新聞記事データベースを利用した外来語の使用頻度の推移調査

柏野和佳子　　山口昌也　　桐生りか　　田中牧郎
国立国語研究所

１．はじめに

国立国語研究所では 2002 年 8 月に「外来語」委員会を設置し，主に，省庁の行政白書に使用されている外来語のうち，一般への定着が不十分でその意味が分かりにくいと思われる語を対象に，分かりやすく言い換えるなどの対応策を世の中に提案している[1]。

本稿では，委員会が第 1 回，第２回の検討対象語として選定した 109 語について，その使用状況を明らかにするために，読売新聞社の 1986 年から 2001 年までの 16 年分の記事データを用いて，各年の使用度数とその総計とを出すのに加え，年ごとに使用頻度を算出し，16 年間にわたる使用頻度の推移の型を調べた。このような大規模な経年調査による語彙調査は従来ほとんど行われていなかったが，定量的に使用状況を明らかにすることは外来語の盛衰を論じる上でも重要であると考えている。

最初に，新聞記事データベースを使って使用状況を調査する上で問題になった，使用状況のとらえ方と対象語の同定について考察する。語彙調査にとってまずは対象データと得ようとする数値の吟味は欠かせない。その上で，使用頻度の推移の型別に分析した結果を述べる。推移の型は大きく 4 とおりに整理できた。年々使用頻度の増えるものが多い一方で，同じような使用頻度のまま推移してきているもの，このまま定着せずに消えていくかのように見える，年々使用頻度の減るもの，年によって使用頻度の増減が入れ替わるものとがあった。

２．調査方法と問題点への対処

2.1. 読売新聞記事データの概要

本調査には読売新聞の 1986 年から 2001 年までの 16 年分の記事データを使用した。これは現在入手可能な新聞記事データのうちで，最も年数の多いものである。データの概要は次のとおりである。

- 読売新聞社の記事データベースに蓄積してあるもののうち，著作権が読売新聞社に無いものを除いた全記事。
- 1990 年以降は東京本社発行記事に加え，地方支社（大阪，西部，中部）発行記事が入っている。
- 1999 年以降はさらに県版（地域版）の記事が入っている。
- 新聞紙面記事との違いは，内容を表すものに見出しを再編集している点である。また，紙面上で見開き２ページにまたがってレイアウトされた記事をデータベーに採録する際には，ページごとに採録せず，右ページの記事としてまとめて採録し，左ページのデータとして見出しのみ，もしくは本文の一部のみを採録している点である。

なお，読売新聞記事データには本文や見出しの他に，キーワード（統制語及び自由語），分類コード，面種コードなどの書誌情報も入っているが，今回の使用状況調査で用いたのは本文のみである。調査に用いた本文の文字数のグラフを図１に示す。図１より記事データ量が年々増加したことが明らかである。特に，支社発行記事が加わった 1990 年と，県版（地域版）記事が加わった 1999 年とを境にデータ量が増加している。

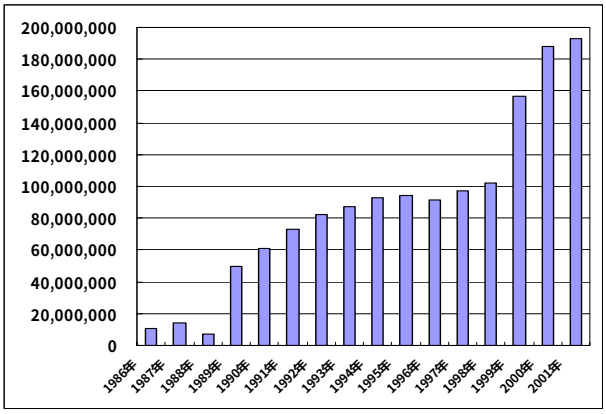

図１ 読売新聞記事データ文字数

2.2. 使用状況のとらえ方の問題

2.2.1. データ量増加の影響

読売新聞記事データを使用する際，使用状況をとらえる上で問題になる点の一つ目は，年により記事データの量が異なるという点である。このため本稿では，使用状況の指標として，使用度数，すなわち出現回数だけではなく，より使用状況を的確に表す数値として，使用頻度，すなわち使用度数を収録単語数で正規化した値に着目した。ただし，平均単語長は年によらずほぼ一定とみなせるので，本稿では収録単語数の代わりに，より算出の容易な収録文字数を用いて正規化することとした。

