論　文

日本大学生産工学部研究報告B
2006 年 6 月 第 39 巻

子供話し言葉コーパスの特徴語抽出に関する研究

中條清美*・西垣知佳子**・内山将夫***・中村隆宏****・山﨑淳史*****

Extracting Outstanding Children's Spoken Words from Children's Spoken Corpora

Kiyomi CHUJO, Chikako NISHIGAKI**, Masao UTIYAMA***,*
*Takahiro NAKAMURA**** and Atsushi YAMAZAKI******

An initiative began, in 2002, to teach English to young learners in Japanese primary schools. To provide the primary school English teachers with an important core vocabulary, the authors have created, from an on-going research project and analysis of 30 picture dictionaries, a base list of the 5,259 words considered to be most relevant to students' everyday lives. Children learn the majority of word meanings during their experiences with both oral and written language ; therefore the authors decided to analyze the real speech data of children. In this study, the following four children's spoken data sources were examined for their suitability for children's speech data : CHILDES (Child Language Data Exchange System) ; Moe, et al. (Vocabulary of First-Grade Children, 1982) ; PoW (The Polytechnic of Wales Corpus) ; and COLT (The Bergen Corpus of London Teenager Language). As a result of that examination, the authors discovered that the CHILDES produced what they consider to be the optimal vocabulary list for everyday words.

キーワード：話し言葉コーパス，子供用語彙，生活語彙，絵辞書，小学校英語教育

1．はじめに

公立小学校では平成 14 年度から「総合的な学習の時間」の中で「英語活動」を行なうことが可能になった。文部科学省の調査によると平成 16 年度には，92.1％の公立小学校で英語活動が実施され[注1)]，「小学校で英語を学ぶことが既成の事実となっている」（田中他，2005：67）と言われている[1)]。

英語学習の入門期としてきわめて重要な小学校英語であるが，指導内容について，文部科学省（2001）の「小学校英語活動実践の手引き」には[2)]，小学校では「子供の日常生活の中の身近な英語を扱うことに重点をおくこと」と記されており，「日常生活英語」の指導が重要な課題のひとつになっていることがわかる。しかし，小学校英語は「教科」ではないため，具体的な指導目標，指導内容，言語材料などに関する明確なガイドラインはなく，教師自らが教材を作成し，指導を行なっている。そのよ

*日本大学生産工学部教養・基礎科学系助教授
**千葉大学教育学部助教授
***情報通信研究機構主任研究員
****小学館
*****日本大学大学院生産工学研究科博士前期課程数理工学専攻 2 年

うな現状において，教材開発の指針となる「小学生のための指導語彙リスト」の選定が必要であることが指摘されている[3]。

そこで，我々は，「小学生のための指導語彙リスト」作成に向けたプロジェクトの一環として,「子供の日常生活の中の身近な語彙」が多く含まれている海外と国内で出版された絵辞書30冊に収載される語彙を収集，分析し，英語教育関係者および研究者が参照，活用できる基本統計量を付加した「英語絵辞書の語彙表」を公刊した[4]。

上述の研究で分析した絵辞書は文字言語を分析したものであった。しかしながら，「子供の語彙」とは子供が使う語 'those words which he uses'（Burroughs 1957：3）と定義されることを考えると[5]，絵辞書のような文字言語だけでなく，家族，友達，教師との会話を扱った音声言語，すなわち話し言葉も考慮に入れる必要があろう[6]。より実用性の高い「小学生のための指導語彙リスト」を選定するには，文字言語であり読み物である絵辞書の語彙だけでなく，実際に子供が使う話し言葉にも注意を払う必要がある。そして最終的には，子供の言語習得を導くためのインプットとなる読み物と話し言葉の両者を網羅し，統合した「英語絵辞書と子供の話し言葉に基づいた日常生活語彙表」を作成したいと考えた。

本研究は，そのような「英語絵辞書と子供の話し言葉に基づいた日常生活語彙表」を作成するための基礎研究であり，子供話し言葉コーパスの特徴語を抽出することを目的とする。具体的に行なったことは，1）現在，利用可能な子供話し言葉コーパス4種を収集し，2）それぞれから大人話し言葉と比較して特徴的な語を抽出して4種の「子供話し言葉特徴語リスト」を作成する。続いて，3）4種の特徴語リストが「小学生のための日常生活語彙表作成」という目的達成のために適切なリストであるかを検討し，最終的に，4)「小学生のための日常生活語彙表の作成」という目的に最も適った「子供話し言葉特徴語リスト」を選定することである。

2．言語資料

2.1　4種類の子供話し言葉コーパス

本研究で子供話し言葉を抽出するために収集された言語資料は，Child Language Data Exchange System（以下，CHILDES），Vocabulary of First-Grade Children（以下，MOE），The Polytechnic of Wales Corpus（以下，PoW），The Bergen Corpus of London Teenager Language（以下，COLT）の4種類である。**表1**にそれぞれの言語資料のタイトル，編纂年，編纂者，コーパスの概要，規模，収集方法，対象者の年齢範囲，データ形式，編纂目的，データの参照先や入手先を示した。なお，各項目の表記は原著の表記に従った。各資料の詳細を以下に記す。

2.1.1　Child Language Data Exchange System (CHILDES)

CHILDESは言語習得研究のための国際的な言語データ共有システムであり，英語をはじめ26ヶ国語の発話データが収められている大規模コーパスである。"The CHILDES Database"よりデータが公開されている[7]。本稿の目的は幼児・小学生の話し言葉を収集することであるので，CHILDESデータベースの中の"English-American Corpora"より，10種のサブコーパス (Bliss, Bohannon, Brown, Carterette and Jones, Evans, Garvey, Gathercole, Kuczaj, Tardif, and Van Kleeck) を選択してダウンロードした。サブコーパスを選択する際の基準は，1）被験者の年齢幅が2歳から5年生（10歳または11歳）の範囲内に含まれる[注2]，2）自然な会話場面で収集されている，の2点である。"English-American Corpora"（推定延べ語数500万語）より，延べ語数1,293,261語，異語数10,725語の言語資料を得た。

2.1.2　Vocabulary of First-Grade Children (MOE)

The Vocabulary of First-Grade ChildrenはMoe, A. J., Hopkins, C. J. and Rush, R. T. が収集した延べ286,108語，異語数6,412語の小学1年生（5歳から8歳）329名の話し言葉のデータである[8]。4つのトピックについてのインタビューの録音から収集された。トピックは，1）Tell me about your favorite game，2）Tell me about your favorite television show，3）Tell me about the best thing or the most exciting thing that ever happened to you，4）Tell me a storyであった (Moe, Hopkins and Rush，1982：14)。

2.1.3　The Polytechnic of Wales Corpus (PoW)

児童言語の統語･意味研究のために，6歳から12歳の児童120名より収集された約65,000語の話し言葉コーパスである。子供たちがレゴ (Lego) で遊んでいる時の会話と 'favourite games' または 'favourite TV programmes' というトピックのインタビューの録音から作成された。データは体系機能文法の枠組みによって構文解析が施され，各行に分析樹が付けられている。International Computer Archive of Modern English (ICAME) より入手できる[9]。

