東京都立大学都市教養学部人文・社会系
東京都立大学人文科学研究科
「人文学報」第519-7号
（日本語教育学）
2023年３月抜刷

複数の異なる文書群からなる対象データの特徴語抽出に関する考察

―幼稚園の配布文書を例に―

長谷川　守寿

複数の異なる文書群からなる対象データの特徴語抽出に関する考察
—幼稚園の配布文書を例に—

長谷川　守寿

1．目的

本研究は、対象データが複数の異なる文章群からなる時、そこから特徴語を抽出する方法について考察するものである。

対象データと参照データから特徴語を抽出する際の指標については寺嶋（2009）の考察があり、そこでは対数尤度比はコーパスサイズの影響を受けにくいとされ、この指標の有用性が述べられている。しかし、後述するように同じカテゴリーには入るが、含まれる内容も異なる、複数のデータしか得られない場合、対数尤度比を用いるにはどのような対応が必要となるのか考察している研究は管見の限り存在しない。

本研究で対象とする文書群とは、幼稚園の配布文書である。幼稚園の配布文書には個人名などが多いためプライバシーに対する配慮が必要とされ、結果として提供元を確保することが難しく、提供していただけるのは限定された量のデータとなってしまい、大量のデータを入手することが困難である。そこで、複数の園から提供していただいているのが現状である。書き手による語彙の違いや特徴語の抽出の際には一つの幼稚園に限定してきたが、今後『幼稚園配布文書コーパス』に含めるデータを増やしていく際に、単純に合併していいのかという疑問に答えることを目的とし、複数の異なる文書群を含む対象データと一つの参照データから、偏りなく特徴語を抽出する方法を探索的に明らかにする。

2．先行研究

髙見（2003）は「複数のコーパスの中で特定のグループに特徴的に多く用いられている語を、統計的手法を用いて一律に特定する方法論」（p.73）を示したもので、イギリスの新聞コーパスから高級紙で多用される語と、大衆紙で多用される語の例を示し、対数尤度比を用いた方法論の妥当性を示している。

同様の手法を日本語教育用の語彙リスト作成に用いた研究として、森（2016）が挙げられる。森（2016）は「学校お便りコーパス」と「連絡帳コーパス」をそれぞれ対象データとし、参照データを『現代日本語書き言葉均衡コーパス』（以後、BCCWJと呼ぶ）に含まれる図書館サブコーパス内の固定長データ（以下、LB_fixedと呼ぶ）として、それぞれのデータの特徴語を抽出し、子どもを持つ外国人を対象とした日本語教育用の語彙シラバスを考察している。

嶋（2016）は、看護師の申し送り場面を文字化したデータを対象データ、LB_fixedを参照データとし、対数尤度比を求め、これを特徴度として特徴度の高い語彙を抽出し、看護師国家試験の語彙との比較から、外国人看護師候補者にとって習得の必要性のある語彙を考察した。

松田（2016）は、理工系専門基礎科目コーパスを構築し、そのコーパスを対象データ、LB_fixedを参照データとし、文字1gram、文字2gram、形態素解析された語から対数尤度比を求め、特徴的な１文字、２文字、語を抽出し、それらを元に理工系留学生のための文字・語彙シラバス案を提示している。

石黒（2016）は、日本語教育分野の修士論文42本を対象データ、LB_fixedを参照データとして対数尤度比を求め、これを特徴度として特徴語を抽出し、複合語による分析結果などともあわせ、日本語教育専攻の大学院留学生のための語彙シラバスの提示語例を示している。

このように、対象データに等質と考えられる、まとまったデータが選ばれ、参照データとの比較から対数尤度比を求め、そこから特徴語を検討している研究は多くあるが、等質と言えないデータしか入手できない場合を考察した研究は管見の限りない。そこで、等質と言えないデータの場合、どのような対応が必要となるのか探索的に検討を加える。