2.2.2. 用例の重複

[1] 外来語委員会の詳細や対象語の一覧は以下より参照可能。
http://www.kokken.go.jp/public/gairaigo/index.html

二つ目の問題点は，重複記事と，それにより抽出される重複用例の扱いである。度数を数える際に重複を排除して数えるべきかが問題になった。ここでいう重複記事とは主に次の要因によって生じているものを指す。問題にしたのはこのうちの①〜③（2.1.節参照）である。

① 本社発行と支社発行とがあるため。
② 複数の県版（地域版）があるため。
③ 見開きページで作られた記事をデータベースに登録する際に意図的に部分的に二重に採録するため。
④ １週間あるいは１か月のダイジェストとして再掲載するため。
⑤ 選挙広報，１０大ニュース募集，など，意図的に複数回掲載するため。

今回の対象語で重複例を調べたところ，多いところでは「ケア」で 23512 件中，236 件（約 1%），「シェア」で 13637 件中，91 件（約 0.6%）見つかった。この数字を大きいと見るか小さいと見るかの判断は難しい。しかしながら，もし重複を用例単位で除こうとすれば調査母体が対象語ごとに異なってしまう。記事単位で除こうとすれば調査の対象記事データがかなり限られてしまう。よって，結論としてこれくらいの数字であれば重複を無視した方が今回の我々の調査には適当であると判断した。

2.3. 対象語の同定の問題

しばしば指摘されているように外来語には表記のゆれのあることが多い。そこで考えられる表記のゆれを想定して検索することで対処した。

例：アイデンティティ 全 2462 件
「アイデンティティ」472 件
「アイデンティティー」1987 件
「アイデンティーティー」1 件
「アイデンティー」1 件
「アイデンテティ」1 件

また，記事データに対し形態素解析を施さずに，文字列マッチングのみによって用例を抽出したこともあり，部分的に文字列が同じである別語の用例が抽出されることが多々あった。これには人手によって排除することによって対処した。

例：スクリーニング 全 311 件
「ハウスクリーニング」
「ガラスクリーニング」など 71 件を排除

３．調査結果

対象 109 語について読売新聞記事データにおける各年の使用度数, 使用頻度, および使用度数の総計を出した。

16 年間の使用度数の総計が 1 万を超えたのは比較的定着に向かっていると思われる 6 語である。度数の総計とともに示すと、「ケア 23512，シェア 13637，ビジョン 11170，アクセス 10673，ベンチャー10558，ガイドライン 10251」である。（これ以降も必要に応じて度数の総計を語に添えて示すことにする。）半数以上の 65 語は使用度数の総計が 1000 にも満たなかった。そのうち 100 にも満たなかったものが 9 語 ある。「ストックヤード 99，フィルタリング 87，タスク 67，ログイン 29，ハーモナイゼーション 22, サマリー13, バックオフィス 8, エンフォースメント 5，トレーサビリティ 5」である。一般に分かりにくいと思われる外来語は，おおむね，新聞においてもその使用は多くない語であることを調査結果により確認することができた。

４．使用頻度の推移の分析

16 年間における使用頻度の推移を調べた。このとき，度数 50 未満の 6 語（前章参照）は初出が 1989 年以降の新語であり度数が少なすぎて推移の傾向をとらえ難いため別扱いにした。残り 103 語について使用頻度の増減傾向を大まかにとらえ, 表１に示すとおり, 大分類で４つ, 細分類で７つの型に分けた。以下，これらの型別に例を挙げて分析する。なお，分析に際しては，文化庁の協力によって「外来語」委員会が調査した「理解率」を適宜参照する。