2.1.4　The Bergen Corpus of London Teenager Language (COLT)

ロンドンの13歳から17歳の少年少女の自然な会話を録音した約50万語のコーパスである。被験者の3～5日間のすべての会話が録音され，teenagerの友人，学校，家族に関する話題の会話が収集された。データには品詞タグが付与されている。ICAMEより入手可能で，分析結果はStenstrom，Anderson and Hasund（2002）に報告されている[10),注3]。

表 1　調査した 4 種類の子供話し言葉コーパス

	CHILDES	MOE	PoW	COLT
編纂年	2000	1982	1978-84	1993
編纂者	MacWhinney, B.	Moe, A., Hopkins, C. and Rush, T.	Fawcett, R. and Perkins, M.	a research team at the University of Bergen
概要	transcript and media data collected from conversations with young children	oral vocabulary of first-grade children	sample of children's speech in a play session or an interview	spontaneous coversations of London teenagers
規模	approximately 300 million characters (300 megabytes)	286,108 words containing 6,412 different words	65,000 words approximately	approximately half a million words
収集方法	contribution and data-sharing	interviews with 329 first grade children, from 1974 to 1981	small cells of 3 children were recorded at play with Lego bricks, and each child also interviewed by the same 'friendly' adult on his/her favourite games and TV programmes	recruits carried SONY Walkman with lapel microphones and recorded three to five days all the conversations they were engaged
対象者	the speakers involved are often young mono-lingual, normally developing children conversing with their parents or siblings	first grade children (ages 5 ; 10-8 ; 4), north-central Indiana, equal distribution of girls and boys	120 children aged between 6-12, children's English from Pontypridd, S. Wales, informal register	between 13- to 17-year old boys and girls from different parts of London and with varying socioeconomic backgrounds
データ形式	all of the data is transcribed in CHAT and CA/CHAT formats	alphabetical and frequency listing of the vocabulary	the parsed corpus is available in machine readable form, 65,000 words approximately, in 11,396 lines	includes prosodic transcription and was tagged with CLAWS C 6 tagset
目的	to provide tools for studying conversational interactions, child language acquisition	use of oral vocabulary lists in language arts and reading instruction; to provide support for language experience programs	psycholinguistic research into development of childrens' English between ages of 6 and 12, investigating the growing use of a variety of syntactico-semantic structures	to collect a reasonably large corpus of teenage language and make it available for research
参照先	http://childes.psy.cmu.edu/	Moe, A., Hopkins, C. and Rush, T. (1982) *Vocabulary of First-Grade Children*. Springfield : Charles C Thomas Publisher	Pow Corpus Manual http://khnt.hit.uib.no/icame/manuals/pow.htm ICAME	Stenstrom, A., Andersen, G. and Hasund, I. (2002) *Trends in Teenage Talk*. Amsterdam : John Benjamins Publishing Company ICAME

2.2　比較基準に用いた大人の話し言葉コーパス

「子供の話し言葉」を観察するには，「大人の話し言葉」と比較して，子供の話し言葉に特徴的に用いられた語（特徴語）を特定する必要がある。本稿では，統計指標を利用して，子供話し言葉コーパスに出現する語の生起頻度と，その語の「大人の話し言葉」コーパスでの生起頻度とを比較し，その出現の度合が顕著な語を子供話し言葉の特徴語として抽出するという手法を用いた。そのよう

な比較の基準となる大人の話し言葉コーパスには，British National Corpus の spoken component 1,036 万語を使用した。そこから求めた頻度 10 以上の 8,462 語（延べ語数 9,126,606 語）を大人話し言葉リスト（British National Corpus Spoken High Frequency Word List，以下 BNC SHFWL）として用いた[11]。

3．研究の方法

3.1　語彙リストの作成

4 種類の子供話し言葉コーパスの特徴語を抽出するために，次の手順に従って，それぞれのコーパスごとに語彙表を作成した。

Step 1：CHILDES，PoW，COLT には多様なタグが付与されているため，それらを除去し，目視による校正を 2 度行なった。MOE は語彙表が出版されているのでスキャナーを使用してデータを入力し，校正を 2 度行なった。

Step 2：コーパスごとにデータを単語単位に分割し，それらを基本形に集約した語彙表を 4 種類作成した。この工程には CLAWS とレマタイズプログラムを使用した[12),13)]。基本形に基づく本研究の語彙表では，cat-cats や go-goes-went-gone-going などの屈折形とその頻度はそれぞれ cat と go に集約された。

Step 3：作成された 4 種類の語彙表には出現頻度の低い語が多く含まれた。効率的な語彙学習の観点からはこれらの低頻度語の必要性は高くない。そこで，各語彙表において頻度 5 以上の語のみを分析の対象とすることにした。頻度 5 という境界値はコーパスサイズと，頻度 4 以下の語を除去した残りの語数とのバランスを考慮して決定した。ただし，PoW はコーパスの規模が小さいため，頻度 2 以上の語を対象とすることにした。このように語彙表を縮小することによって，次の Step 4 に示すような語彙表の精度向上のための人手による処理が可能になった。

Step 4：作成された語彙表には London，Marie，nineteen のような地名，人名，数字や，話し言葉に特有の ah，er，em，uh-uh などの間投詞や無意味語が多く含まれていた。多目的の言語使用に対応する教育用語彙表にはこれらは不要と考えられる[注4)]。CLAWS を利用したタグ付けによる品詞情報を参照しながら，数詞，固有名詞，無意味語等を人手で除外した。

Step 5：作成された語彙表には，特定のテキストにのみ高頻度に出現する語彙，例えば，ding（ゴーンという音），whack（ぴしゃり）のような語が含まれている。このような語彙には効率的な語彙学習を目指す教育用語彙選定には適切でないと考えられるものも多く含まれる。我々は，外国語としての英語教育における学習語彙は，現代英語を代表する 1 億語の BNC における頻度 100 以上の 13,994 語（延べ語数 86,112,272 語）[14)]の範囲内で十分であると考えている[注5)]。そこで，4 種の子供話し言葉の語彙表および大人話し言葉リスト (BNC SHFWL) から BNC 13,994 語に含まれていない語を除外した。

最終的に特徴語抽出に用いた語彙表の異語数，延べ語数は**表 2** のようになった。

3.2　子供話し言葉の特徴語の抽出

それぞれの子供話し言葉コーパスでの語彙の出現状況を，約 1,000 万語からなる BNC の大人話し言葉での出現状況と比較して，子供話し言葉に特徴的に使われている語彙を抽出した。使用した統計指標は次の 9 種である。1）頻度 (Freq : frequency)，2）ダイス係数 (Dice : Dice coefficient)[15)]，3）コサイン (cosine)[16)]，4）補完類似度 (CSM : complementary similarity measure)[17)]，5）対数尤度比 (LLR : log-likelihood ratio)[18)]，6）カイ二乗値 (Chi2 : chi-square test)[19)]，7）イエーツの補正公式 (Yates : chi-square test with Yates's correction)[20)]，8）自己相互情報量 (PMI : pointwise mutual information)[21)]，9）マクネマー (McNemar : McNemar's test)[22)]。これらの指標を利用して計算を行なう背景には，