３．方法

３．１．対象

本研究の対象は、長谷川（2022）で対象とした幼稚園と長谷川・西尾（2019）で対象とした幼稚園の配布文書を文字化したデータである（以後それぞれA、Bと呼ぶ）。Aは年少・年中・年長の３学年を対象とした公立の幼稚園のデータで、自然豊かな環境で友達と様々な遊びをし、触れ合うことに重きを置き、Bは年少・年長の２学年を対象とした公立の幼稚園で、歩くことに重きを置いている印象を受ける。

筆者らの研究グループは、複数の幼稚園から提供していただいた文書を保有しているが、園によって提供していただける資料が異なる。例えば、Aには毎月園全体に配布されるお知らせから、それぞれのクラスで配られるもの、さらには保護者会資料も含まれる。それに対してBから提供していただいた文書は、毎月園全体に配布されるお知らせと入園のしおりである。このように配布元による違いが見られ、また園長先生が書いた文書、クラスの担当教諭が書いた文書、PTA役員が書いた文書など書き手が様々な文書も含まれるため、それらを文書群と呼ぶこととする。

３．２．前処理

本研究では、語の特定には形態素解析[(1)]の結果判明する、その語の小分類までの品詞[(2)]、語彙素、語彙素読みを用いることとする。これは、品詞・語彙素が同じで、語彙素読みが違う語が出現してくるためであり、三つの情報を使用することにより、正確な語の抽出と語数の集計が可能となる。例えば「表（ひょう）」と「表（おもて）」の場合、品詞はともに「名詞-普通名詞-一般」で、語彙素も「表」であるが、語彙素読みが「ヒョウ」と「オモテ」で異なる（以後、（ ）内は語彙素読み）。そのため（１）（２）のような例から正確に語数を集計するには、三つの情報を

使用することが必要となる。

（1） 名前は表の記名場所には書かずに、防災頭巾の内側に （A）
（2） 近いうちに係を募集するために氏名記入の表を貼り出しますので、 （B）

調査手順として、まずBのデータに対して長谷川（2022）と同様の方法でクリーニングを行う。表記を変更することで形態素解析の結果を修正できるものは表記を修正し、表記の変更で形態素解析の結果を修正できないものは、形態素解析の結果を直接人手で修正する。

表記の修正で形態素解析の結果を修正できたものに関しては、長谷川（2022）と同様の語が多く、直接人手で修正した語には、「門」を含む語が多数挙げられる。例えば、「入り口・正門」は「正門（マサカド）」と解析されてしまうので、「入り口・正門（セイモン）」のように語彙素読みを修正したものもあれば、「動物・門前（モンゼン）」を「動物・門（モン）・前（マエ）」、「動物・門脇（カドワキ）」を「動物・門（モン）・脇（ワキ）」等のように、語の区切りを修正し、それに伴い、適宜語の追加を行ったものもある。他に区切り方を修正した語には、「泥団・子作り」を「泥・団子・作り」にするなど、主に遊びに関する語に多く見られた。

またBのデータ内で気づいた誤字には、例えば「絡麗」（「綺麗」の誤り）や「川添い」（「川沿い」の誤り）のようなものがあったので、適宜修正した。その後、長谷川（2022）で対象としたAのデータでも見落としがないか確認し、必要に応じて修正を加えた。

その結果対象データとして、Aは延べ102,539語、Bは延べ45,832語と判明した。参照データとなるLB_Fixedは、6,685,183語[(3)]である。

3．3．「異なる」文書群と判断する理由

ここでタイトルにあげた「異なる」文書群と判断した理由を述べる。

まず、名詞を例に特徴語を比べる。A・Bそれぞれを対象データ、LB_fixedを参照データとした際、形態素解析の結果「名詞-普通名詞-一般」と品詞付けされた語に対し、特徴度を求める。特徴度には対数尤度比を補正[(4)]した数値を用いることとする。表1は、A・Bそれぞれの特徴語を特徴度の高い順に上位10語までを挙げたものである。「会」「園」が語として抽出されるのは、LB_fixedの語彙表と比較するため、同じ基準である短単位で語に区切っていることによる（「降」が語として抽出されるのは、「降園」という語が形態素解析用辞書UniDicに存在しないためである）。表1を見ると、同じ幼稚園という教育機関であるため、「学級」や「園庭」のように共通する語もあるが、Aの「遊び」「友達」や、Bの「年少」「親子」など異なる語も見られ、これが同じように扱っていいのか、判断に迷う所以である。