表１：使用推移の型による分類

大分類	細分類	語数
増加型	右上がり型	44
平行型	平ら型	17
減少型	右下がり型	13
	山型	10
	山型の後平ら型	4
凹凸型	突出型（特定の年のみ増加）	10
	上記以外の型	5

4.1. 増加型

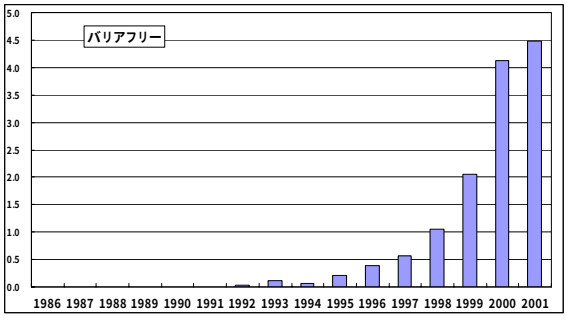

図２「バリアフリー」文字使用頻度(×10^{-5})

使用頻度が増加しているこのタイプにもっとも多くの

語が分類された。最多度数の「ケア 23512」から「ストックヤード 99」まで 44 語あった。このうち，きれいに右肩上がりに増加傾向を示す「バリアフリー3656」の例を図２に示す。

この増加型を示す語というのは，近年，その語の表す概念が取り上げられることが多くなってきたものと言えそうである。たとえば,「ケア 23512, デイサービス 4585, バリアフリー3656，ノーマライゼーション 277，ノンステップバス 217，セカンドオピニオン 205，ユニバーサルサービス 112」といった，福祉や医療に関わる語がこの増加型であった。今回の対象語 109 語のうち，福祉や医療に関する語でこの増加型に入らなかったのは「インフォームド・コンセント 1732」(山型)「メンタルヘルス 408, スクリーニング 311」(突出型) の 3 語のみである。「インフォームド・コンセント」は 1997 年までは増加傾向を示しており,「スクリーニング」は BSE 問題に絡み 2001 年に突出して使用が増えていることをあわせて考えると，ほとんどが増加傾向を示したことになる。また，インターネットの普及とともに使用が増えてきていると思われる「アクセス 10673，コンテンツ 1422，アーカイブ 227」などもこの増加型を示した。

使用が増えているということは定着に向かっていると見ることもできる。ただし，中には定着が不十分なうちに，その使用だけが増加している語もないわけではなさそうである。理解率を見ると，度数総計の多い「デイサービス, ケア, バリアフリー」といった語は, 順に 77.2%, 75.6%，72.7%，と高いのに対して，度数総計がさほど多くない「インフォームド・コンセント，ノーマライゼーション, コンテンツ, アーカイブ」は順に 23.2%, 12.2%, 23.0%，8.0%，と低い。なお,「アクセス」は総計が多い割に理解率は 57.7%とそう高くない。

4.2. 平行型

この型を示す語は全部で 17 語あるが，約半分の 8 語が度数総計 2,000 以上で, 残り半分の 9 語が度数総計 100～1300 であり，それぞれで異なる特徴をもつ。

前者に該当するのは，図３に示す「シェア 13637」の他,「ビジョン 11170，バックアップ 4154，マーケティング 2748，アイデンティティ 2462」などである。

これらは早くからある程度定着傾向を見せており，毎年同じように使用されてきていると見られる。おおむね理解率の調査でも高い数値がでている。ただし，この中では「アイデンティティ」のみ理解率が 20.9%と低い。この語は使用頻度の総計や推移の型では定着傾向を示しつつも，語の表す概念の難しさによって，実際のところは定着し難い語のようである。

後者に該当するのは, 図４に示す「コア 1034」のほか,「ライブラリー1290，モータリゼーション 227，デリバリー168，フェローシップ 139」などである。理解率は高くはない。これらは目立って使われることはあまりないが，かといって，使われなくなる様子も見られない語と言えそうである。

