

第3章 新聞に見る基幹外来語

中山 恵利子
桐生 リか
山口 昌也

1 はじめに

本稿で対象にするものは、「新聞1紙の10年分の語彙」という限られた範囲からではあるが、その中でさまざまな紙面にわたり何度も繰り返し使われている、いわば基本的な外来語とでもいうべきものである。ただし、本稿では「基本外来語」ではなく、「基幹外来語」という用語を用いる。本研究は、新聞1紙10年分に現れる、使用範囲が広く使用率が高い外来語の実態をそのまま把握することを目的としており、どのような分野にこの結果を応用するかについてはその先の問題となるため、真田(1977)の「基幹語彙」の定義に従い、「基幹外来語」とする。

ここで真田(1977)の定義を簡単にまとめておく。「基本語彙」はある目的のために、それに応じた言語資料を使って語彙調査をし、その結果に語の質的な面を加味して語を選定したものである。それに対して「基幹語彙」とは、ある特定の語集団を対象として語彙調査を行い、その調査から直接に得られるその語集団の骨格としての部分集団のことである。したがって、「基本語彙」とは、ある目的のために必要な言語資料の「基幹語彙」に、語の質的な面を加味したもの、と換言できよう。

特定の言語活動の実態をそのまま切り取る「基幹語彙」や、ある目的のために語彙調査に基づき選定される「基本語彙」については、国立国語研究所を始めとして研究されてきているが、外来語だけを対象とした研究となると、教育のために選定者の判断に基づいて選ばれた「基礎語彙¹」が中心になり、「基幹語彙」についての研究は、管見の限りでは、存在しない。本来、現代日本語の使用実態を反映させた「基幹語彙」研究のためには、あらゆる日本人のあらゆる言語生活をデータ化することが理想的であるが、現状ではそのようなデータは存在しないため、本研究では現代日本人の社会生活の一部を反映していると考えられる新聞を利用して、「新聞に見る基幹外来語」のリストを作成した。また、そのリストの他に、語義や複合語形や出現頻度の経年変化など、「基幹外来語」の実態をいくつかの観点から切り取り、図表を付した。

これらのリストと図表は、国立国語研究所の「外来語」のホームページ上での公開を予定している。情報が大量であるため利用する人にとっては電子化データのほうが扱いやすいからである。

以下の各章で、まず、リストとそれに付随する図表の種類を俯瞰し、次に、リスト作成に利用した新聞記事データベースの概要、リストと図表の作成方法について説明し、最後に、リストと図表の応用の可能性について述べる。

2 リスト・図表の種類

●「新聞に見る基幹外来語」リスト一覧²（リスト1～4は本報告書にも掲載）

リスト1： 一般用法のある語

リスト1-1 頻度順

リスト1-2 エントロピー値順

リスト1-3 面種別頻度順

リスト2： 固有名詞用法のある語

リスト3： アルファベット表記語

リスト4： 和語・漢語としての用法のある語

リスト5： 頻度100以上のカタカナ表記語・アルファベット表記語

リスト6： 基幹語彙としてのカタカナ表記語・アルファベット表記語

●「新聞に見る基幹外来語」図表一覧

図表A： 語義分布表

図表B： 複合語情報

図表C： 経年変化グラフ

本稿では、上記の「リスト1～6」を「リスト」と、「図表A～C」を「図表」と呼ぶこととする。

3 新聞記事データベースについて

リスト作成には、「CD-毎日新聞」(日外アソシエーツ)という新聞記事データベースの1994～2003年版(10年分)を利用した。このデータベースは、毎日新聞の東京版・大阪版を収録したフルテキストデータからなる。それぞれの記事に対して、掲載日、朝夕刊の別、掲載面、ページなどの情報が付与されている。

「CD-毎日新聞」に収録されるすべての記事のうち、大阪版を除いた東京版の記事を対象とする。大阪版を除いたのは、大阪版1996～2001年の紙面に、東京版にはない「健康」という面が設定されていて同一に扱うことができないためである。

「CD-毎日新聞」1994～2003年・東京版(朝夕刊)の総量は、記事にして860,660、総文字数は452,307,833である。各年の総文字数を図1に示す。

なお、社会面や経済面といった面の種類のことを以下、「面種」と呼ぶ。面種は16面あり、各面における総文字数は、図2のとおりである(図2の「スポ」は「スポーツ」の略)。

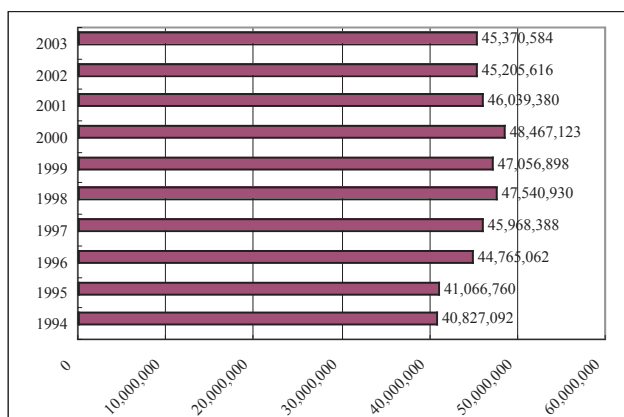
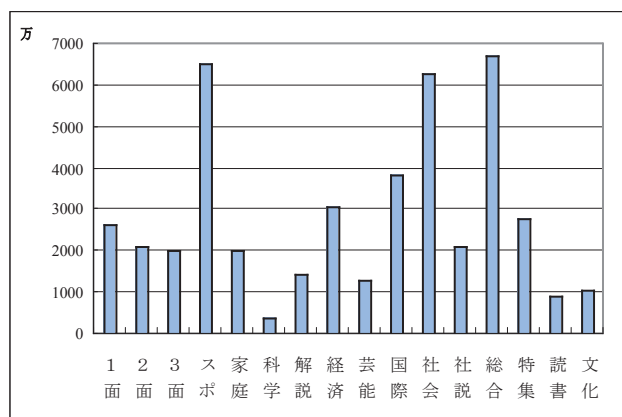