表１．名詞の特徴語上位10語（網掛けは共通する語。括弧内は特徴度）

幼稚園	A		B	
順位	名詞	頻度	名詞	頻度
1	友達	336（1628.55）	年少	220（2009.18）
2	学級	208（1526.63）	会	325（1263.65）
3	遊び	267（1512.83）	親子	143（ 958.91）
4	会	486（1501.70）	園庭	85（ 822.24）
5	園	199（1390.49）	降	77（ 736.70）
6	園庭	136（1111.71）	園	92（ 715.14）
7	学期	140（1016.60）	登園	71（ 708.62）
8	年中	104（ 872.17）	遠足	76（ 663.71）
9	弁当	152（ 837.07）	学級	79（ 637.07）
10	子供	405（ 795.98）	広場	98（ 628.52）

そこで、「動詞−一般」と品詞付けされた語を対象に表１と同様の処理をした。その結果が表２である。

表２．動詞の特徴語上位10語（網掛けは共通する語。括弧内は特徴度）

幼稚園	A		B	
順位	動詞	頻度	動詞	頻度
1	遊ぶ	370（1870.69）	言う	45（505.33）
2	楽しむ	230（ 978.90）	遊ぶ	87（365.12）
3	言う	266（ 594.32）	つく	155（228.52）
4	作る	251（ 311.72）	楽しむ	59（202.90）
5	取り組む	70（ 268.05）	引き落とす	21（191.52）
6	過ごす	86（ 258.99）	取り組む	32（128.00）
7	食べる	141（ 224.37）	育てる	36（120.83）
8	味わう	61（ 214.13）	過ごす	36（107.10）
9	触れ合う	36（ 196.71）	歩く	56（ 92.27）
10	感ずる	126（1 73.91）	引き取る	19（ 82.37）

表２を見て分かる通り、「遊ぶ」「楽しむ」「取り組む」「過ごす」など共通するものもあるが、その一方Ａでは「作る」「食べる」「味わう」が特徴語として抽出されているのに対し（主に「お弁当」に関連する文書に存在する）、Ｂは「引き落とす」「引き取る」のように費用の引き落としや引き取り訓練など、事務的な連絡に使用される語が抽出されている。以上のことから、二つのデータを同じ幼稚園の配布文書というジャンルには含まれるが、「異なる」文書群と考える。

3．4．手順

以後、形態素解析の結果「名詞－普通名詞－一般」と品詞付けされる語に限定して、複数の異なる文書群から特徴語を抽出する方法を考察する。

まず、3．3節で行ったようにAとBからそれぞれ特徴語を抽出し、それらを「特徴語リストA」、「特徴語リストB」と呼ぶこととする。次にAとBを合併（以後、データを合わせることを「合併」と呼ぶこととする）し、A＋Bを作成し（以後、「＋」は合併したデータを表すこととする）、そこから「特徴語リストA＋B」を抽出する。「特徴語リストA」、「特徴語リストB」、「特徴語リストA＋B」を、3．4．3節で述べる評価方法で検討し、どちらかに偏った結果になった場合は、別の手法を採る。

3．4．1．頻度が0である語の扱いについて

ここでは、対象データには含まれるが、参照データには含まれない語の扱いについて述べる。

本研究で使用した形態素解析用辞書の語彙項目としては存在するため、対象データでは語として抽出されるが、LB_fixedではデータに出現していない語（頻度0）がある。例えば、幼稚園での遊びやそれに使われるものを表す「七並べ」「色紙（いろがみ）」は、対象データには1語ずつ出現し、LB_fixedには1語も出現しない語である（仮にb＝0とする）。

