

第4章 新聞記事における語彙の時間的変化分析

——語種との関係を中心に——

山口 昌也

1 はじめに

本研究の目的は、新聞記事を対象に、語の出現頻度に関する時間的変化を実態調査し、語種との関係から変化の特徴を明らかにすることである。

これまで、国語研究所の語彙調査（例：現代雑誌九十種調査（国立国語研究所（1964））をはじめとして、各種の語彙調査や言語研究（橋本（2003）、佐竹（2002））において語の出現頻度が調査されてきたが、その多くは、時間的に単一の調査であったり、10年以上の間隔をおいた調査であり、出現頻度の時間的変化の実態やその要因については、明らかにされてこなかった。しかし、出現頻度の時間的変化を分析することは、語彙における重要性の判別や、統計情報に基づいた自然言語処理システムの構築にとって必要であると考えられる。

そこで、本研究では、語の出現頻度に関する時間的変化の特徴分析の第一段階として、語種に着目し、語種比率、および、語ごとの出現頻度の時間的変化に関して実態調査した結果を報告する。分析対象として、語種を選択したのは、出現頻度の変化は、語としての定着の度合に影響されると考えられるが、語種は定着の度合を知る一つの指標になると考えたからである。

分析対象は、(1)広範な読者が存在し、社会的影響が大きいこと、(2)内容が幅広く、豊富な語彙を含むことから、毎日新聞9年分(1994～2002年)とした。この資料に対して、形態素解析、および、語種付与を行い、分析用のデータとする。なお、本論文における「形態素」とは、後述の形態素解析システム『茶釜』の辞書見出しを基準とする。

これらのデータを対象に、(a)語種比率の時間的変化、(b)形態素の出現頻度の時間的変化と語種との関係进行分析する。(a)については、各年ごとに語種(和語、漢語、外来語、混種語)の比率を実態調査する。さらに、語種比率の時間的変化の有無を調査、分析する。

一方、(b)については、二つの観点から分析する。一つは、形態素ごとに出現頻度の時間的変化の大きさを測定し、語種との関係进行分析するものである。時間的変化の指標としては、出現頻度の変動係数を用いる。もう一つは、使用語彙としての時間的変化と語種との関係进行分析するものである。具体的には、年ごとの使用語彙を比較することにより、共通語彙、語彙の差分(新出する形態素、および、使われなくなった形態素)の実態を調査し、語種との関係を考察する。

この後、2節では、分析対象の資料に対して行った形態素解析、語種付与について説明する。3節では、各種の分析とその結果を示す。最後に、4節で本論文のまとめを行う。

2 分析対象のデータ

1節で述べたように、本研究の分析対象は、毎日新聞9年分(1994～2002年)とした。本研究の分析で使用した資料は、「CD-毎日新聞データ集」(日外アソシエーツ)に収録されている記事全文である。調査の時間的間隔は、1年とした。

この資料に対して、形態素解析を行い、各形態素に語種を付与した。本節では、その方法・結果について説明する。

2.1 形態素解析と語種付与

形態素解析には、奈良先端科学技術大学院大学松本研究室で開発している形態素解析システム『茶釜』(ChaSen ver. 2.2.9, IPADIC ver. 2.4.4(松本ら(2002)))を使った。解析時の『茶釜』の設定は、デフォルト設定から次の変更を行っている。

- ・品詞「記号」列は、品詞「記号」の1形態素とする
- ・品詞「名詞-数」列は、品詞「名詞-数」の1形態素とする

語種付与には、語種辞書『かたりぐさ』¹を用いた。『かたりぐさ』は、IPADIC ver. 2.4.4の各見出しに対する語種情報を収録している。付与した語種は、和語、漢語、外来語、混種語の4種類である。

ただし、固有名詞（『茶釜』の品詞では、「名詞-固有名詞」）、数詞（『茶釜』の品詞では、「名詞-数」）、記号、未知語（『茶釜』では、辞書に登録されていない文字列を形態素として切り出した場合、その形態素に「未知語」という品詞を付与する）については、語種付与は行っていない。