図1 各年の総文字数「CD-毎日新聞」

図2 各面の総文字数「CD-毎日新聞」
1994年～2003年の10年分

スポーツ面・社会面・総合面のように、毎日、複数ページにわたって設けられる面の記事量は、当然のことながら突出し、その面自体の出現回数が少ない、科学面などの記事量は極端に少ない結果となっている。

4 リスト・図表の作成方法

4.1 リスト作成方法

「CD-毎日新聞」1994～2003年(東京版)の全ての記事から、次の基準(正規表現)によって、10年間の総頻度が100以上である、カタカナもしくはアルファベットに数字、一部の記号を含む2文字以上の文字列を頻度情報とともに抜き出した。

[0-9 アーヴ a-z A-Z] [0-9 アーヴー a-z A-Z. /・:]*[0-9 アーヴー a-z A-Z]

※ただし、数字と . だけからなる文字列は除く。

したがって、「10メートル」「15:00」「イメージ」「ニューヨーク・タイムズ」「OL」「JRグループ」のような文字列が抽出される。そこから、「15:00」、「A4」、また「1カ国」の一部である「1カ」のような、

1語とはなりえない文字列を対象から取り除いた³。また、「アイデンティティー」と「アイデンティティ」、「バイオリン」と「ヴァイオリン」のような異表記の統合を行い、頻度を足した。上記の基準では連続して取り出す文字列に数字が含まれているため、「10メートル」「100メートル」等は「メートル」とは別にそれぞれ独自に抽出される。それらについても「10メートル」「100メートル」等の頻度を「メートル」単独の頻度に足し込んだ。さらに、「JRグループ」のようなアルファベット表記語とカタカナ表記語による複合語は「JR」と「グループ」に分け、別々に頻度を足しこんだ。その結果、8226語とその頻度の情報を抽出した（頻度順「リスト5」）。

次に8226語から、新聞に見る、基幹語彙としてのカタカナ表記語とアルファベット表記語を選び出した。「基幹語彙」と選定する際の指標は次の2つである。

- (1) 新聞における使用頻度が高いこと（多用されている）
- (2) 新聞の多くの面に使用されていること（汎用されている）

(1)については10年間の総出現頻度を、(2)については各面における出現頻度の分布のばらつきを示すエントロピー値を利用して、多用される語、汎用される語をそれぞれ求めた。

語 w の面種 i における頻度を f_{wi} 、総頻度を F_w とすると、語 w のエントロピー値 H_w は次の式で表される。

$$H_w = - \sum_i \frac{f_{wi}}{F_w} \log \frac{f_{wi}}{F_w}$$

なお、エントロピー値は、16面全てに同確率で出現する場合は最大値「4」となり、一つの面にしか出現しない場合は最小値「0」となるものである。したがって、エントロピー値が高いほど、あらゆる面に同数ずつ出現する確率が高く、汎用される語と考えられる。エントロピー値の計算に際しては、面の間の「頻度1が持つ重み」の格差を解消するため、面ごとの記事量の違いを考慮した。つまり、各面における出現頻度を新聞全体に占める各面の大きさに応じた値に換算し、それをエントロピー値の計算に用いた。

換算値を f'_{wi} としたときの、換算の式は以下のとおりである。

$$f'_{wi} = f_{wi} \div \frac{\text{面}i\text{の文字数}}{\text{全体の文字数}} \times 100$$

今回は、およそ1000語をリストやそれに付随する図表の作成が可能な語数の目安とし、(1)(2)両方の基準を満たす語を選定した。(1)では総出現頻度の高い順に全体の25%（頻度617以上の2057語）を、(2)では、エントロピー値の高い順に全体の50%（エントロピー値2.71857以上の4113語）を取り出し、双方に重複して現れる語を求めると1304語であった。

次に、1304語の各語につき新聞における実例を500例、ランダムサンプリングにより抽出した⁴。その用例を見ると、上記の正規表現で示した文字列抽出基準により、本来切り取られるべきでない箇所で文字列が切り取られてしまい、1304語の対象語に入ってしまった語が見つかった。「イレブン」「インターナショナル」「サン」「スタンダード」「セブン」の5語である。これら5語は実際には「基幹語彙」としての抽出基準をクリアしていないため、対象から外した⁵ので、対象語は1299語となった。この1299語が「新聞に見る、基幹語彙としてのカタカナ表記語・アルファベット表記語」であり、それを頻度順に並べたものが「リスト6」である。この1299語からアルファベット表記語68語を取り出し（「リスト3」）、残りの1231語のカタカナ表記語について、その用例を分類した。以下、分類の手順を示す。