このような場合、例えば嶋（2016）は対象データで5回以上出現している語を特徴的な語（この場合は名詞）とし、石黒（2016）では対数尤度比を計算できないとしている。しかし本研究では、対象データの頻度に関係なく、高見（2003）同様に、対数尤度比を求める計算式の中でblnb＝0として処理することとする（lnは自然対数を求める関数で、この部分だけ0として扱う）。よって、特徴語に含まれるか否かは、対数尤度比によることとなる。

3．4．2．特徴度について

3．3で述べた通り、特徴度には対数尤度比を補正した数値を用いるが、特徴度の有意水準とその臨界値は、高見（2003：89）によれば、5％で3.84、1％で6.63、0.5％で7.88、0.1％で10.83である。有意水準として1％を採用している研究には、田中・近藤（2011）や森（2016）などが挙げられる。

本研究では、手法の有効性を確認するために、考察対象となる特徴語の数を制限する。そこで高見（2003）・嶋（2016）・松田（2016）同様に0.1％を採用し、対数尤度比が10.83より大きい語を特徴語、つまり対象データに有意に偏って出現する語とする。

3．4．3．抽出結果の評価方法

本研究では「特徴語リストA」「特徴語リストB」「特徴語リストA＋B」の三者を比較する上で、ある範囲内における語の変化の検討では、合併したことの特徴を捉えることが出来ないと考える。例えば、20位までの特徴語を比較した場合、合併することによって順位が上昇する語、順

位する語、さらには20位以下になる語などあり、ある範囲に含まれる特徴語を考察しても、合併による変化は捉えられない。

そこで、複数のデータが存在し、それらのデータから特徴語を抽出する際、特徴語リストに含まれる語が、ある特定のデータに偏っているのは望ましくなく、どのデータからも同じように特徴語リストに採用されることが望ましいと考える。具体的にはAから多く採用され、Bからはあまり採用されない（逆も同様）というのは望ましくないと考え、どちらからも偏りなく採用されるのが望ましいと考える。そこで、採用の状況を「特徴語リストA」「特徴語リストB」「特徴語リストA＋B」の語数に基づいて評価することとする。

4．結果

4．1．そのまま合併した場合

AとBをそのまま合併したデータA＋Bと、参照データの比較から抽出された「特徴語リストA＋B」の内訳を調べたものが表3である。「特徴語リストA＋B」において、「特徴語リストA」にも「特徴語リストB」にも含まれる場合は「AとB共通」とし、「特徴語リストA（またはB）」だけに含まれる場合は「A（またはB）から採用」されたとした。「AとBの合併により」とは、ABの合併により特徴語リストに含まれるようになった語のことで、例えば幼稚園児の遊びの一種である「双六（すごろく）」がある。これは、Aでは3回（特徴度4.08）、Bでも3回（特徴度8.11）であったが、合併することにより6回（特徴度10.94）となったような語を指す。これらは「特徴語リストA」にも「特徴語リストB」にもないので、A・Bから採用された語とはしないこととする。

表3．「特徴語リストA＋B」の内訳（単位：語）

内訳	AとB共通	Aから採用	Bから採用	AとBの合併により	合計
	182	313	107	25	627

このように、「AとBの合併により」特徴語リストに採用された語もあるため、「特徴語リストA＋B」から採用の状況を観察するのではなく、「特徴語リストA」「特徴語リストB」の観点から「特徴語リストA＋B」に何語採用され、何語不採用だったかを観察することとする。

そこで「特徴語リストA」「特徴語リストB」から「特徴語リストA＋B」に採用された語数を示したのが表4である。「採用された語」は、表3の「AとB共通」と「A（またはB）から採用」の合計である。