なお、「年月」（「トシツキ」[和語]/「ネンゲツ」[漢語]）のように、複数の語種の可能性を持つ語については、一意に決定することが困難な場合があるため、本論文においては、「その他」の語種として表記し、分析対象とはしない。さらに、助詞、助動詞は語種付与対象としているが、すべての語が和語であり、出現頻度も大きいので、分析の対象から除外した。

延べ形態素数を年ごとにまとめた結果を図1²に示す。縦軸は、延べ形態素数、横軸は年である。全年・全形態素中における、和語、漢語、外来語、混種語の比率は、合計で41.9%だった。これらが、本論文における分析対象の形態素となる。なお、助詞、助動詞をあわせると、68.1%である。また、複数の語種の可能性がある「その他」となったのは、1.3%だった。

異なり形態素数を年ごとにまとめた結果を図2に示す。縦軸は、異なり形態素数、横軸は年である。全年・全形態素での異なり語数も同時に示した（横軸が「合計」の部分）。分析対象となる和語、漢語、外来語、混種語の比率の合計は、年あたりの平均で30.1%、全年・全形態素では、13.1%だった。全年・全形態素で、分析対象4語種の異なり語数の比率が少ないのは、数詞、未知語の異なり語数が大きくなるためである。

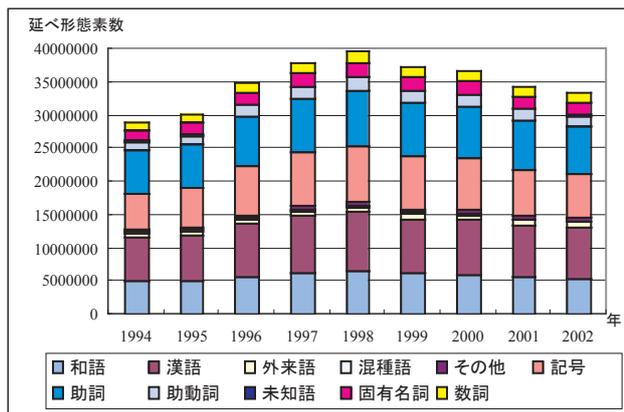


図1 延べ形態素数

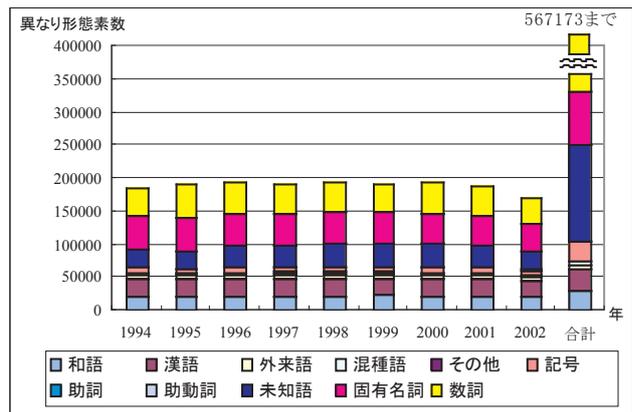


図2 異なり形態素数

2.2 形態素解析の問題点と分析への影響

本論文では、形態素解析に『茶釜』を利用しているが、語彙調査を行う場合には、(a)解析誤り、(b)未知語、(c)形態素の辞書の問題がある。本節では、分析に先だって、これらの問題が分析に与える影響について考察しておく。

まず、解析誤りの問題に対しては、解析精度を見積もることにより対処する。解析精度の見積りは、毎日新聞2002年1月分（2642740形態素）から1万形態素をランダムサンプリングして、人手でチェックすることにより行った。その結果、精度は約98%となった。誤りの多く（約64%）は、固有名詞に対するもので、『茶釜』の辞書に登録されていないことに起因していた。品詞が未知語として判定される場合もあるが、多くは、既存の見出しで過分割される傾向にあった。したがって、解析誤りの多くは、延べ形態素数、および、異なり形態素数を増加させるように影響すると考えられる。