まず、500用例の中の対象語を単純語形か複合語⁶形かで分類した。その結果、500用例全てが複合語形である語は「エネ」の1語であり、1例でも単純語形があるものが1230語であった。

次に、1230語について、用例中の単純語形のみを「一般用法、固有名詞、和語・漢語」に分類した。「一般用法」とは一般名詞、形容詞、副詞などの用法のことである。分類の際の基準は以下のとおりである。

- ① 固有名詞は、人名（動物の呼称等）、地名、その他に分類する。その他には会社名、組織名などが入る。例：「マイクロソフト」、「ビートルズ」

②和語・漢語に由来すると思われる固有名詞であっても、固有名詞に分類する。

例：「トヨタ」→固有名詞（その他）, 「タマ」→固有名詞（人名）

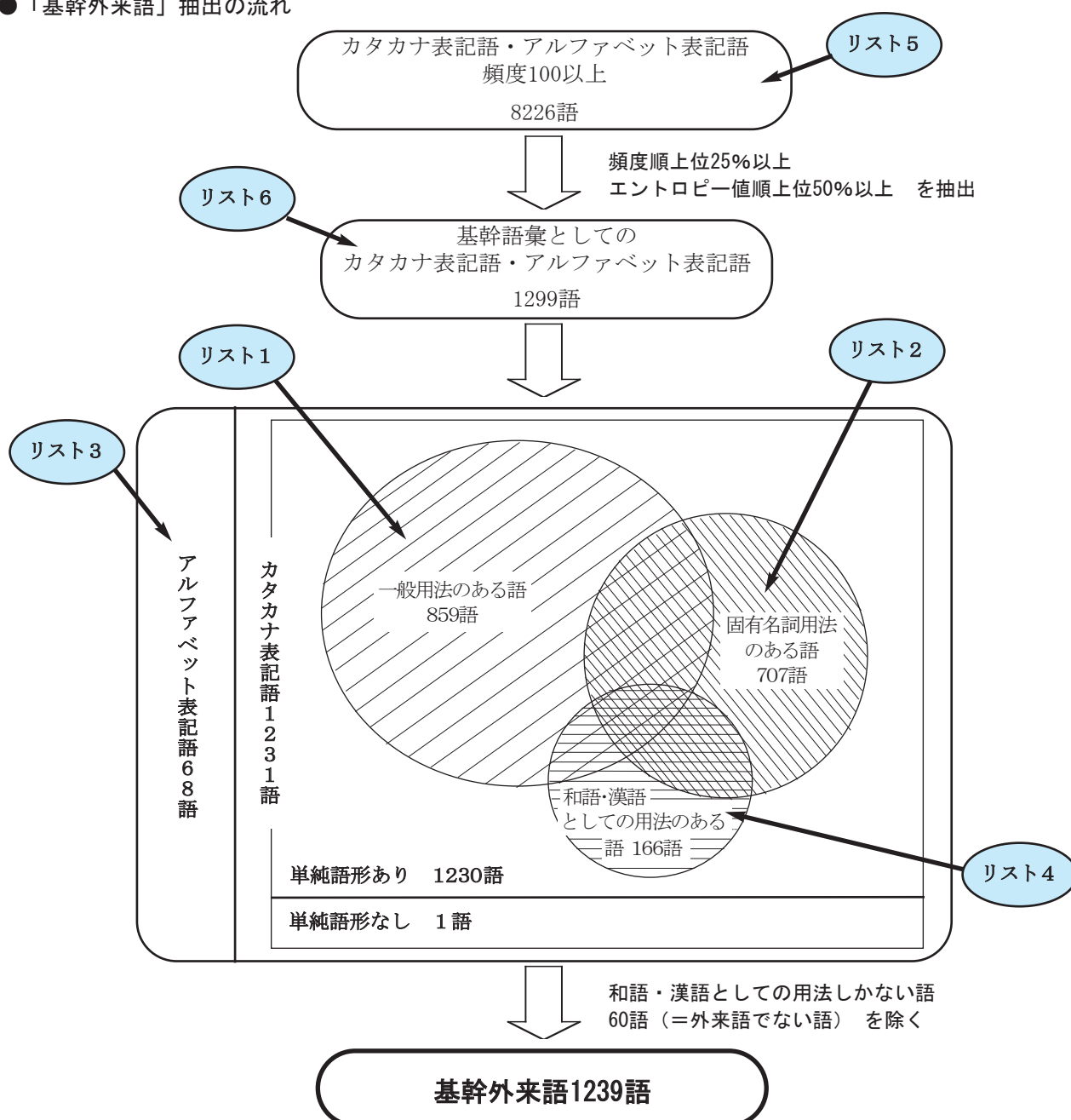
この分類により、「一般用法」が1例でもある語を以下、「一般用法のある語」と呼ぶ。

また、「固有名詞用法」が1例でもある語を「固有名詞用法のある語」と、「和語・漢語としての用法」が1例でもある語を「和語・漢語としての用法のある語」と呼ぶ。したがって、これらの語には重複があるが、それについては次節で述べる。