表 2　特徴語抽出に使用した語彙表

	コーパス名	分析の対象	異語数（語）	延べ語数（語）
調査した語彙表（子供話し言葉）	CHILDES	頻度 5 以上	2,427	1,213,601
	MOE	頻度 5 以上	1,467	280,017
	PoW	頻度 2 以上	1,084	53,519
	COLT	頻度 5 以上	2,226	403,745
比較した語彙表（大人話し言葉）	BNC SHFWL	頻度 10 以上	8,462	9,126,606

もし単語 α の子供話し言葉における出現状況が，大人話し言葉である BNC SHFWL における出現状況よりも顕著であるならば，その単語 α は子供話し言葉の特徴的な単語であろうという考え方がある。各指標はそのような顕著性の度合を測定するために利用されている。各指標は定義式によって求められるが，9 種の統計指標の各々の定義式が異なるため，同一の単語 α であっても異なる指標値が与えられる。各統計指標の工学的な説明は内山他（2004）[23]，定義式は Chujo and Utiyama（2006）を参照されたい[24]。

以上の 9 種の統計指標を用いて，4 種の子供話し言葉の各語の出現状況を，BNC 大人話し言葉リストでの出現状況と比較した指標値を求めた。その値に従って降順に単語をソートし，特徴語リスト 36 種（4 資料× 9 指標）を作成した。

各指標の上位に順位付けられた特徴語の実例を観察するため，4 種の子供話し言葉コーパスごとに，9 指標によって抽出された最上位 70 語の特徴語リストを作成した。70 語という語数は本稿の 1 ページの紙幅で一覧できる語数を基準にして決めた。

3.3　特徴語リストの検討

本研究の目的は，子供話し言葉コーパスの特徴語を抽出することであった。その結果を踏まえ，最終的には，文字言語から抽出された「英語絵辞書の語彙表」と，音声言語から抽出された「子供話し言葉の特徴語リスト」を統合して，「小学生のための日常生活語彙表」を作成したいと考えている。そのためにはこのような日常生活語彙の選定に用いる適切な「子供話し言葉の語彙表」を捜す必要がある。そこで，4 種の子供話し言葉コーパスに 9 種の指標を適用して得られる特徴語リスト 36 種（4 資料× 9 指標）の抽出傾向を観察した。

それらの観察に基づき，まず，1）4 種の子供話し言葉コーパスのうちどのコーパスを使うと，本研究の目的に適った特徴語を抽出できるかを検討し，目的達成に最適な子供話し言葉コーパスを絞り込んだ。次に，2）選定した子供話し言葉コーパスに適用した 9 種の統計指標のうち，どの指標によって選定された特徴語リストが最も本研究の目的に適っているかを検討した。

3.4　絵辞書の語彙との比較による特徴語リストの検討

3.3 で導かれた結果をもう 1 つ別の角度から検証する。本研究は「小学生のための日常生活語彙表」開発のための基礎研究である。そのような語彙表に収載する語彙数を何語に設定するかの目安として，「英語の使える日本人」の育成を念頭においた小野（2005）や竹蓋・水光（2005）は[注6]，「500 語」を妥当な数と考えている[25),26]。

そこで，本研究では 500 語の学習語彙表を作成すると仮定して，すでに資料作成が完了している「絵辞書の語彙表」の上位 500 語と子供話し言葉の特徴語リストの上位 500 語がどの程度似ているのか，両者の共通語の割合を観察することにした。絵辞書の上位 500 語とは中條他（2005）において「ある単語が海外で出版された英語絵辞書の収載語として選択される確率」を基準にして順位付けられたものであり[注7),27]，子供話し言葉の 500 語は 9 種の統計指標上位に順位付けられたものである。

4．結　果

4.1　4 種の子供話し言葉の特徴語

子供話し言葉コーパス CHILDES，MOE，PoW，COLT に特徴的に使われている語彙を抽出した結果の一部を**表 3**，**4**，**5**，**6** に示した。紙幅の関係で，9 種の指標より求められた特徴語のうち指標値の高いものから上位 70 語を示した。各特徴語の指標値の表示は省略した。また，特徴語の指標値が同一値の場合は abc 順に順位付けた。なお，Freq/Dice，Chi2/Yates は上位 50 位に同じ語が現れたため同じ列に示した。

表 3，**4**，**5**，**6** に示した特徴語の上位 70 語は，各表ごとに同一の子供話し言葉コーパスから抽出されたものであるが，使用した指標によって上位に順位付けられた特徴語はかなり異なっている。一見して，指標が左から右にいくにつれて語彙レベルの高い語彙が抽出されていることがわかる。特徴語抽出に関する我々の先行研究から，9 種の指標のうちどの指標を使用するかによって，4 段階の異なる難易度レベルの特徴語の抽出が可能であることが判明している。頻度とダイス係数によって抽出される特徴語は「補習レベル」，コサインと補完類似度は「初級レベル」，対数尤度比，カイ二乗値，イエーツの補正公式は「中級レベル」，自己相互情報量とマクネマーは「上級レベル」の学習者に適した特徴語を抽出すると考えられる[28)~31)]。このような 4 段階のレベル分けは**表 3**，**4**，**5**，**6** の各表の上位 70 語にもほぼ該当しているようである。以下，それぞれの子供話し言葉コーパスごとに 9 種の指標別に得られた特徴語上位 70 語を観察する。

4.1.1　CHILDES の特徴語

表 3 に示した CHILDES の特徴語を指標別に観察していく。9 種の指標については，特徴語の抽出傾向と語彙レベルの似ているものをまとめて 4 群に分けて観察した。

⑴　頻度，ダイス係数 (Freq/Dice)

大人の話し言葉[32)]と順位は少し異なるものの，頻度リストに通常現れる高頻度の機能語が上位に抽出されている。

⑵　コサイン (Cosine)，補完類似度 (CSM)

基本的な動詞である go，want，see，put，like，play，make，eat，look，let，fall，break，watch，turn，sit，write，know，come，take，tell，fix などが多数現れて