次に、未知語の問題である。全年・全形態素に対して、品詞として「未知語」と分析された形態素の比率は、年あたりの平均で、延べ形態素数が0.45%、異なり形態素数が17.1%である。延べ形態素数に与える影響は小さく、また、未知語は解析誤りに含まれるため、上で見積もった解析精度を分析の際の判断材料とすることができる。

一方、異なり形態素数に与える影響は、17.1%と大きいですが、解析精度の見積もりからは直接判断できな

い。そこで、2002年の未知語（異なり29458、延べ175643）の異なりから200形態素をランダムサンプリングし、未知語に占める語種の内訳を見積もった。その結果、和語、漢語、外来語、混種語の形態素数は、それぞれ、21(10.5%)、4(2.0%)、31(15.5%)、3(1.5%)形態素だった（括弧内は、未知語全体に占める割合）。残りは、固有名詞140、既存形態素の分割誤り1形態素である。この結果から、和語、漢語、外来語、混種語の異なり語数に占める割合は、それぞれ 1.8、0.34、2.7、0.26%加算したもの（年あたりの平均）になると推測される。

形態素の辞書には、(i)異表記の統一を含め、同語異語判別をしていない、(ii)単位が不統一である、といった問題がある。(ii)は、形態素を含んだ形態素が辞書に登録されている問題である。例えば、「日本人」は「日本人」全体で辞書に登録されているため、1形態素と解析される。一方、「韓国人」は、1形態素では登録されておらず、「韓国」と「人」とが別々に登録されているため、2形態素に分割される。

本論文の分析では、これらの問題については、特別な処置をしておらず、分析に与える影響はわからない。ただし、(iii)については、形態素という観点からすると、分割されたほうが好ましいため、本来解析されるべき形態素数に対して、少なくなるほうに解析されると予想される。

3 時間的変化の分析

3.1 語種比率の時間的変化

3.1.1 分析方法、および、測定結果

本節では、語種比率の時間的変化について分析する。この分析の目的は、語種によって、出現頻度の変化に違いがあるのか、また、特定の変化（例えば、漢語の比率が増大しているなど）が起きているのかを知ることである。

調査対象の語種全体に占める各語種の延べ形態素数、異なり形態素数の百分率を、それぞれ、表1、2に示す。「合計」欄は、全年・分析対象4語種での測定結果である。

表1 語種比率の変化（延べ形態素数）[%]

	1994	1995	1996	1997	1998
和語	39.67	38.99	39.13	39.08	39.35
漢語	53.87	54.69	54.27	54.41	54.13
外来語	4.95	4.78	5.06	4.98	5.01
混種語	1.50	1.54	1.54	1.52	1.50
	1999	2000	2001	2002	合計
和語	39.64	39.84	40.01	38.6	39.37
漢語	53.86	53.51	53.26	54.85	54.09
外来語	4.99	5.14	5.22	5.10	5.03
混種語	1.51	1.52	1.51	1.45	1.51

表2 語種比率の変化（異なり形態素数）[%]

	1994	1995	1996	1997	1998
和語	38.67	38.54	38.7	38.68	38.56
漢語	45.29	45.61	45.48	45.5	45.58
外来語	9.60	9.46	9.46	9.39	9.49
混種語	6.45	6.39	6.35	6.43	6.37
未知語	(15.68)	(14.74)	(16.84)	(17.79)	(18.56)
	1999	2000	2001	2002	合計
和語	38.86	38.6	39.03	38.14	39.81
漢語	45.35	45.6	45.22	45.89	44.46
外来語	9.41	9.45	9.37	9.71	8.55
混種語	6.39	6.35	6.38	6.25	7.19
未知語	(18.06)	(17.73)	(17.63)	(17.27)	---

3.1.2 分析

表1より、延べ形態素数の語種比率の時間的変化は、どの語種についても小さいことがわかる。語種比率の最大値と最小値との差は、和語、漢語、外来語、混種語、それぞれ、1.4, 1.6, 0.4, 0.1%であった。また、語種比率の変化に特定の傾向は見られない。