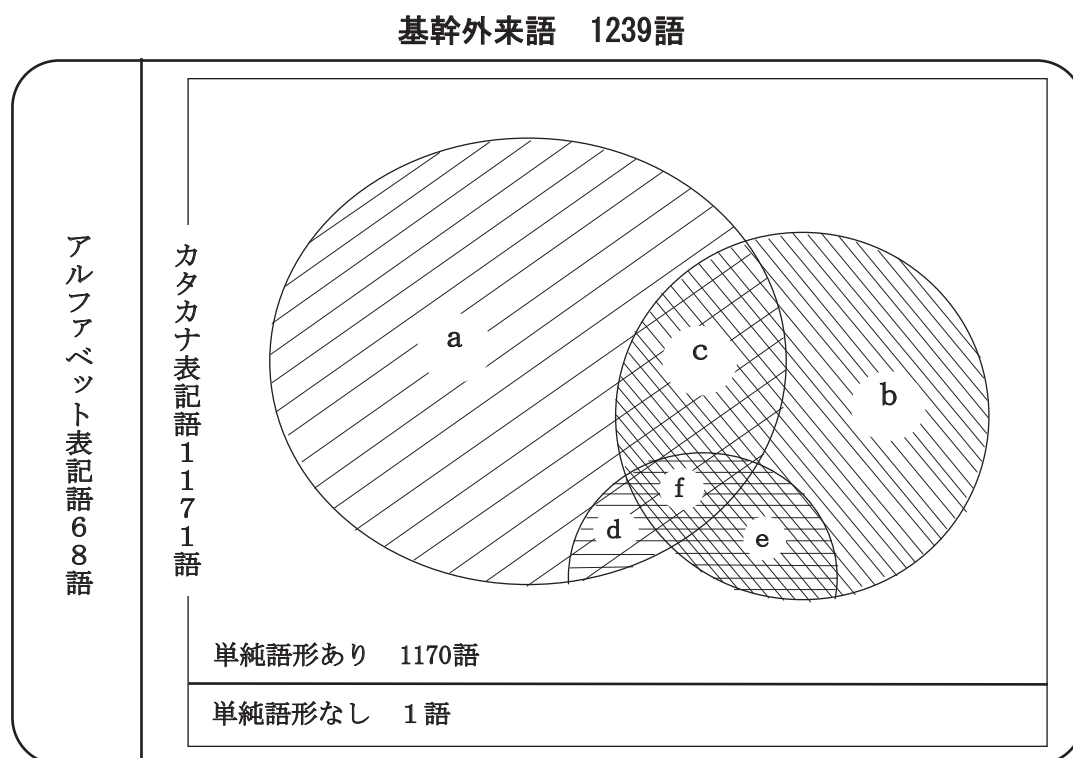
分類結果は次のとおりである。

一般用法のある語は859語（一般用法のみの語472語）であり、「新聞に見る基幹外来語（一般用法のあるもの）」（「リスト1」）として次節における語別の図表作成の対象とする。固有名詞用法のある語は707語（固有名詞用法のみの語251語）で、これは「リスト2」に、和語・漢語としての用法のある語は166語（和語・漢語としての用法のみの語62語）であり、これは「リスト4」に掲載する。

●「基幹外来語」抽出の流れ



●「基幹外来語」内訳図



- a : 一般用法のみの語 453語
 b : 固有名詞用法のみの語 249語
 c : 一般と固有名詞の両方の用法のある語 362語
 d : 一般と和語・漢語の両方の用法のある語 10語
 e : 固有名詞と和語・漢語の両方の用法のある語 62語
 f : 一般と固有名詞と和語・漢語の3つの用法のある語 34語

★一般用法のある語 = $a + c + d + f = 859$ 語

★固有名詞用法のある語 = $b + c + e + f = 707$ 語

4.2 図表作成方法

4.1で述べた手順により得られた、一般用法のある語859語について、新聞における使用実態を反映した3つの図表を作成した。

図表A : 語義分布表

図表B : 複合語情報

図表C : 経年変化グラフ

以下、それぞれの図表について、具体的に説明していく。

図表A : 語義分布表

「図表A」は、500用例における対象語を意味分類し、用例数を付したものである。

4.1で述べたとおり、リストを作成する段階で、500用例における対象語を複合語形と単純語形に分け、それぞれの件数を出し、単純語形を「一般用法」「固有名詞（人名・地名・その他）」「和語・漢語」に分類し、用例数を算出した。その一連の結果も「図表A」には掲載してある。ここでは、さらにそのうちの「一般用法のある単純語形」について、用例を意味によって分類し、その意味と用例数とを付した。

分類に用いた語義は、『コンサイスカタカナ語辞典』（第3版）に従ったが、必要に応じて筆者らで設定しなおしたものもある。これらの語義に基づき、500用例を分類した。辞書に掲載されていない語義が見つかった場合には、新語義を立て、その新語義の冒頭に＊を付した。これら全ての作業結果を一覧にしたのが「図表A」である。

下記の表には、別語と多義語のある語として「ガード」を、「一般用法」「固有名詞」「和語・漢語」の3つの分類項目にまたがる語として「メス」を例に挙げる。「メス」のように複数の分類項目にまたがる語も少なくない。

ここで、「図表A」とリストとの関係を確認しておこう。「図表A」において複数の分類項目にまたがる語は複数のリストに掲載されることになる。たとえば、「メス」は「一般用法のある語」として「リスト1」に、「固有名詞用法のある語」として「リスト2」に、「和語・漢語としての用法のある語」として「リスト4」に、というように各リストに重複して掲載されている。