表 3　CHILDES 特徴語の一部（上位 70 語）

	Freq, Dice	Cosine	CSM	LLR	Chi2, Yates	PMI	McNemar
1	be	you	you	what	what	antenna	auditorium
2	I	I	I	you	you	ape	automobile
3	you	be	what	do	daddy	auditorium	bumpy
4	do	what	do	here	no	automobile	catfish
5	it	do	go	I	here	awesome	frown
6	a	it	a	no	do	backyard	hiss
7	the	a	here	daddy	eat	barefoot	horseback
8	what	go	want	eat	I	baton	stairway
9	that	not	no	go	play	beetle	toad
10	to	that	this	want	baby	blot	underpants
11	not	here	see	play	want	bouncy	wavy
12	go	to	it	where	go	buckle	chimpanzee
13	and	no	where	put	where	buffalo	crescent
14	have	the	put	baby	put	bumpy	glider
15	this	want	not	see	truck	cannon	gruesome
16	he	this	one	too	pencil	carnival	leash
17	in	where	can	let	fish	catfish	nutritious
18	get	daddy	like	truck	too	cavity	snowy
19	can	see	play	why	doggy	chimpanzee	springtime
20	on	can	make	big	people	cinnamon	awesome
21	we	put	why	pencil	toy	closet	blot
22	will	he	eat	fish	let	creek	fort
23	want	eat	how	toy	big	crescent	fuzzy
24	see	play	look	doggy	cowboy	croak	croak
25	they	get	let	people	why	dame	dame
26	like	like	daddy	milk	see	deer	dynamite
27	here	on	some	cowboy	milk	destroyer	edible
28	there	one	too	fall	hat	doubly	mule
29	she	why	big	this	fall	dozen	tuna
30	know	make	little	make	papa	dune	watercolor
31	no	baby	baby	one	airplane	dynamite	bouncy
32	one	how	up	hat	next	edible	destroyer
33	put	and	down	how	ride	elf	doubly
34	of	look	on	break	honey	fingernail	pounce
35	where	in	fall	papa	break	fort	prairie
36	for	let	name	airplane	fix	french	rhino
37	up	too	break	next	snake	frown	wiper
38	look	will	out	ride	bout	fuzzy	buckle
39	with	have	car	look	lion	glider	creek
40	make	big	watch	fix	make	goblin	goblin
41	how	some	turn	like	bunny	goldfish	quiver
42	right	she	boy	honey	how	grapefruit	sardine
43	some	up	now	boy	one	gruesome	soy
44	all	we	over	some	shoe	harpsichord	spaceship
45	now	know	sit	ball	hurt	hiss	pastel
46	come	there	milk	bout	turtle	horseback	sucker
47	think	truck	he	shoe	this	kangaroo	baton
48	out	little	pencil	snake	ball	leash	dune
49	why	pencil	truck	hurt	mama	leopard	carnival
50	say	right	write	little	boy	lobster	spinach
51	at	now	kind	lion	finger	make-believe	antenna
52	let	fish	dog	watch	juice	measles	cavity
53	down	down	fish	bunny	monkey	mosquito	barefoot
54	so	out	off	chair	last	mule	shoemaker
55	just	come	who	a	dose	nap	snarl
56	play	toy	ball	name	chair	nutritious	surf
57	good	fall	toy	dog	candy	pap	pap
58	well	milk	take	finger	pretend	pastel	recital
59	take	people	chair	turtle	dinosaur	polar	subway
60	when	doggy	book	last	watch	pounce	beetle
61	too	with	right	cause	bug	prairie	harpsichord
62	who	break	come	mama	peanut	propeller	make-believe
63	big	cowboy	hat	juice	look	putt	wrench
64	eat	hat	color	monkey	nap	quiver	backyard
65	little	all	tell	pretend	duck	railroad	buffalo
66	then	name	pretty	pretty	dog	recital	polar
67	if	take	girl	dose	cause	reindeer	measles
68	daddy	good	cause	candy	mouth	rhino	wont
69	off	they	another	mouth	name	sardine	deer
70	over	boy	fix	nap	like	sheep	propeller

表 4　MOE 特徴語の一部（上位 70 語）

	Freq, Dice	Cosine	CSM	LLR	Chi2, Yates	PMI	McNemar
1	and	and	and	and	and	ape	bumpy
2	I	then	he	then	people	bumpy	collie
3	be	he	then	he	then	closet	railroad
4	the	I	go	bear	bear	collie	sheep
5	it	people	I	play	mama	deer	spaceship
6	he	the	get	people	porridge	dynamite	vampire
7	a	bear	she	one	he	kangaroo	nap
8	to	go	they	go	play	lacy	rhino
9	you	get	one	baby	baby	motel	sucker
10	they	they	like	eat	papa	nap	dynamite
11	get	be	play	get	candy	panther	lacy
12	go	one	all	she	next	racer	kangaroo
13	then	play	there	mama	cartoon	railroad	reindeer
14	that	it	little	porridge	yes	reindeer	deer
15	not	a	bear	little	ride	rhino	motel
16	have	she	up	guy	guy	sheep	racer
17	do	to	on	dog	eat	spaceship	panther
18	in	mama	down	game	one	spinach	woody
19	she	porridge	out	brother	wolf	sucker	trash
20	we	baby	house	ride	tag	trash	ape
21	on	papa	eat	house	brother	vampire	closet
22	there	you	too	papa	game	woody	spinach
23	one	like	big	jump	baseball	yes	peck
24	of	eat	run	real	ice	candy	pup
25	like	next	baby	run	jump	papa	garbage
26	this	candy	when	next	dog	porridge	octopus
27	all	there	dog	cartoon	fish	people	beret
28	say	cartoon	game	like	patrol	mama	cape
29	can	on	people	hit	snow	baseball	jeep
30	know	little	dad	ice	go	hello	spooky
31	up	guy	girl	candy	duck	next	dolphin
32	out	ride	real	too	monster	basketball	refrigerator
33	see	yes	no	catch	real	witch	elevator
34	when	all	man	snow	pig	cartoon	sitter
35	if	dog	around	fish	hit	gorilla	ligament
36	come	game	guy	ball	grandma	patrol	wizard
37	down	brother	brother	girl	little	robot	pistol
38	but	wolf	home	yes	witch	cone	skipper
39	little	jump	see	bed	last	quit	lizard
40	play	in	somebody	wolf	catch	wolf	recess
41	with	ice	a	big	ball	tag	gorilla
42	well	not	boy	dad	bunny	haunt	spinner
43	so	tag	bed	last	sister	bunny	curse
44	what	up	watch	sister	jail	sweater	heather
45	just	house	over	tag	get	wizard	gruff
46	because	fish	try	boy	she	robber	dresser
47	will	snow	sometimes	pig	flower	web	lightning
48	thing	real	stuff	duck	cream	monster	smear
49	bear	that	come	grandma	house	goat	parachute
50	put	baseball	hit	sometimes	run	hood	vanilla
51	at	patrol	catch	patrol	bed	spooky	bye
52	house	we	put	flower	ghost	piggy	vacation
53	take	run	jump	down	girl	spinner	penguin
54	some	hit	ball	watch	hide	gym	robber
55	back	down	back	they	bat	tiger	hug
56	no	duck	ride	around	mouse	vacation	frosty
57	time	monster	thing	cream	dad	bat	eagle
58	too	too	fall	baseball	basketball	beret	ultra
59	big	do	win	monster	goat	goose	paste
60	other	catch	card	chair	whoever	joker	haunt
61	about	pig	name	fall	boy	jail	chute
62	make	out	chair	win	too	motorcycle	graveyard
63	eat	ball	sister	card	sometimes	ghost	accidental
64	think	last	red	all	chair	duck	yell
65	over	grandma	mama	tree	bug	bug	ox
66	run	this	other	stuff	fun	lion	unlock
67	right	big	ice	somebody	big	ramp	chess
68	for	sister	porridge	hide	fall	puppet	lone
69	or	girl	last	red	tree	lizard	puppet
70	try	bed	sit	up	hello	recess	piggy