次に、表2を見ると、異なり形態素数も延べ形態素数と同様、時間的な変化は小さく、また、特定の変化傾向は見られないことがわかる。ただし、2.2節で述べたように、未知語の影響を考慮する必要がある。具体的には、未知語が全形態素に占める比率の変化と未知語を構成する語種の比率である。前者については、表2の「未知語」欄に示す（未知語欄だけ、全形態素数に対する比率であることに注意）。この表2と図1より、未知語の異なり形態素数は、延べ形態素数と相関が高い（相関係数0.88）ことがわかる。また、2.2節で示したように、後者の未知語を構成する語種の比率は、和語、外来語の比率が高いことがわかっている。したがって、和語と外来語、特に未知語に比べて語種比率の小さい外来語は、延べ形態素数の変化に伴う影響を多く受け、それが外来語の異なり形態素数の比率に反映されると考えられる。

3.2 出現頻度の変動係数の分布

3.2.1 分析方法

本節では、出現頻度の変化が語種ごとに異なるかどうかを分析する。その方法として、各形態素に対して、9年分の出現頻度の変動係数を求め、その分布を調査する。変動係数 c_x は、 $c_x = \sqrt{v_x}/\bar{x}$ で表される。なお、 \bar{x} は出現頻度の平均、 v_x は出現頻度の分散である。変動係数は、ばらつきの程度を表す指標であり、変動係数を使うことにより、平均値の異なるデータの分散を相対的に比較することができる。なお、変動係数の計測にあたっては、低頻度の形態素の影響を排除するために、全年の総出現頻度が18以上³の形態素を対象とした。

3.2.2 変動係数の分布の測定

ここでは、全年の総出現頻度数によって、次に示す四つの集合にわけて分布を計測した。「集合1」は、語種別の変動係数を把握することを目的とする。「集合2～4」の三つの集合は、それぞれ低頻度、中頻度、高頻度の形態素における変動係数の分布を分析することを狙って、設定した。

- 集合1 : 18 以上
- 集合2 : 18 以上 100 未満
- 集合3 : 100 以上 1000 未満
- 集合4 : 1000 以上

集合1の各語に対して変動係数を求め、語種ごとにその分布をプロットした結果を図3に示す。同様に、集合2～4に対する結果を図4～6に示す。横軸は変動係数で、各階級の区間を0.1としている。縦軸は、語種別の全度数に占める割合である。

さらに、各語種、集合ごとに変動係数の平均値と分散を求めた結果を表3に示す。最左欄は、集合番号を表す。また、表3の「個数」欄には、個々の集合に含まれる異なり形態素数を示す。

3.2.3 変動係数の分布の分析（全体的傾向）

まず、変動係数の分布に関する全体的傾向を把握するために、集合1の測定結果について考察を加える。考察を加える前に、変動係数の値を使うと、形態素のどのような性質を捉えられるのかを確認しよう。表4、5は、語種別に、変動係数の上位5位、下位5位の形態素を列挙したものである。

表4で変動係数上位5位の形態素を見ると、特定の期間にしか出現しなかった形態素が上位を占めている。例えば、表4の漢語1位の「鼻音」は言語学に関連する用語であり、1995年の一つの記事にだけ出現した形態素である。一方、表5を見ると、出現頻度の時間的変化が小さく、表4中の形態素よりも、汎用性の高い形態素が挙げられていることがわかる。