【図表A：「ガード」「メス」】

katakana (代表形)	外来語								和語 漢語		
	複合語	単純語	外来語（一般）				外（固有名詞）				
			別語		多義		語義別	人名		地名	その他
ガード	148	352	①	guard	[1]	防護（する） ボクシングなどの防御の構えもこちら	266				
					[2]	守衛，警備員，番人．〈大〉⇒ガード - マン．	5				
					[3]	バスケット， アメフトなどの選手	71				
			②	道路をまたぐ鉄道橋．〈昭〉				10			
メス	8	262	洋式小刀の 1 種． 外科用小刀， 解剖刀．〈江〉				259	3			230

図表B： 複合語情報

「図表B」に掲載する複合語には、次の3種類の形がある。1つめは、「デジタル＋放送」「電子＋メール」のように和語・漢語との複合語を作るもの、2つめは、「イメージ＋アップ」「キャッシュ＋カード」のようにカタカナ表記語同士の複合語を作るもの、3つめは、「Eメール」「NHKドラマ」のようにアルファベット表記語との複合語を作るものである。

以下に「デジタル」の例を示す。

【図表B】

○「デジタル」和語・漢語との複合語

複合語	単独 ⁷	含む ⁸	複合語	単独	含む
デジタル放送	34	102	デジタル加入者線	4	15
デジタル化	82	101	デジタル方式	12	12
地上波デジタル	1	33	デジタル多用途ディスク	11	11
地上波デジタル放送	22	24	非対称デジタル加入者線	11	11
デジタル技術	19	19	デジタル通信	1	11
地上デジタル	5	17	デジタル情報	9	10
デジタル多チャンネル	2	17	地上デジタル放送	9	10
デジタル家電	8	15	デジタル多チャンネル放送	5	10

＊以下、省略

○「デジタル」カタカナ表記語同士の複合語

複合語	頻度	複合語	頻度
デジタルカメラ	822	デジタル・ビデオ・ディスク	140
デジタルトレンド	560	デジタルデバイド	131
サービスデジタル	187	デジタルホン	106
デジタルテレビ	179	デジタルビデオカメラ	104

○「デジタル」アルファベット表記語との複合語

複合語	頻度	複合語	頻度
B Sデジタル	932	C Sデジタル	212

1つめの和語・漢語との複合語については、500用例中にどのような複合語が何例現れるかを調べた。2つめのカタカナ表記語同士の複合語については、本リスト作成の母体である頻度100以上のカタカナ文字列8226から対象語（文字列）を含む文字列をその頻度の情報とともにすべて抜き出し、ごみ⁹を取り除いた上で表に記載する方法を用いた。3つめのアルファベット表記語とカタカナ表記語の複合語は、4.1で述べたとおり、抽出した文字列の整理の過程でアルファベット表記語とカタカナ表記語とに分割して扱ったものをここで再度複合語として掲載しなおした。

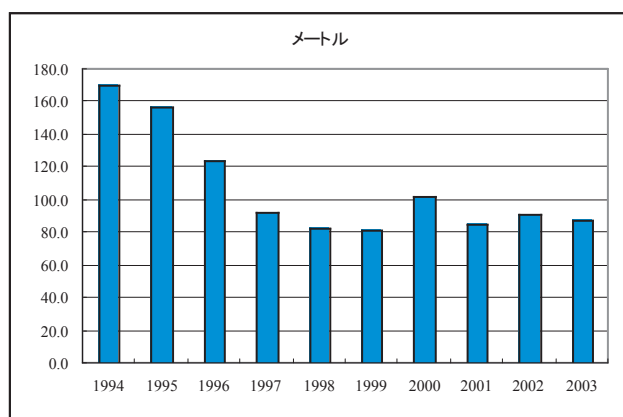
以上、3つの複合語情報を各語についてまとめたものが「図表B」である。

図表C： 経年変化グラフ

「図表C」の経年変化とは、各語の1994～2003年の出現頻度の推移である。3章で述べたとおり「CD-毎日新聞」の各年における記事量にはばらつきがあるため、グラフ化の際には単純な出現頻度ではなく、出現頻度をその年の文字数で割り、 10^6 倍にした数値を使用している。

【図表C】

○「メートル」経年変化グラフ



5 リストや図表の応用の可能性

5.1 リスト

「リスト1」は、新聞における外来語の使用実態を反映させたリストである。新聞に見る基幹語彙は何かという問いに答えるのは頻度順の「リスト1-1」である。また、面種別の頻度を載せたリスト（「リスト1-3」）からは、たとえば経済面でよく使われる語が何かがわかる。さらに、一般用法のある語だけを

対象としているリスト1だけでなく、固有名詞用法や和語・漢語としての用法を持つ語も含んだ「リスト5」や、新聞コーパスから抽出した8226語のアルファベット表記語・カタカナ表記語に頻度やエントロピー値、面種別頻度の情報を付した「リスト6」を付した。これらのリストがあれば新聞コーパスを利用した外来語研究に何らかの貢献ができるであろう。

また、固有名詞が一例でもある語を掲載した「リスト2」は、たとえば地名で並べ替えをすれば、地名に使われている語が取り出せるので、目的に応じた利用ができる。

「リスト4」は、外来語ではないが、和語・漢語のカタカナ表記のリストである。カタカナ表記される語は外来語だけではないということが、このリストから明白になる。新聞における和語・漢語のカタカナ表記のリストは他に類を見ないものであろう。