表4より、「特徴語リストA」からは527語中495語（共通182＋採用313）、93%（495／527）の語が採用され、「特徴語リストB」からは331語中289語（共通182＋採用107）、87%（289／331）の語が採用されたことになる。この結果をχ^2検定にかけると、「特徴語リストA＋B」に採用されたか否かに関して、採用の程度に「特徴語リストA」と「特徴語リストB」の間に有意な偏りがあった（$\chi^2(1)=10.470, p<.01, \phi=.115$）。さらに残差分析を行うと、「特徴語リストA」か

ら「採用された語」と「特徴語リストB」から「採用されなかった語」が有意に多く、「特徴語リストB」から「採用された語」と「特徴語リストA」から「採用されなかった語」が有意に少ない結果（$p < .05$）となった。

このように、そのまま合併したのではA・B間で採用に偏りが生ずる結果となる。

表4.「特徴語リストA＋B」に採用された語とされなかった語の内訳（単位：語）

	採用された語	採用されなかった語	合計
A	495	32	527
B	289	42	331

4.2. 延べ語数を元に調整した場合

上記のように、延べ語数の異なるAとBをそのまま合併した場合、「特徴語リストA＋B」への採用にはA・B間で偏りのある結果となった。そこで、延べ語数が異なることが問題の一つと考え、延べ語数を元に調整し、延べ語数の少ないBを対象に、延べ語数BをA／B倍し、かつそれぞれの語の頻度もA／B倍することとする。つまりBの延べ語数をAと同じ値にし、他の語の頻度も同様に調整することとする。このように頻度に調整を加えたBのことを以後「調整B」と呼び、参照データとの比較から「特徴語リスト調整B」を抽出する。その後、Aと調整Bを合併したものを「A＋調整B」とし、参照データとの比較から「特徴語リストA＋調整B」を抽出する。

まず、「特徴語リストA＋調整B」の内訳を示したのが表5である。

表5.「特徴語リストA＋調整B」の内訳（単位：語）

内訳	Aと調整B共通	Aから採用	調整Bから採用	Aと調整Bの合併により	合計
	228	231	236	26	623

表3と比べた場合、全体の語数は627語から623語と大きな変動は見られなかった。ただし、表3では、延べ語数の多いAからの採用が延べ語数の少ないBからの採用よりも多かったのに対し、調整を加え延べ語数は同じになっているAと調整Bでは、表5からわかるように「調整Bから採用」が「Aからの採用」よりも多くなっている。この点については、今後の課題で触れる。

「特徴語リストA」と「特徴語リスト調整B」から、「特徴語リストA＋調整B」に採用された語の内訳を示したのが、表6である。「特徴語リスト調整B」（510語）は、Bの語数を調整し、再度特徴度を求め抽出したものであるが、「特徴語リストA」（526語）とあまり変わらない結果となった。ただし、「特徴語リストA」では、採用されなかった語の数が増えた（32語から67語）のに対し、「特徴語リスト調整B」からは採用された語が増える（289語から464語）結果となった。

4.1と同様に、この結果をχ^2検定にかけると、「特徴語リストA」「特徴語リスト調整B」から「特徴語リストA＋調整B」への採用の程度に関して、Aか調整Bかで有意な関連が見られな

かった（$\chi^2(1)=3.311, p>.05, \phi=.060$）。

このように延べ語数を元に調整することで、どちらかに偏ることなく、特徴語リストに採用されることになる。

表6.「特徴語リストA＋調整B」に採用された語とされなかった語の内訳（単位：語）

	採用された語	採用されなかった語	合計
A	459	67	526
調整B	464	46	510

5. まとめ

筆者らのグループが作成している『幼稚園配布文書コーパス』を拡大していく際に、異なる文書群のものをそのまま合併にしてよいのかという疑問に対して、「名詞-普通名詞-一般」に限定した調査であるが、そのまま合併した場合には、語数の多いデータに偏って抽出されることとなり、少ない文書群の語数を調整した場合、偏りのない特徴語リストを抽出することができ、そのまま合併するのではなく、必要な対応をして扱う必要性を明らかにした。