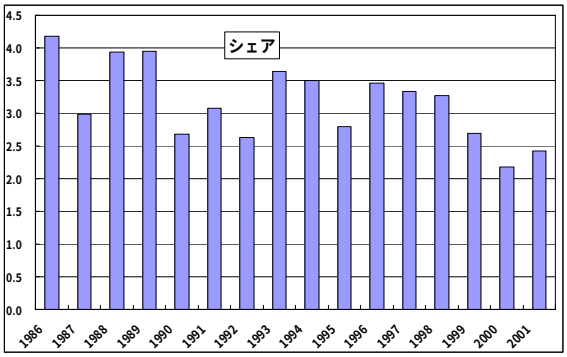

図３「シェア」使用頻度(×10-5)

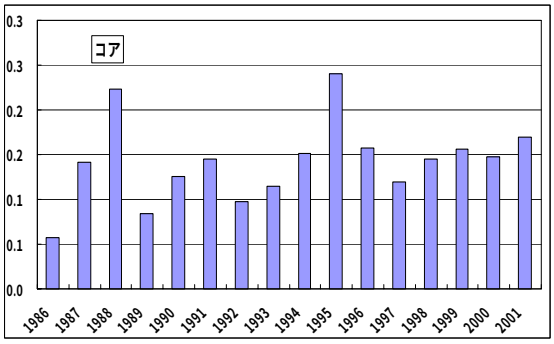

図４「コア」使用頻度(×10-5)

4.3. 減少型

一般には外来語の使用は増えているという印象が強いが，今回の対象語の中には減少傾向を示す語があった。この 16 年の早いうちより減少しているものが「アメニティ 576，フレックスタイム 492，サーベイランス 408」など 13 語，一度増えていき，その後減少に転じているものが「ガイドライン 10251，マルチメディア 6739，フィルタリング 87」など 10 語，一度増えて減った後に平らに推移しているものが「アナリスト 2756，トレンド 1234」など 4 語あった。図５～図７に一例ずつ示す。

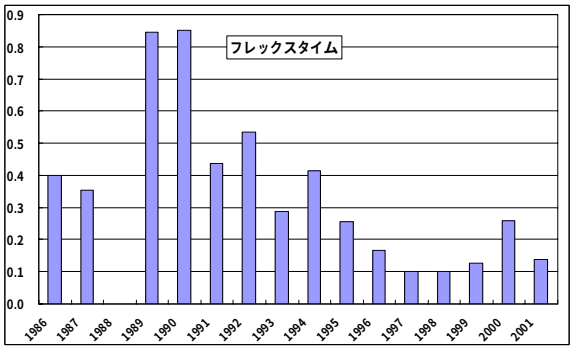

図５「フレックスタイム」使用頻度(×10-5)

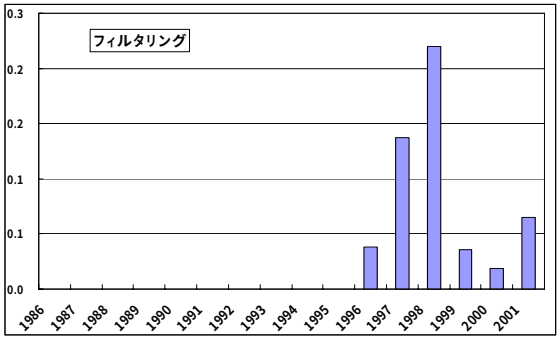

図6「フィルタリング」使用頻度(×10-5)

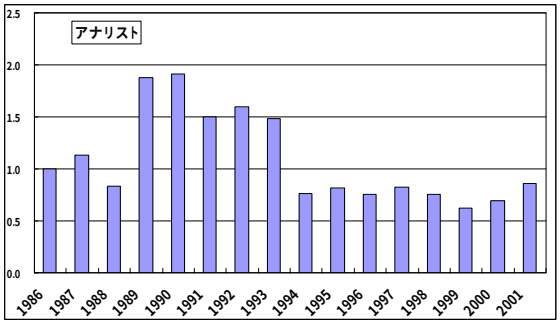

図7「アナリスト」使用頻度(×10-5)