5.2 図表

「図表A」からは、実際に新聞に用いられているが辞書に掲載されていない、新しい語義（図表中、語義の冒頭に*を付与）がわかる。たとえば「ミルク」は、『コンサイスカタカナ語辞典』では「①牛乳.〈明〉②⇒コンデンス-ミルク.」となっているが、新聞での用例を見ると、乳児に与える飲料を指す例が多数見つかる。また、ウシに限らず、ヤギやヒツジ、マウスなどの母乳を指して「ミルク」とする例もあった。したがって「ミルク」の語義は、「①牛乳.〈明〉②⇒コンデンス-ミルク. ③人や動物の赤ちゃんに与える飲料. ④牛以外の動物の乳.」の4つとなる。

【図表A：「ミルク」】

katakana (代表形)	外来語								和語 漢語	不明 その他 ごみ など	
	複 合 語	単 純 語	外来語（一般）				外（固有名詞）				
			別語 ＜語義＞	多義＜語義＞		語 義 別	人名	地名			そ の 他
ミルク	185	308		[1]	牛乳. 〈明〉	140	1		2		7
				[2]	⇒コンデンスミルク.	1					
				[3]	* 人や動物の赤ちゃんに与える飲料.	160					
				[4]	* 牛以外の動物の乳.	4					

「ミルク」以外にも、「ダイヤル」には「伝言ダイヤル」や「募金ダイヤル」のような、「受付専用電話番号」という意味で使用されている用例がある。「新聞に見る基幹外来語」にも少しではあるが、新しい語義が生じていることがわかる。これらの情報は、辞書作成や教育などに役立つ情報であろう。

また、「図表A」からは、新聞ではどの語義がどのくらい使われているかもわかるため、教育現場ではどの語義をどのような例文を用いて教えるかを考えるときの目安となる。たとえば、「クール」という言葉からは「(温度・色などが) 冷たいさま、涼し気なさま.」という語義をまず思い浮かべる人がいるだろうが、実際の用例を見ると、この語義の用法はほとんどない。「(性格・態度・表情などが) 冷静で、ものに動じないさま、興奮していないようす.」という意味の、「人」に関する用法のほうが圧倒的に多いことがわかる。

「図表B」の複合語情報からは、「タイミング」「プレッシャー」のようにほとんど複合語を作らない語のほかに、「コース」「タイプ」など複合語のバリエーションが豊富な語や「タブー」「オフ」のようにバリエーションが少ない語などがわかる。また、複合語の各件数からはその対象語の典型的な複合語例がわかる場合も少なくない。たとえば、「タブー」は複合語のバリエーションが少ない上に、「タブー視」という複合語が複合語形の8割を占めており、「タブー」を教えるときに「タブー視 (する)」などの形を一緒に教えておくことが有効だと言える。

【図表B】

○「エネ」和語・漢語との複合語

複合語	単独	含む	複合語	単独	含む
省エネ	212	450	省エネ住宅	6	6
エネ庁	11	22	省エネ推進	6	6
新エネ	7	22	改正省エネ法	5	6
省エネ対策	19	21	エネ庁長官	4	4
省エネ技術	17	21	省エネ策	4	4
省エネ効果	18	19	省エネ性能	4	4
省エネ法	10	18	省エネ投法	4	4
省エネ基準	11	16	新エネ法	3	4
省エネ型	11	15	高エネ研	3	3
省エネ努力	11	12	省エネ機器	3	3
省エネ化	11	11	省エネ作戦	3	3
省エネ性	3	8	省エネ政策	3	3
省エネ意識	6	6	省エネ投球	3	3

*単独の頻度3未満（90語）は省略

※カタカナ表記語同士の複合語，アルファベット表記語との複合語なし

○「コース」和語・漢語との複合語，カタカナ表記語同士の複合語

複合語	単独	含む	複合語	単独	含む
初コース	9	9	コース取り	2	2
東京コース	8	9	コース紹介	2	2
滑降コース	3	8	コース適性	2	2
周回コース	1	7	円コース	2	2
コース設定	4	4	研修コース	2	2
男子滑降コース	3	4	広島C C八本松コース	2	2
別コース	1	4	散歩コース	2	2
コース料理	3	3	宿泊コース	2	2
難コース	3	3	出世コース	2	2
平たんコース	3	3	専用コース	2	2
観光コース	2	3	袖ヶ浦コース	2	2
基礎コース	2	3	同コース	2	2
小回りコース	2	3	日帰りコース	2	2
特別コース	2	3	府中コース	2	2