表 5　PoW 特徴語の一部（上位 70 語）

	Freq, Dice	Cosine	CSM	LLR	Chi2, Yates	PMI	McNemar
1	I	people	I	roof	people	a-level	I
2	be	roof	a	window	roof	avalanche	and
3	the	dun	get	door	dun	diver	it
4	a	window	not	put	window	dungeon	a
5	it	I	put	house	shutter	hairdressing	not
6	and	gate	we	people	gate	living-room	do
7	not	shutter	there	build	brick	multi-colored	avalanche
8	you	brick	look	gate	ladder	pike	dungeon
9	do	a	one	dun	door	series	hairdressing
10	we	door	make	look	fence	spaceship	multi-colored
11	have	be	do	brick	garage	subway	subway
12	that	ladder	can	play	build	wont	diver
13	get	fence	house	I	house	shutter	pike
14	they	house	on	make	wont	people	a-level
15	there	build	door	garage	put	dun	living-room
16	to	put	window	fence	aerial	monopoly	series
17	in	garage	no	a	monopoly	robber	spaceship
18	on	the	like	ladder	cartoon	roof	wont
19	can	not	play	red	play	bout	bout
20	go	get	build	one	red	rebound	rebound
21	will	we	go	get	tree	cartoon	salute
22	he	it	roof	car	wheel	aerial	thriller
23	this	wont	this	no	look	ladder	rainy
24	put	do	they	tree	bus	brick	spear
25	look	look	little	shutter	make	rainy	amusement
26	one	and	car	there	bridge	fence	skillful
27	of	play	man	bus	a-level	marble	mob
28	like	there	here	white	living-room	gate	swivel
29	make	make	up	man	series	amusement	ventilation
30	what	aerial	good	wheel	spaceship	mob	astronomy
31	all	monopoly	need	sometimes	white	slant	haunt
32	know	red	big	garden	car	cowboy	headmistress
33	house	one	it	bridge	garden	headmistress	robber
34	up	cartoon	gate	little	I	salute	shopkeeper
35	if	can	red	aerial	diver	thriller	farmhouse
36	no	they	all	blue	pike	garage	tribe
37	now	have	people	big	one	playground	slant
38	with	tree	thing	we	sometimes	den	ledge
39	good	on	brick	can	marble	inn	ulcer
40	door	you	garage	not	no	spear	den
41	she	wheel	some	cartoon	blue	pavement	flatten
42	thing	that	now	monopoly	a	anymore	murderer
43	for	no	dun	game	yellow	window	skull
44	window	bus	sometimes	yellow	man	shark	anymore
45	then	car	bus	wall	get	franc	shark
46	here	go	white	like	layer	skillful	franc
47	play	bridge	tree	need	pavement	layer	witch
48	some	white	fence	on	robber	wheel	whale
49	come	this	find	wont	avalanche	swivel	adventure
50	but	like	off	color	dungeon	ventilation	hood
51	just	man	ladder	fit	hairdressing	daisy	daisy
52	right	a-level	garden	here	multi-colored	astronomy	alley
53	little	living-room	another	square	subway	haunt	soccer
54	see	series	down	layer	game	shopkeeper	crocodile
55	build	spaceship	let	stick	there	grease	shutter
56	at	garden	game	marble	wall	skull	telescope
57	want	in	blue	pavement	little	cane	tiger
58	out	will	start	fall	big	chimney	muddle
59	think	sometimes	wheel	do	cowboy	farmhouse	runway
60	need	little	color	inside	color	bridge	hinge
61	man	blue	stick	good	fit	tribe	brownie
62	roof	he	bridge	find	climb	hairdresser	dawn
63	down	big	wall	another	square	overlap	preferably
64	car	diver	fit	football	path	kit	pillar
65	say	pike	finish	hole	playground	ledge	spade
66	when	yellow	small	climb	bout	doll	crumb
67	big	marble	enough	hat	rebound	settee	slam
68	so	up	fall	path	can	rainbow	grease
69	off	all	yellow	sister	we	ulcer	cane
70	where	game	square	finish	inn	telescope	amusing

表 6　COLT 特徴語の一部（上位 70 語）

	Freq, Dice	Cosine	CSM	LLR	Chi2, Yates	PMI	McNemar
1	be	I	I	people	people	annex	anonymity
2	I	be	you	I	next	anonymity	graduation
3	you	you	do	you	I	chariot	grenade
4	it	do	go	go	fuck	flowery	mellow
5	do	people	he	she	you	funky	flowery
6	the	it	not	do	shit	graduation	funky
7	not	not	she	fuck	go	gravitational	shrine
8	to	go	be	next	last	grenade	volcano
9	and	he	what	like	she	mellow	gravitational
10	that	she	like	he	do	metabolic	chariot
11	have	what	it	no	like	percent	series
12	a	to	get	what	no	series	metabolic
13	he	get	know	shit	tape	shrine	annex
14	go	and	no	last	shut	volcano	bodyguard
15	get	have	just	tape	he	people	lush
16	what	like	why	mum	what	next	malleable
17	she	that	really	not	mum	quid	melody
18	they	know	fuck	shut	sir	informer	bout
19	in	the	people	know	piss	kinetic	renewable
20	know	a	up	why	wicked	mathematics	nude
21	of	next	mum	sir	bitch	dick	ass
22	will	fuck	out	get	mathematics	bitch	ping
23	can	just	want	piss	quid	bout	sly
24	like	no	man	really	percent	renewable	knack
25	on	they	come	just	sad	wicked	mania
26	we	can	tape	man	why	juicy	sap
27	say	shit	tell	wicked	dick	tit	percent
28	this	on	can	sad	bastard	melody	cob
29	just	in	off	though	not	immature	stun
30	so	last	look	bitch	moon	basketball	ecstasy
31	but	say	alright	everyone	know	last	informer
32	there	will	shit	listen	crap	cunt	console
33	well	up	though	crap	everyone	ecstasy	slit
34	up	so	okay	bastard	stupid	bodyguard	hello
35	all	why	shut	moon	man	geezer	moth
36	think	really	talk	stupid	get	lush	acne
37	for	this	last	mathematics	really	malleable	arcade
38	with	mum	how	record	though	sly	indigestion
39	come	tape	sir	hate	listen	fruit	blazer
40	then	come	who	dick	hate	whore	ammonia
41	right	but	one	funny	just	flirt	catapult
42	see	out	hear	alright	truly	vodka	dope
43	about	want	next	miss	joke	hello	illiterate
44	if	well	listen	boy	fancy	tertiary	strangle
45	no	shut	see	quid	funny	shit	warrior
46	want	all	record	joke	record	anorexic	freak
47	out	of	school	tell	microphone	nude	scrum
48	one	see	boy	percent	miss	bastard	immature
49	really	then	here	fancy	cunt	cob	stereotype
50	when	right	girl	girl	boy	expel	anorexic
51	at	sir	bit	it	suck	moon	malaria
52	look	there	everyone	up	homework	piss	orgy
53	why	one	miss	microphone	alright	workman	chubby
54	good	look	funny	hear	conversation	bunk	lyric
55	mean	man	right	truly	girl	ping	alleyway
56	now	we	please	off	arse	raisin	amazingly
57	how	about	good	conversation	boom	truly	bullshit
58	who	tell	walk	laugh	laugh	suck	hug
59	or	with	all	be	cool	acne	whore
60	as	when	when	talk	tit	arcade	kinetic
61	tell	piss	so	friend	tell	sad	ozone
62	here	off	then	out	lesson	triumph	basketball
63	thing	think	piss	walk	swear	scrum	courtyard
64	off	how	stupid	homework	annex	blazer	grin
65	where	who	dad	reckon	bum	crap	doc
66	okay	though	friend	okay	reckon	ugly	grievance
67	fuck	alright	where	cunt	hear	gravity	cent
68	time	good	sit	lesson	kinetic	ass	blaze
69	down	okay	hate	cool	bore	slag	greasy
70	take	wicked	something	dad	friend	dawn	groove