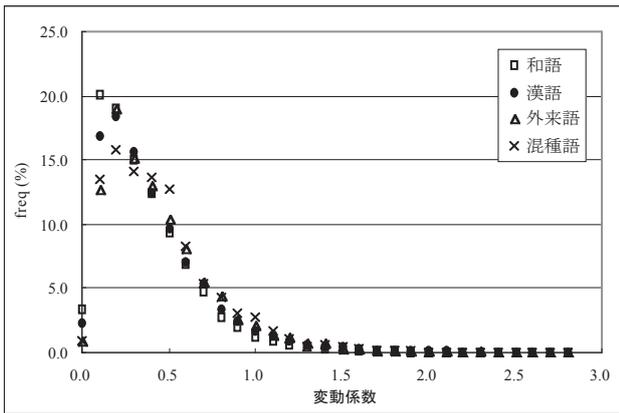


図3 出現頻度の変動係数 (出現頻度18以上)

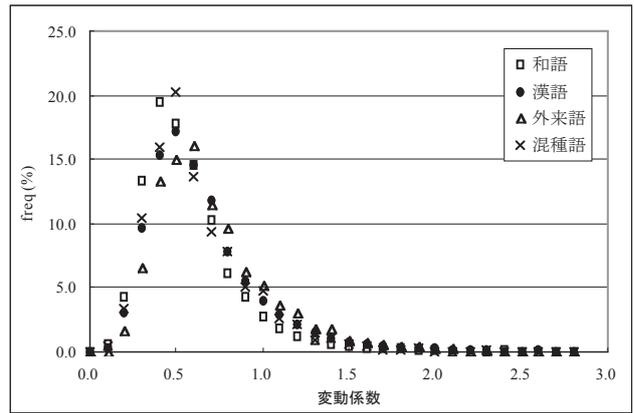


図4 出現頻度の変動係数 (出現頻度18以上100未満)

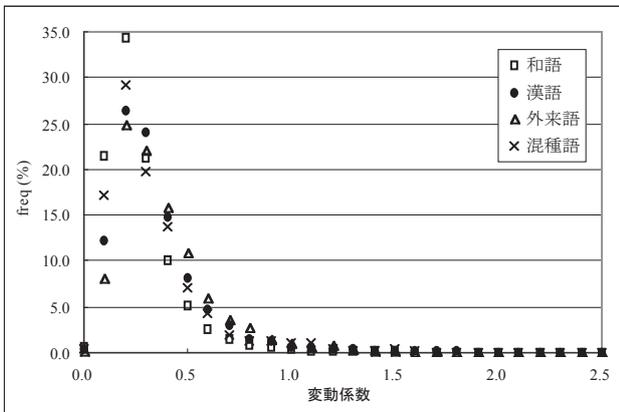


図5 出現頻度の変動係数 (出現頻度100以上1000未満)

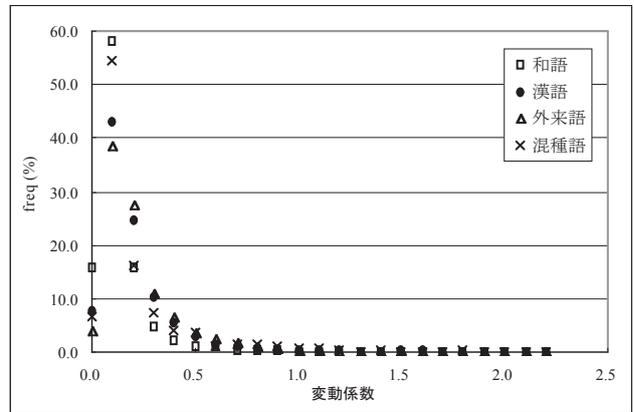


図6 出現頻度の変動係数 (出現頻度1000以上)

次に、語種別の特徴を見てみよう。図3を見ると、外来語と混種語は、和語・漢語に比べ、変動係数0～0.1の出現割合（縦軸）が低く、0.4以上の出現割合が高いことがわかる。例えば、変動係数0.1では、外来語と混種語の出現割合は13%付近なのに対して、漢語が約17%，和語は20%である。一方、変動係数が0.4になると、両者の関係が入れ替わり、外来語(13%)と混種語(14%)の出現割合のほうが、和語(12%)・漢語(12%)よりも高くなる。

