,970	9.8
接頭詞	244	3.1	868,076	0.8
接尾辞	229	2.9	3,346,976	3.2
形容詞	146	1.9	1,588,226	1.5
形状詞	130	1.7	1,314,004	1.3
副詞	118	1.5	1,830,329	1.7
代名詞	38	0.5	1,516,372	1.4
連体詞	30	0.4	997,276	1.0
感動詞	29	0.4	161,716	0.2
接続詞	20	0.3	481,094	0.5
合計	7870	100.0	104,612,423	100.0

語数は記号・空白・補助記号を除いて集計した。表 1 は、幼稚園サンプルでの頻度の多い順に BCCWJ 全体の品詞構成表(BCCWJ_token_suw)とあわせて、まとめたものである(異なり語数の比較は行わない)。

BCCWJ に比べると、名詞が少なく動詞が多いという傾向が見られる。これは、配布文書が園児の園内での様子を報告することが多いためであると推測され、それゆえか形容詞・形状詞の割合も高い。また接頭辞の割合も高く、244 語中 233 語が「お家」や「ご家族」などにつく「御」であり、文書の丁寧さという特徴が見てとれる。

表 2. BCCWJ 内のサブコーパスの品詞比率

ジャンル	OB	OC	OM
名詞	28.1	28.0	29.0
動詞	15.2	14.4	15.7
形容詞	1.8	2.3	0.9
形状詞	1.3	1.3	1.5
接頭辞	0.6	0.8	1.2

表 2 に名詞の比率が 20 %台である BCCWJ 内のサブコーパス（全て固定長）のいくつかの品詞の比率を示す。傾向として OB(ベストセラー)、OC（Yahoo!知恵袋）、OM（国会会議録）に近いと思われるが、動詞、形容詞・形状詞、接頭辞が多いサブコーパスは見当たらない。

４．２．名詞

名詞（語彙素）とその出現頻度を観察する。名詞は「名詞-数詞」の頻度が多いため、「名詞-普通名詞」に限定して比較する。サンプルデータと BCCWJ における上位 11 語までの「名詞-普通名詞」をまとめたのが表 3 である（10 位は 2 つ。網掛け部分は共通する部分。以下同)。

「事」など共通している語もあるが、やはり「子供」「子」「一緒」など、特徴的な語彙が見られた。本調査は、コーパスの一部を使用した限定的・試行的なものであるが、名詞に関しては共通する語彙が少ないということがいえる。なお、「家」は「お家（うち)」の形での出現だけである。

表 3. 名詞とその出現頻度（上位 11 語）

幼稚園サンプル		BCCWJ	
名詞	頻度	名詞	頻度
事	57	事	740,949
子供	50	物	276,332
子	45	年	246,629
一緒	32	時	162,023
会	31	人	156,890
日	27	為	152,779
組	25	月	152,090
母	24	中	116,641
弁当	24	自分	111,079
家	23	所	108,040
皆	23	方	96,578

４．３．動詞

動詞には「動詞-一般」と「動詞-非自立可能」があるが、両方を合わせて検討する。サンプルデータと BCCWJ における上位 10 語までの動詞をまとめたのが、表 4 である。

表 4. 動詞とその出現頻度（上位 10 語）

サンプルデータ		BCCWJ	
動詞	頻度	動詞	頻度
為る	186	為る	2,563,860
居る	95	居る	1,121,183
成る	80	有る	956,900
下さる	54	言う	803,148
来る	46	成る	564,043
出来る	39	来る	238,115
見る	38	思う	229,081
行く	36	見る	223,222
遊ぶ	36	行く	220,816
頂く	34	出来る	209,481

名詞に比べ共通する語が多い（10 語中 7 語）が、「下さる」(6)「遊ぶ」(7)「頂く」(8)等、敬意表現や園児の様子を描写する表現も見られた。

(6) 月曜日に上靴袋に入れて持たせてください。
(7) 色々変身させて遊んでいます。
(8) 持ってきていただきたいと思います。

４．４．形容詞

ここでは、形容詞に形状詞（いわゆるナ形容詞）を含めて考察する。幼稚園サンプルと BCCWJ における上位 10 語の形容詞をまとめたのが表 5 である（10 位が 2 つ）。

表 5. 形容詞とその出現頻度

サンプルデータ		BCCWJ	
形容詞	頻度	形容詞	頻度
様	49	無い	458,395
良い	28	様	373,477
楽しい	25	良い	198,994
幼稚	16	多い	55,818
嬉しい	13	高い	40,495
上手	12	そう	38,826
大きい	12	大きい	37,111
無い	11	可能	36,362
寒い	9	強い	28,368
様々	6	悪い	27,301
美味しい	6	少ない	25,175