いる。名詞はあまり抽出されていない。

⑶　対数尤度比 (LLR)，カイ二乗値，イエーツの補正公式 (Chi2/Yates)

コサイン，補完類似度と同様に動詞が多く現れており，上述の動詞に加え，pretend，do，ride，hurt も抽出されている。上位の語に名詞が多く出現し，子供の話し言葉らしい，動物 (fish，doggy，snake，lion，bunny，turtle，monkey，bug，duck，dinosaur) や人に関する語 (daddy，baby，people，cowboy，papa，boy，mama) が目を引く。20 位以降には日用品 (pencil，toy，hat，ball，shoe，chair)，食品 (milk，honey，juice，candy，peanut)，乗物 (truck，airplane)，50 位以降に身体 (finger，mouth) 関連の語が現れる。Doggy，bunny，daddy などの低年齢の児童向けの語彙がこれらの指標で抽出されているように見える。

⑷　自己相互情報量 (PMI)，マクネマー (McNemar)

上位の語は機能語と動詞に代って，圧倒的に名詞が多くなる。自己相互情報量とマクネマーの指標は子供の身の回りの生活に結びついた多彩な名詞や形容詞を豊富に抽出していると言える。例えば，20 位までの抽出を見ただけでも，動物関連では ape，beetle，buffalo，catfish，chimpanzee，toad，生活関連の名詞には antenna，automobile，auditorium，backyard，stairway，形容詞は awesome，bouncy，bumpy，wavy，gruesome，nutritious，snowy といった，より詳細な描写を可能にする語彙が抽出されている。これらの語は全般的に単語の長さが長く，語の示す対象が認知レベルの高いものであり，母語話者の子供話し言葉の中でも，年齢の高い児童によって使用された語彙のように見える。日本人の小学生向けの語も多少含まれているが，どちらかといえば，これらの名詞や形容詞は中級以上の英語学習者が語彙の幅を広げるのに役立ちそうである。

以上の⑴から⑷の 4 群の指標の観察により，⑶のグループの指標（対数尤度比，カイ二乗値，イエーツの補正公式）が「子供らしい」特徴語を「最も良く」抽出していると考えられる。

4.1.2　MOE の特徴語

指標別の抽出傾向はほぼ CHILDES と同じである。そこで上記の CHILDES の特徴語の観察に見られたように，4 群に分けた指標の中で，子供らしい特徴語を最も良く抽出していると思われる対数尤度比 (LLR)，カイ二乗値，イエーツの補正公式 (Chi2/Yates) の特徴語を中心に観察してみたい。

インタビューのトピックが“Tell me a story”であったためと考えられるが，物語に登場するような身近な人物に関する語 (people，baby，mama，guy，brother，papa，girl，dad，sister，boy，grandma)，物語に登場する動物やモンスターなどの語 (bear，wolf，pig，bunny，duck，mouse，goat，witch，ghost，monster) が抽出されている。また，“favorite games”に関する語とそれらに伴う動詞 (tag，card，baseball，hide［and seek］，ball，bat，game，basketball，play，ride，jump，run，hit，catch，win) があり，“favorite television show”に関する語と動詞 (cartoon，watch) もある。また，食品に関する語 (eat，porridge，ice，candy，cream) も抽出されている。

自己相互情報量 (PMI)，マクネマー (McNemar) では上位に特定の動物を示す語 (ape，collie，deer，kangaroo，panther，reindeer，rhino，sheep) が抽出された。

4.1.3　PoW の特徴語

PoW の特徴語には子供たちがレゴ (Lego) で遊んでいる時の会話で用いられた語が顕著に抽出されている。指標別の傾向はほぼ CHILDES や MOE と同様であるので，比較的中庸の特徴語を抽出している対数尤度比 (LLR)，カイ二乗値，イエーツの補正公式 (Chi2/Yates) の特徴語を中心に観察したい。

特徴語には，レゴの色に関する red，white，blue，yellow，color，multi-colored や，レゴの大きさや形に関する big，little，square，layer，そしてレゴ遊びで交わされる会話を連想する put，build，brick，make，need，fit，here，fall，do，inside，good，find，another などがある。レゴで作っているものは家とその周囲の建造物であろうかと想像させられるような roof，window，door，house，gate，garage，fence，ladder，tree，garden，bridge などが並んでいる。また，“favorite games”というトピックに関係すると思われる monopoly，cartoon，play，game，finish，marble，football，playground なども抽出されている。

4.1.4　COLT の特徴語

COLT の特徴語は上述した CHILDES，MOE，PoW の 3 種の特徴語とは異質である。指標別の抽出傾向は COLT についても他の 3 種のコーパスとほぼ同様であるので，ここでも，対数尤度比 (LLR)，カイ二乗値，イエーツの補正公式 (Chi2/Yates) を中心に特徴語を観察してみる。

ティーンエイジャーの会話の最も顕著な特徴は，*slangage* と *smallwords* にあると言われる (Stenstrom，Andersen and Hasund，2002：XI)[33]。実際に，スラングの中でも *proper slang* に分類される man，sad，wicked，cool，quid や *dirty slang* の crap，dick，bastard，bitch，piss，fuck が**表 6** の特徴語の上位 4 位から 40 位の間に順位付けられている。*Swearwords* である shit も 20 位以内にある。また，‘and things like that’ などで用いられる like のような *smallwords* (Stenstrom，Andersen and Hasund，2002：63) も 10 位前後に順位付けられ，特徴度が高い[34]。そしてティーンエイジャーの会話に現れる多彩なトピックの中には mathematics，homework，lesson

など学校に関する話も含まれている。

COLT を本研究の調査対象の 1 つに選んだ理由は，ティーンエイジャーの身の回りの日常生活語彙が抽出できるかもしれないという期待があったからである。しかし，それらは上位にはほとんど抽出されなかった。

4.2　特徴語上位の観察のまとめ

1）最適な子供話し言葉コーパスの選定

「4 種のコーパス別特徴語の抽出傾向」を**表 3，4，5，6** に示した最も特徴度の高い上位 70 語について見た結果は以下のようにまとめられる。上位 70 語だけでは網羅的な比較はできないが，ある程度の傾向は観察できると考える。

- CHILDES からは基本的な動詞や，子供の身の回りの生活に結びついた多彩な名詞と形容詞が豊富に抽出されている。
- MOE からはインタビューのトピックに関連した身近な人物や物語に関する子供らしい名詞や動詞が抽出されている。
- PoW からはレゴ遊びに関連した形容詞や動詞，建造物に関する名詞，そしてゲームに関連した名詞や動詞が抽出されている。
- COLT からはティーンエイジャーの会話の特徴として様々なスラングの実例が抽出されている。

以上の考察から，MOE と PoW で抽出される特徴語はインタビューのトピック等の影響が強く見られること，COLT は本稿の目的に適っていないことが判明した。その結果，「小学生のための日常生活語彙表」に使用する「子供話し言葉コーパス」には，偏りなく一般的な子供の話し言葉を網羅していると考えられる CHILDES が適していると結論された。