これらは，たとえば社会的に認知されるようになるきっかけの政策や運動があり，一時期新聞を賑わせたが，その後話題に上ることが徐々に少なくなり減少していると考えられる[2]。このまま定着せずに消えていくのか，あるいは，推移の型が増減後に平らになっているものに見られるように，ある程度の使用頻度のまま推移していくのか，語によって分かれてくるだろう。なお，今回ここに分類された語は，総計が 2000 以上ある語も含めて，ほとんどが理解率 50%に満たず，理解率から見ても現時点ではまだ定着には至っていないと言える。

4.4. 凹凸型

先の減少型とは異なり，推移に連続した増減があるのではなく，使用頻度が他と比べて非常に突出している年のあるものがある。細分類ではその突出がより顕著なものを突出型としてそれ以外のものと分けた。「ダンピング 3691，インサイダー1546，ライフライン 1485，メンタルヘルス 408」など 10 語を突出型に，「コミットメント 508，ケーススタディ 183」など 6 語をそれ以外に分類した。各例を図８，９に示す。ある特定の年に使用頻度が多くなるのは，その年に盛んに取り上げられた話題や政策に関わる語であったためと見られる。図８に示す「ライフライン」は阪神・淡路大震災のあった 1995 年の使用頻度が突出している。

ここに分類された語は「ダンピング」のみ度数総計が 2000 を超えているだけで，総じて度数総計が少ない。また，理解率を見ても最も高いのが「ライフライン」の 51.8%であり，いずれも低いものばかりである。減少型以上に，凹凸型に分類された語は定着を示していない。

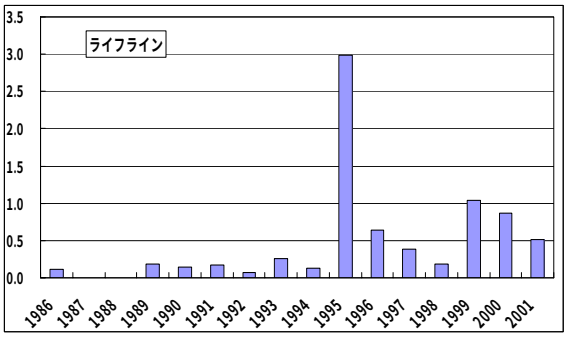

図８「ライフライン」使用頻度(×10-5)

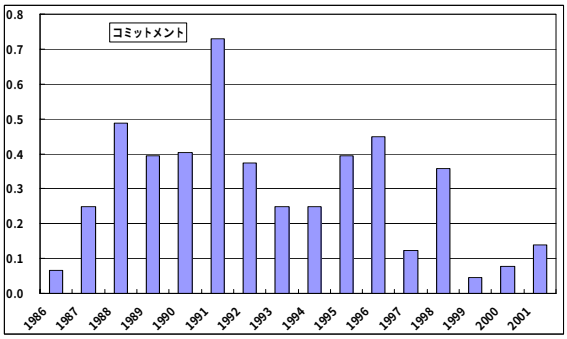

図９「コミットメント」使用頻度(×10-5)

5．まとめ

対象とした 109 語の外来語について 16 年間の新聞記事を用いて使用頻度の経年調査をすることにより，新聞記事における外来語の盛衰の実態を定量的に明らかにすることができた。全体的には使用頻度の推移が増加か平らであるものが多い一方で，使用頻度が特定の事柄に左右されて増減するものや，使用が減少し短命な傾向を示すものも見られた。各外来語が実際にどれくらい定着し，あるいは廃れていくかは，今後，新聞記事はもとより，他のメディアの使用状況も調べていく必要があろう。国立国語研究所では広報紙や雑誌などとその調査範囲を広げている。また，理解率調査のような社会調査データの活用もさらに進めていく予定でいる。

謝辞

読売新聞記事データの使用を許諾してくださった，外来語委員の関根健一氏をはじめとする読売新聞社のみなさま，および，調査を手伝ってくださった西部みちるさん，ほかみなさまに感謝いたします。

[2] 「アメニティ」の減少については関根健一「新聞記事の中のカタカナ語」（『日本語学』22-8、2003 年 7 月）にも言及がある。