*単独の頻度2未満（152語）は省略

複合語	頻度	複合語	頻度
インコース	163	コースレコード	156
フルコース	159		

※アルファベット表記語との複合語なし

○「タブー」和語・漢語との複合語

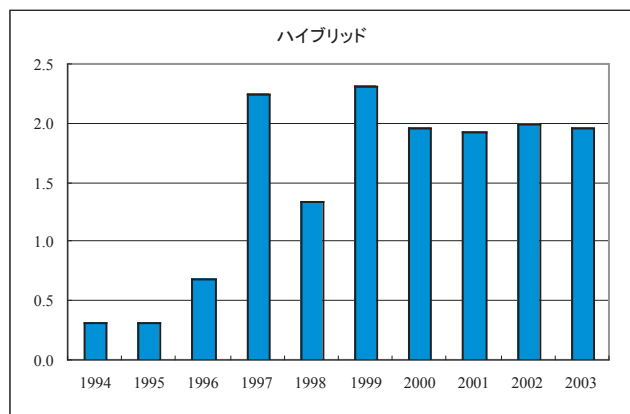
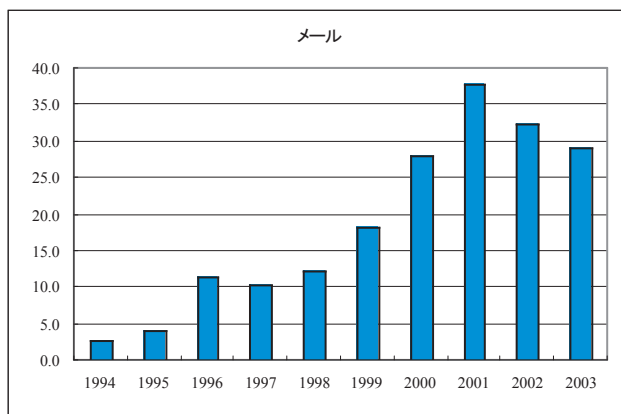
複合語	単独	含む	複合語	単独	含む
タブー視	118	118	タブー感覚	1	1
タブーなし	5	5	タブー侵犯	1	1
タブー破り	4	4	タブー人格	1	1
タブー扱い	3	3	憲法タブー	1	1
タブー化	3	3	社会的・道徳的タブー	1	1
戦後タブー	2	2	性的タブー	1	1
タブー感	1	2	政治的タブー	1	1
25%タブー	1	1	天皇タブー	1	1
タブー意識	1	1			

※カタカナ表記語同士の複合語，アルファベット表記語との複合語なし

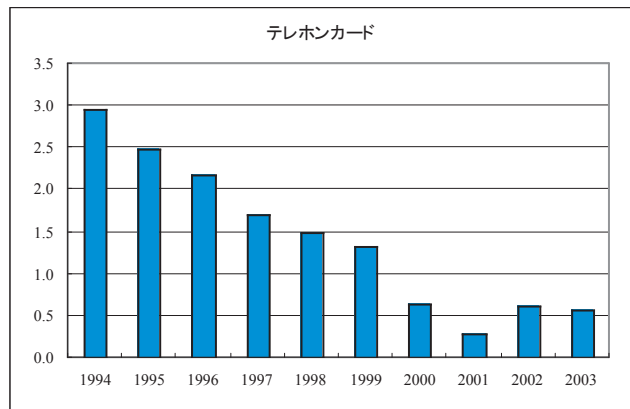
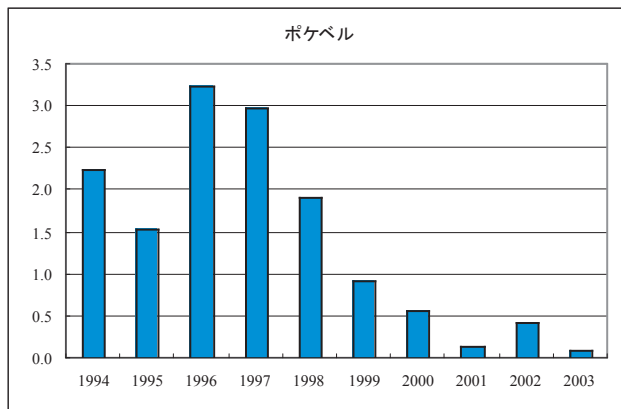
「図表C」は，語別の出現頻度の経年変化のグラフである。増加している語として，「メール」「ハイブリッド」「フリーター」「バリアフリー」などがあり，減少している語として「ポケベル」「テレホンカード」「マルチメディア」「クーラー」などがある。さらに，激減している語として「スチュワーデス」が挙がる。また，一時的に大量に使用された語は，1995年の「プレハブ」，1997年の「ダイアナ」「カメラマン」，1999年と2000年に「ミレニアム」，2001年の「ハンセン」「センタービル」，2002年の「サッカー・ワールドカップ」「フーリガン」などである。これらの語はその年のできごとを代弁している，と言えよう。今回は「基幹外来語」の研究であるので，このような頻度の推移をリストに反映させていないが，減少してしまい，すでにほとんど使用されなくなっている語や一時的に大量に使用された語などは「基本外来語」を考える際には，対象からははずすといった手続きが必要になるう。