2）最適な統計指標の選定

「9 種の統計指標別の抽出傾向」を**表 3，4，5，6** のそれぞれに示された 4 つの指標群に分けて観察した結果は以下のとおりである。

- 頻度とダイス係数は頻度表の上位に通常現れる機能語を中心とする特徴語を抽出しており，4 種のコーパスを通じてよく似た抽出結果となっている。子供話し言葉としての特徴度は低い。
- コサインと補完類似度は，子供の動作や遊びに関する基本的な動詞と，身の回りで使われる初級レベルの名詞を抽出している。ただし，COLT の場合は基本的な動詞は抽出されているが名詞は少ない。
- 対数尤度比，カイ二乗値，イエーツの補正公式は，基本的な動詞をはじめ，動物や人間，物語や遊びに関連する名詞を豊富に抽出し，日本人児童向けに適度にバランスのとれた子供話し言葉の特徴語を抽出しているように見える。ただし，COLT の場合はスラングが多く抽出された。
- 自己相互情報量とマクネマーは名詞や形容詞を豊富に抽出している。しかし語彙の難易度の点で日本人児童向けの語彙は多くなく，どちらかと言えば，中級レベル以上の学習者向けと考えられる特徴語を抽出している。

以上の結果から，「対数尤度比，カイ二乗値，イエーツの補正公式」の指標群の特徴語が本稿の目的に最も適っていると判断された。

以上の考察は主観的な観察である。そこで，「どの子供話し言葉」の特徴語リストに日常生活語彙が多く含まれるか，そして，「どの統計指標」が日常生活語彙をよく抽出しているかを客観的に把握する必要がある。次に，絵辞書の語彙との共通語の割合を求めて類似度を検討し，主観的な観察結果の妥当性を検討した。

4.3　子供話し言葉の特徴語リストと絵辞書の語彙との比較

本研究グループが既に作成した「英語絵辞書の語彙表」の上位 500 語と，今回抽出した 4 種の子供話し言葉の特

表 7　4 種の子供話し言葉コーパスと絵辞書の語彙上位 500 語の重なり

		子供話し言葉コーパス			
		CHILDES	MOE	PoW	COLT
統計指標	Freq/Dice	44.2%	40.4%	31.8%	26.0%
	Cosine	48.8%	39.0%	27.4%	24.0%
	CSM	51.4%	43.6%	30.8%	26.6%
	LLR	48.4%	38.6%	28.4%	19.6%
	Chi2/Yates	47.2%	38.0%	28.8%	18.8%
	PMI	31.6%	33.2%	27.6%	14.6%
	McNemar	6.8%	21.4%	22.4%	9.2%
平均		39.7%	36.3%	28.1%	19.8%

徴語上位 500 語がどの程度似ているのか，両者の共通語のパーセンテージを求めた（**表 7**）。絵辞書の語彙は日常生活語彙を多く含むことから，絵辞書の語彙との共通語の割合が高いコーパスは，子供の日常生活をよりよく反映した語彙を含むコーパスと考えられる。

表 7 では，頻度 (Freq) とダイス係数 (Dice)，カイ二乗値 (Chi2) とイエーツの補正値 (Yates) はそれぞれ同一値であったので結果を Freq/Dice, Chi2/Yates のようにまとめて示した。最下段には 9 指標の共通語の平均パーセンテージを示した。

表 7 をまず横方向に見ていく。4 種の子供話し言葉コーパスの中では，全般的に CHILDES の共通部分の割合が一番高い。自己相互情報量 (PMI) とマクネマー (McNemar) の指標については，上級レベルの特徴語を抽出する指標であるため，絵辞書の語彙との共通部分が少ないのは当然の結果と考える。自己相互情報量，マクネマーを加えて単純平均を求めた最下段の平均値でもやはり CHILDES が一番高く，MOE，PoW，COLT と続いた。従って，CHILDES の特徴語が絵辞書の語彙と共通部分が多いと言える。**4.1.1** の特徴語リストの実例の考察から，「小学生のための日常生活語彙表」作成のための子供話し言葉コーパスとして CHILDES が最も適していると判断されたが，その判断の妥当性が再度確認された。

次に，CHILDES の縦列を見る。絵辞書の語彙との共通度の算出結果は，補完類似度 (CSM)，コサイン (Cosine)，対数尤度比 (LLR)，カイ二乗値とイエーツの補正値 (Chi2/Yates) が高く，いずれも 50％前後となっていることから，これらの指標が妥当と判断される。**4.1.1** の特徴語リストの実例の考察から，対数尤度比とカイ二乗値/イエーツの補正値が日常生活語彙の抽出という目的に適った指標と考えられたが，その判断は適切であったことが確認された。

最終的にどの指標を 1 つ選ぶかということであるが，我々はこれまでの特徴語抽出の先行研究の実績を考え合わせて対数尤度比を選ぶことにした[35]。対数尤度比は数学的にも安定した指標と言われており[36]，大学英語教育学会の JACET 8000 基本語彙の開発でも使用されている，広く知られた指標であることも考慮した[37]。

5．まとめ

本研究は，「小学生のための日常生活語彙表」の作成に向けた基礎研究である。4 種の子供話し言葉コーパスでの語彙の出現状況を，1,000 万語からなる British National Corpus の大人の話し言葉での出現状況と比較して，子供の話し言葉に特徴的に使われている単語を抽出した。結果を比較したところ，CHILDES コーパスに対数尤度比を適用して抽出した特徴語が，「小学生のための日常生活語彙表」の作成という目的に最も適した「子供話し言葉の語彙表」であろうと判断された。

2002 年に始まった小学校英語では，文部科学省による明確なガイドラインがなく，「日常生活に関する英語」ということ以外に具体的な指針は示されていない。指導語彙に関しては，小学校教員が個人的経験に基づき主観的に選定し，指導している。そのような現状のもと，語彙選定の際の参考資料となる科学的手法に基づく信頼性の高い「小学生のための日常生活語彙表」の作成が待たれている。本研究の結果は，今後，この分野の研究に貢献するものと考える。

謝辞　本研究は，平成 16～17 年度科学研究費補助金・基盤研究(C)課題番号 16520331 および，平成 17～18 年度科学研究費補助金・基盤研究(C)課題番号 17520401 の援助を受けている。

参考文献

1）田中茂範，アレン玉井光江，根岸雅史，吉田研作（2005）「小中高大一貫の英語教育をデザインするための枠組み」『英語教育』54 (8)：67-72.
2）文部科学省（2001）『小学校英語活動実践の手引』東京：開隆堂出版.
3）佐久正秀，本田勝久（2004）「小学校英語教育における語彙教材の開発に向けて」『第 30 回全国英語教育学会長野研究大会発表要綱』JA 長野県ビル，8/8/2004，598-599.
4）中條清美，西垣知佳子，内山将夫，岩楯弘美，山﨑淳史（2005）「英語絵辞書の語彙」『日本大学生産工学部研究報告』38：77-105.
5）Burroughs, G.E.R. (1957) *A Study of the Vocabulary of Young Children*. University of Birmingham, School of Education.
6）Honig, B. (2001) *Teaching Our Children to Read*. Thousand Oaks, California : Corwin Press, Inc.
7）MacWhinney, B. (2000) *The CHILDES Project : Tools for Analyzing Talk. 3rd Edition. Vol. 2 : The Database*. Mahwah, NJ : Lawrence Erlbaum Associates. http://childes.psy.cmu.edu/data/.
8）Moe, A., Hopkins, C. and Rush, T. (1982) *Vocabulary of First-Grade Children*. Springfield : Charles C Thomas Publisher.
9）International Computer Archive of Modern English (ICAME) (1999) The ICAME Corpus Collection on CD-ROM, Version 2. http://nora.hd.uib.no/icame/newcd.htm.