以上のようなグラフ上の特徴から、外来語・混種語は、出現頻度の変動係数の大きい形態素の割合が和語・漢語よりも大きいことがわかる。また、変動係数の平均値（表3（集合1））で見ても、外来語・混種語のほうが、和語・漢語よりも大きい。具体的には、和語・漢語の変動係数の平均値がそれぞれ0.420, 0.463なのに対して、外来語・混種語はそれぞれ0.494, 0.505となっている。

さらに、変動係数の分散は、和語が 8.49×10^{-2} と最も低く、その他の語種は、 $10.4 \times 10^{-2} \sim 10.8 \times 10^{-2}$ である。したがって、和語は、平均的に見て、他の語種よりも出現頻度の時間的変化のばらつきが少ないことがわかる。

表3 変動係数の平均値, および, 分散

集合	和語			漢語			外来語			混種語			全体		
	平均	分散	個数	平均	分散	個数	平均	分散	個数	平均	分散	個数	平均	分散	個数
1	4.20	8.49	18018	4.63	10.8	22736	4.94	10.7	5013	5.05	10.4	2856	4.51	9.99	49829
2	6.29	8.55	7150	7.07	11.6	7326	7.63	12.0	1416	6.75	9.58	1321	6.77	10.4	17519
3	3.26	3.90	7509	4.05	6.59	9097	4.37	6.30	2460	3.82	6.30	1224	3.79	5.71	20784
4	1.84	1.81	3359	2.61	4.52	6313	2.87	4.70	1137	2.61	5.47	311	2.40	3.86	11526

平均: $\times 10^{-1}$, 分散: $\times 10^{-2}$

表4 語種別の変動係数（上位5形態素）

順位	和語		漢語		外来語		混種語	
1	撞く	2.83	鼻音	2.83	シンドローム	2.48	浜納豆	2.66
2	下仕え	2.72	殯宮	2.83	チームスピリット	2.37	ブレンド米	2.46
3	方違え	2.68	着帯	2.83	スペースラブ	2.36	番茶も出花	2.35
4	日掛け	2.65	骨炭	2.83	リターンキー	2.33	猟場	2.20
5	麦笛	2.48	血粉	2.83	クーロン	2.31	こま犬	2.15

表5 語種別の変動係数（下位5形態素）

順位	和語		漢語		外来語		混種語	
1	眠れる	0.0395	一度	0.0457	スローガン	0.0473	気負い	0.0602
2	贈る	0.0395	上手	0.0462	パフォーマンス	0.0535	場所	0.0611
3	中入	0.0414	気	0.0480	テレビ	0.0574	反する	0.0691
4	送れる	0.0418	満員	0.0488	タイトル	0.0594	半年	0.0714
5	間もなく	0.0418	語	0.0493	リズム	0.0603	気の毒	0.0754

3.2.4 変動係数の分布の分析（集合別の分析）

次に、出現度数の程度ごとに変動係数の分布を分析するために、集合2, 3, 4を個別に見ていくことにする。

まず、語種の観点を入れずに、変動係数の全体的な傾向を見てみよう。表3の「全体」の項目を見ると、変動係数の平均値は、集合2, 3, 4の順に下がっていく。これは、出現頻度が高い形態素の集合ほど、変動係数が小さいこと、つまり、出現頻度の変動が小さくなることを意味している。さらに、分散も同様の傾向にあり、出現頻度が高い形態素の集合ほど、形態素ごとの変動のばらつき自体も小さくなることわかる。このことは、図4～6の分布の山の最大値が左にずれていくとともに、山の幅が狭く（分散が小さく）なっていることからわかる。