もちろん用例検索等の対象とする場合は、そのまま合併することに問題はないが、少なくとも上記のように、語の頻度を調べ対数尤度比を求めるような量的調査を行う場合には、異なる文書群のデータとして別々に扱う必要がある。

この結果から、今後『幼稚園配布文書コーパス』を拡大していく際には、BCCWJのジャンルに相当するような扱いが必要なのではないかと考える。BCCWJは出版（生産実態）サブコーパス・図書館（流通実態）サブコーパス・特定目的サブコーパスからなり、さらにその下にいくつかのレジスターから構成される（山崎 2006）。規模は全く小さいが『幼稚園配布文書コーパス』にも構造化が必要になり、目的に応じて使い分けていくことが必要となると思われる。

6. 今後の課題

今後に残された課題は多い。本研究においては様々な変数を先行研究に倣ってきた。例えば、参照データをLB_fixedとして特徴語を抽出してきたが、参照データを別のデータ、例えばBCCWJの生産実態を反映したデータに変更した場合、抽出される特徴語はどのように変わるのか考察することは、方法論の妥当性を示すためにも必要となろう。石川（2022）は五種類の参照データを用いて、対象データとの比較を行い、結果がどのように変化するか観察を行ったものであるが、対象データの特徴を明確にするためにも、さらにはLB_fixedを使用することの妥当性を示すためにも、参照データの変更や複数化は検討する必要がある。

次に対象とした品詞である。「名詞-普通名詞-一般」については、今後この方法で調整し、抽出していくのが望ましいと考えられるが、中分類までの「名詞-普通名詞」の場合や、動詞のような他の品詞の場合も同様の結果が得られるか、調査が必要になる。

この他にも特徴度の有意水準として0.1％を採用したが、１％を採用した場合や５％を採用した場合、同様の結果が得られるかは不明であり、これらも今後の課題の一つである。

次に本研究で用いた調整についてである。４．２で触れたように、調整した場合、調整を加えたデータから採用された語が多い結果となり、これは直感に反するかと思われる。どのような語が採用されるのか詳細な調査が必要となるが、一因としてそれぞれの頻度に対し一律に調整しており、語数と頻度の分布としては、頻度１の語がない不自然な形となっていることが考えられる。そこで頻度の多寡に応じた調整を行い、全体の頻度を延べ語数の多いデータに合わせる方法も検討したい。

筆者らの研究グループでは、幼稚園配布文書を収集する際には、なるべく多様な幼稚園の配布文書を集めるという意図の元、依頼を行ってきた経緯がある。３．３で述べたようにＡ・Ｂともに特徴があるデータのため、今回のように語数を調整する方法で、どちらにも偏らないデータの採用が可能になった可能性も考えられる。本稿で用いた手法が、共通点が多いと考えられるデータ、例えば同じ地域にある３年制の公立幼稚園のようなデータの場合にも有効に働くのか、慎重な調査が求められる。

外国につながる保護者のため、より有用性の高い語彙表や文型集を作成することは、筆者らの研究グループの目的の一つである。語彙表作成時には、実際に海外につながる保護者に対する理解度等の調査が必須となるが、その際の調査対象とする語を厳選するためにも、『幼稚園配布文書コーパス』をさらに拡大・充実させていくことが重要であり、今後は少量のデータも含めていく必要があると考える。実際に、幼稚園単位ではなく、あるクラスを担当された教諭の方から、その方が書いた文書だけを提供していただいたケースもある。これはＢよりさらに少なく、Ａに対して３分の１程度のデータであるが、提供していただいた貴重なデータを有効に活用するためにも延べ語数を調整する方法が有効なのかを確認しながら、特徴語を抽出していく手法を今後も検討していきたい。