【図表C】

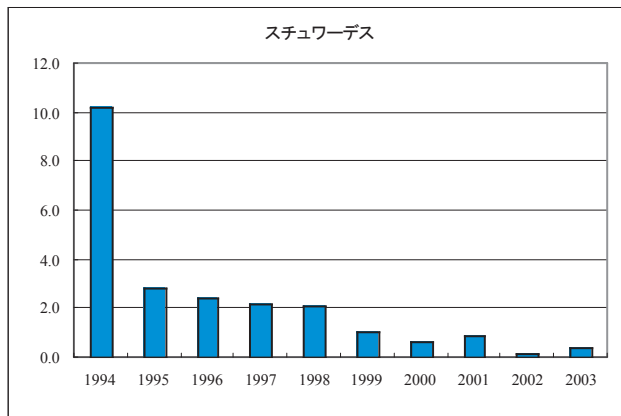
○増加している語「メール」「ハイブリッド」



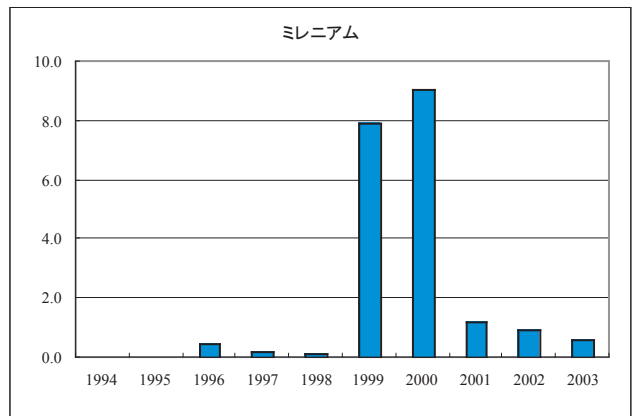
○減少している語「ポケベル」「テレホンカード」



○激減している語「スチュワードス」



○一時的に大量に使用された語「ミレニアム」



6 おわりに

本研究は、「新聞1紙」「10年分」という限りのあるものではあるが、過去にこれだけの大量のデータから抽出した外来語のリストやそれに関する図表が公開されたことはなく、辞書作成や研究や教育などにも活用される基礎的なデータとなることを願っている。

今後は、複合語情報を語義別に分けたり、経年変化による分類を試みたりして、今回のデータをさらに充実させていく予定である。また、「基幹外来語」とは言っても、時事的な影響を避けられない新聞をデータに、流行り廃りの影響を避けられない外来語を対象語として扱う限り、情報の新しさは必要になってくるので、リストのほうも更新の可能性を探っていきたい。

注

- 1 基礎語彙については、真田(1977)に従わず、慣例に従う。『日本語学キーワード事典』によれば、「主に言語教育を目的とし、日常の言語生活に最低限必要であると、個人の理論や言語観察に基づいて選定された語の集合」を指す。
- 2 リスト4, 5, 6は和語・漢語を含んでいたが、多用・汎用される語集団ではなかったりして「基幹外来語」とはいえないが、ここでは便宜上リスト6までを「新聞に見る基幹外来語」リストと呼ぶ。
- 3 「A4」は1語と考えられるが、本稿ではアルファベット1文字は対象外とするため、「A4」のようなアルファベット1文字と他のものとの複合語も除く。
- 4 サンプルは、毎日新聞1994～2003年（東京版）の全記事のうち、タイトル部分を除いた本文のみを対象として行った。これは、タイトル部分は文脈が短く、それだけでは語義の区別がつかない可能性を考慮したものである。
- 5 たとえば、「セブン」と「イレブン」は「セブン-イレブン」という形の語が「-」で別々の語として切り取られてしまい、「セブン」と「イレブン」の頻度が多くなってしまったものである。「セブン」も「イレブン」も「セブン-イレブン」の用例を除くと、1304語の基準（頻度617以上、エントロピー値2.71857以上）を満たさなくなる。
- 6 ここでいう複合語とは、和語・漢語とカタカナ表記語との複合語である。カタカナ表記語同士の複合語は正規表現で示した文字列抽出基準により1語として切り出されている。例えば、「マン」の複合語は「営業マン」「商社マン」などであり、「サラリーマン」は1語としてリストの項目になっている。つまり、ひとかたまりのカタカナ文字列として切り出されたものは単純語として扱っている。「エイズ」も「エイズウイルス」も単純語である。ただし、図表Bの複合語情報には「エイズウイルス」は「エイズ」の複合語として掲載している。
- 7 「単独」とは、その複合語形がそれ以上他の語と結合せず、単独で使用されている場合（「デジタル放送が…」における「デジタル放送」）を示す。
- 8 「含む」は、その複合語形がさらに他の語と結合し、より長い複合語を構成している場合（「地上デジタル放送が…」における「デジタル放送」）を指す。
- 9 ここでいう「ごみ」とは、たとえば、対象語「キー」に対する「キーブ」「スキー」「ミッキーマウス」のようなものである。これらは「キー」を含む文字列として一旦は抽出されるが、「キー」の複合語ではないため複合語の資料から除外することとなる。

参考文献

- 小池清治他編（1997）『日本語学キーワード事典』朝倉書店
真田信治（1977）「基本語彙・基礎語彙」『岩波講座日本語9』87-132，岩波書店
三省堂編修所編（2005）『コンサイスカタカナ語辞典第3版』三省堂

次ページより，以下，「新聞に見る基幹外来語」リスト1～4を掲げる。

- リスト1： 一般用法のある語 859語
 リスト1-1 頻度順
 リスト1-2 エントロピー値順
 ※語頭の「*」は，一般用法のみの語（453語）を表す
 リスト1-3 面種別頻度順
 ※各面種につき，出現頻度順上位100語を掲載
- リスト2： 固有名詞用法のある語 707語
 ※頻度順
 ※語頭の「*」は，固有名詞用法のみの語（249語）を表す
- リスト3： アルファベット表記語 68語
 ※頻度順
- リスト4： 和語・漢語としての用法のある語 166語
 ※頻度順
 ※語頭の「*」は，和語・漢語としての用法のみの語（60語）を表す

なお，本報告書には掲載していない，リスト5，6，図表A，B，Cも合わせ，すべてのリストと図表は，国立国語研究所「外来語」ホームページで，電子ファイルによる公開を予定している。