「幼稚」が 16 回出現するのは、茶まめでは「幼稚園」を「幼稚（形状詞）／園（接尾辞）」と解析するためである。動詞ほど共通する語を持つわけでもないが（11 語中 4 語）、名詞よりは多い。「楽しい」「嬉しい」「美味しい」などは園児の様子やお弁当の記述の際に使われていた。

４．５．語種

幼稚園サンプルと BCCWJ の語種構成をサンプルでの頻度順に集計したのが表 6 である。

和語の使用比率は BCCWJ と変わらないが、漢語・固有名詞が少なく、記号が多いという特徴が見られる。“！”や引用を示す“「”“」”、さらに“♪”(9)“☆”(10)“(^・^)”(11)等も多用されているが、記号の多さも幼稚園の配布文書の特徴と考えられる。

(9) お話いただけると嬉しいです♪♪
(10) 「大きくな～れ☆大きくな～れ☆」
(11) プレゼントだったようです(^・^)

表 6. 幼稚園サンプルと BCCWJ の語種構成

	幼稚園サンプル		BCCWJ	
語種	頻度	比率	頻度	比率
和語	6,447	67.9	71,518,076	68.4
記号	1,624	17.1	256,511	0.2
漢語	1,129	11.9	26,106,082	25.0
外来語	204	2.1	2,945,153	2.8
混種語	68	0.7	1,125,516	1.1
固有名詞	21	0.2	2,661,023	2.5
その他	0	0	62	0.0
合計	9,493	100.0	104,612,423	100.0

漢語の少なさについては今後「やさしい日本語」とも関連するため、さらなる調査が必要である。

４．６．N-gram の抽出

UniDic-MeCab を用いて区切られた形態素単位の N-gram を、ChaKi.NET を用いて抽出した結果 5-gram（12）が得られた。

（12） お願いします。（13 回）・ありがとうございました。（10 回）・いきたいと思います（9 回）・てきました。（9 回）・ていました。（7 回）・ておいてください。（7 回）

この結果を BCCWJ の N-gram を明らかにした「言語モデル配布ページ」と比較する。これは BCCWJ のコアデータを短単位に区切り、さらに短い単位に変更しているため、単純に比較できないが、他の組み合わせがなく語として確かなもののみ抜き出し、得られた結果は 7-gram（13）、5-gram（14）、4-gram（15）である。

（13）と回答した人の割合（36 回）

（14）ているのですが（32 回）・しなければならな（32 回）・ではありません（58 回）

（15）されている（345 回）・ については（320 回）・においては（221 回）・しています（221 回）

BCCWJ の文末表現には、前置き・逆接（～ですが）、義務（～なければならない）・断定（～ではありません）・機能表現「～について・において」等が多く見られるが、幼稚園の配布文書には園側の意思表明「いきたいと思います」・依頼「ておいてください」、それが達成されたことに対する感謝を表す表現「ありがとうございました」が抽出され、大きく異なる。

５．考察

幼稚園の配布文書を BCCWJ と比較すると、品詞構成、語種構成とも異なり、品詞毎に多用される語にも特徴が見られた。また N-gram を比べると、BCCWJ と異なる表現が多用されている。

なお、BCCWJ と比較するため UniDic を用いたが、語彙表を作成するには表 3 のように細かすぎるため、代替の方法も考える必要がある。

本発表は、静岡県の公立幼稚園を対象としたが、今後は更にデータを増やし、都内の公立幼稚園や私立幼稚園も対象に配布文書の特徴を探りたい。地域性や運営母体の違いにより、文書の内容や使用語彙が違ってくる可能性が考えられるからである。多様性のある文書の内容を取り入れることで、幼稚園の教師と NNS を含む保護者に役立つ教材作成に活かしていきたいと考える。

参照データ

『現代日本語書き言葉均衡コーパス』語彙表ver1. 0（http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html）
言語モデル配布ページ（2015/11/04 取得）
（ http://plata.ar.media.kyoto-u.ac.jp/gologo/lm.html）

参考文献

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、pp.99-102、首都大学東京・都立大学・日本語・日本語教育研究会

森篤嗣（2015）「子どもをも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する－日本語教育に役立つ多面的な語彙シラバスの作成－』、pp.49-60、データに基づいた日本語教育のための語彙・文法研究会