注

（１）形態素解析には、形態素解析用辞書にUniDic、形態素解析用エンジンにMeCabを用いた。

（２）形態素解析の結果、語に付与される品詞には、その品詞により、大分類、中分類、小分類された結果が付与される。例えば、「表（ひょう）」には、大分類「名詞」、中分類「普通名詞」、小分類「一般」が付与される。また「遊ぶ」には、大分類「動詞」、中分類「一般」が付与される。それらを「-」で結合した形で、その品詞を表すこととする。なお、動詞のように小分類がない品詞もある。

（３）「『現代日本語書き言葉均衡コーパス』語彙表ver.1.0 解説」より。2023/01/12確認。URLは以下。https://ccd.ninjal.ac.jp/bccwj/data-files/frequency-list/BCCWJ_frequencylist_manual_ver1_0b.pdf

（４）（対象データでの当該語の頻度a）×（参照データの延べ語数－b）が、（参照データでの当該

語の頻度b)×(対象データの延べ語数－a)よりも小さい場合、－1を乗じる補正を行う。

参考文献

石川慎一郎（2022）「L2日本語学習者を対象とした中間言語対照分析における参照基準の拡張「多言語母語の日本語学習者横断コーパス」（I-JAS）と「小中高大生による日本語絵描写ストーリーライティングコーパス」（JASWRIC）の連動分析の試み」『計量国語学会第六十六回大会予稿集』計量国語学会、pp.42-47

石黒圭（2016）「第8章　日本語教育専攻大学院留学生のための語彙シラバス」森篤嗣編『現場に役立つ日本語教育研究2　ニーズを踏まえた語彙シラバス』、くろしお出版、pp.159-178

嶋ちはる（2016）「第11章　外国人看護師のための語彙シラバス」森篤嗣編『現場に役立つ日本語教育研究2　ニーズを踏まえた語彙シラバス』、くろしお出版、pp.213-229

高見敏子（2003）「「高級紙語」と「大衆紙語」のcorpus-drivenな特定法」『北海道大学大学院国際広報メディア研究科言語文化部紀要』44、pp.73-105

田中牧郎・近藤明日子（2011）「第1章第2節　教科書コーパス語彙表」『言語政策に役立つ、コーパスを用いた語彙表・漢字表等の作成と活用』、特定領域研究「日本語コーパス」言語政策班、pp.55-64

寺嶋弘道（2009）「日本語教育語彙を選定するための統計的指標―尤度比検定、カイ2乗検定、イェーツの補正公式の特徴―」『ポリグロシア』17、pp.71-83

長谷川守寿（2022）「幼稚園の配布文書における特徴語について」、『人文学報』518-7、東京都立大学人文科学研究科、pp.37-52

長谷川守寿・西尾広美（2019）「『幼稚園配布文書コーパス』の構築と文書作成者による語彙の違いについて」、『人文学報』515-7、首都大学東京人文科学研究科、pp.51-66

松田真希子（2016）「第7章　理工系留学生のための文字・語彙シラバス」森篤嗣編『現場に役立つ日本語教育研究2　ニーズを踏まえた語彙シラバス』、くろしお出版、pp.139-158

森篤嗣（2016）「第9章　子どもを持つ外国人のための語彙シラバス」森篤嗣編『現場に役立つ日本語教育研究2　ニーズを踏まえた語彙シラバス』、くろしお出版、pp.179-195

山崎誠（2006）「『現代日本語話し言葉均衡コーパス』の基本設計について」『特定領域研究「日本語コーパス」平成18年度公開ワークショップ（研究成果報告会）予稿集』、特定領域研究「日本語コーパス」総括班、pp.127-136

付記

本稿は、日本学術振興会科学研究費補助金（基盤研究C、「保護者としての日本語」の確立に向けた『幼稚園の配布文書コーパス』の構築と分析、課題番号：20K00730、研究代表者：長谷川守寿）の研究成果の一部である。