10) Stenstrom, A., Andersen, G. and Hasund, I. (2002) *Trends in Teenage Talk*. Amsterdam : John Benjamins Publishing Company.
11) 中條清美，西垣知佳子，内山将夫，中村隆宏，山﨑淳史（2006）「BNC 口語 3 分野からのレベル別 ESP 語彙の抽出」『日本大学生産工学部研究報告』39.
12) CLAWS7 (1996) http://www.comp.lancs.ac.uk/computing/users/eiamjw/claws/claws7.html.
13) 中條清美，内山将夫（2005）「語彙分析入門：lemma リストの作成」第 26 回英語コーパス学会ワークショップ，昭和女子大学，10/22/2005.
14) Chujo, K. (2004) "Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatised High Frequency Word List." In : J. Nakamura, N. Inoue, and T. Tabata (Eds.), *English Corpora under Japanese Eyes*, Amsterdam : Rodopi, 231-249.
15) Manning, C.D. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge : The MIT Press.
16) Manning, C.D. and Schütze, H. (1999)，前掲論文.
17) Wakaki, M. and Hagita, N. (1996) "Recognition of Degraded Machine-printed Characters Using a Complementary Similarity Measure and Error-Correction Learning." *IEICE Trans. Inf. & Syst*. E79-D, 5.
18) Dunning, T.E. (1993) "Accurate Methods for the Statistics of Surprise and Coincidence." *Computational Linguistics*, 19(1) : 61-74.
19) Hisamitsu, T. and Niwa, Y. (2001) "Topic-Word Selection Based on Combinatorial Probability." *NLPRS-2001* : 289-296.
20) Hisamitsu, T. and Niwa, Y. (2001)，前掲論文.
21) Manning, C.D. and Schütze, H. (1999)，前掲論文.
22) Rayner, J.C.W. and Best, D.J. (2001) *A Contingency Table Approach to Nonparametric Testing*. New York : Chapman & Hall/CRC.
23) 内山将夫，中條清美，山本英子，井佐原均（2004）「英語教育のための分野特徴単語の選定尺度の比較」『自然言語処理』11 (3)：165-197.
24) Chujo, K. and Utiyama, M. (2006) "Selecting Level-Specific Specialized Vocabulary Using Statistical Measures." *System*, 34(2).
25) 小野博（2005）「小学校における身につく英語学習法の開発」『日本児童英語教育学会第 26 回全国大会資料集』中部大学，6 /12/2005，66-69.
26) 竹蓋幸生，水光雅則（2005）『これからの大学英語教育』東京：岩波書店.
27) 中條清美，西垣知佳子，内山将夫，岩楯弘美，山﨑淳史（2005），前掲論文.
28) 中條清美，内山将夫（2004）「統計的指標を利用した特徴語抽出に関する研究」『関東甲信越英語教育学会研究紀要』18：99-108.
29) 内山他（2004），前掲論文.
30) 中條清美，内山将夫，長谷川修治（2005）「統計的指標を利用した時事英語資料の特徴語選定に関する研究」『英語コーパス研究』12：19-35.
31) Chujo, K. and Utiyama, M. (2005) "Selecting Level-Specific BNC Applied Science Vocabulary Using Statistical Measures." *Selected Papers from the Fourteenth International Symposium on English Teaching*, English Teachers' Association/ROC Taipei, 195-202.
32) 中條他（2005），前掲論文.
33) Stenstrom, A., Andersen, G. and Hasund, I. (2002), 前掲書.
34) Stenstrom, A., Andersen, G. and Hasund, I. (2002), 前掲書.
35) Chujo, K., Nishigaki, C. and Utiyama, M. (2005) "Selecting 500 Essential Daily-Life Words for Japanese EFL Elementary Students from English Picture Dictionaries and a Children's Spoken Corpus." *Proceedings of Inaugural International Conference on the Teaching and Learning of English in Asia*, Penang, Malaysia, 11/15/2005.
36) 久光徹，丹羽芳樹（1997）「統計量とルールを組み合わせて有用な括弧表現を抽出する手法」『情報処理学会自然言語処理研究会資料』NL-122：113-118.
37) 大学英語教育学会基本語改訂委員会（2003）『大学英語教育学会基本語リスト JACET List of 8000 Basic Words』
38) 西垣知佳子，中條清美，岩楯弘美（2005）「絵辞書で学ぶ日常生活語彙－国内・海外絵辞書の語彙比較－」『日本児童英語教育学会第 26 回全国大会資料集』中部大学，6 /11/2005，21-24.

注

注 1) 文部科学省
http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo3/siryo/015/05071201/005/002.pdf

注 2)「2 歳から 5 年生」は幼児および小学生にあたる．我が国では小学生というと 6 年生までを意味するが，米国では 5 年生までが小学校で，6 年生からはミドルスクールという地域も多い．

注 3) The Bergen Corpus of London Teenage Lan-

guage (COLT) は現在 BNC の一部となっており，BNC SHFWL の 5％を占める．BNC SHFWL よりこの部分を除外するのが理想的であり，今後の課題の１つである．

注 4) 固有名詞や数字等は特定のテキストに集中して出現することが多いので，語彙リストの比較の際には除去されることが多い．これらの語は特徴度が非常に高いため，我々が直観的に考える「特徴語」の抽出には障害となる．本研究では，意味のある普遍的な結果を得るため，これらの語をすべての語彙リストから人手で取り除いた．ただし，実際の教育場面での指導には，背景的知識の説明とともに固有名詞の指導は必要と考える．

注 5) この 13,994 語は教育用語彙としては非常に大きな語彙である．例えば，一般的な大学生の語彙数は 2,000 語前後という研究結果がある．また，中学・高校教科書の語彙をすべて学習したとしても 3,000 語前後にすぎないことも明らかにされている．

注 6) 「英語の使える日本人」の育成を念頭において，小野（2005）と竹蓋・水光（2005）では小学校英語教育での指導語彙数として以下の目安をあげている．小野（2005：67）：小学校：500～1,000 語，中・高：3,000～4,000 語，竹蓋・水光（2005：60）：中学以前：生活関連用語 500 語，中学：1,000，高校：2,000，大学：2,000 語，修士：1,000，博士：1,000，社会人：500 語．

注 7) 我々は海外と国内で出版された絵辞書を各 20 冊と 10 冊，計 30 冊収集した．西垣・中條・岩楯（2005）により両絵辞書の収載語彙を比較した結果[38]，両者の編集方針は大幅に異なることが明らかになった．そこで，本稿では海外絵辞書を使用することにした．

（H 18．1 .10 受理）