さらに、語種別に見てみよう。表3からは、次のことが推測できる。

- ・表3を見ると、変動係数の平均値、分散ともに、どの語種についても、出現頻度が大きい語の集合ほど、小さくなる。例えば、外来語の平均値を見ると、集合2, 3, 4の順に、0.763, 0.437, 0.287となる。これは、前節で述べた全体的な傾向と同一である。
- ・変動係数の平均値を語種別に見ると、どの集合でも、外来語が最も大きく、外来語／漢語／混種語／和語の順となる。このことは、外来語の出現頻度は時間的変化の観点から変動が大きく、逆に和語は小さいことを示している。
- ・変動係数の分散を語種別に見ると、どの集合でも、和語が最も小さいことが顕著である。特に、集合4（高頻度の形態素の集合）で他の語種との差が大きく、他の語種の分散が $4.52 \times 10^{-2} \sim 5.47 \times 10^{-2}$ なのに対して、和語の分散が 1.81×10^{-2} となっている。このことは、変動係数の分布の型にも現れており、集合4（図6）の最頻値（モード）は、外来語58.2%、和語38.7%と、その差は約20%にもなる。以上の事柄は、高頻度の和語は時間的変動のばらつきが小さく、安定して使われていることを示している。

3.3 使用語彙の時間的変化

この節では、語種と使用語彙の時間的変化との関係を把握するために、各年の使用語彙の共通部分と差分を測定した結果を示す。表6は、1994年と1995～2002年の各年との共通部分・差分を示したものである。なお、表中の「差1」とは、1994年側だけに出現した形態素数、「差2」は比較対象の年側だけに出現した形態素数である。さらに、表6から、各年の使用語彙における差分（差1, 2の合計）の割合を求めた結果が、表7である。

まず、語種別に、語彙の共通部分に対する差分の割合を見てみよう。図7を見ると、混種語が0.27～0.29と最も大きく、外来語が0.13～0.15で最も小さい。この要因として考えられることは、生起確率の低さから、1年に1度も出現しない語が増加することである。実際、調査期間の9年間に1度も出現しない年のある語の割合は、次のようになる。この結果の割合における語種別順位は、図7の結果と一致する。

和語：0.28，漢語：0.24，外来語：0.20，混種語：0.36

次に、割合の変化について考察する。図7を見ると、どの語種についても、最大値と最小値の差は、0.01～0.02程度であり、大きな変化は認められない。これは、表6を見ても、同様である。ただし、この結果には、未知語が含まれていない。2.2節に示したように、異なり形態素数に占める未知語の割合は17.1%（年あたりの平均）であり、その影響は大きい。『茶筌』が認定した未知語については、分割誤りもあり、それ自体が形態素とは限らないので、さらなる調査が必要であるが、未知語の部分で語彙の変化が起きている可能性はある。その変化の内容についても、語彙量は大きく変化せず、「新陳代謝」が起きていることが推測される。なぜならば、(1)特に外来語については、未知語中に15.5%含まれると推定されている（2.2節参照のこと）にもかかわらず、表6の差1，2に極端な増加がないこと、(2)図2の「合計」のとおり、全年での異なり形態素数は増加しているからである。

表6 使用語彙の共通部分と差分(共通，差1，差2)

	和語	漢語	外来語	混種語	合計
1994	22251, 0, 0	26060, 0, 0	5524, 0, 0	3710, 0, 0	57545, 0, 0
94/95	19411, 2840, 2431	23420, 2640, 2430	5077, 447, 282	3086, 624, 534	50994, 6551, 5677
94/96	19677, 2574, 2579	23591, 2469, 2564	5084, 440, 358	3115, 595, 537	51467, 6078, 6038
94/97	19703, 2548, 2632	23621, 2439, 2652	5075, 449, 348	3121, 589, 590	51520, 6025, 6222
94/98	19667, 2584, 2606	23650, 2410, 2672	5114, 410, 366	3125, 585, 555	51556, 5989, 6199
94/99	19669, 2582, 2725	23527, 2533, 2605	5069, 455, 351	3097, 613, 583	51362, 6183, 6264
94/00	19526, 2725, 2590	23532, 2528, 2598	5065, 459, 352	3077, 633, 562	51200, 6345, 6102
94/01	19497, 2754, 2612	23230, 2830, 2386	4989, 535, 322	3067, 643, 550	50783, 6762, 5870
94/02	18631, 3620, 1881	22692, 3368, 1987	4940, 584, 282	2927, 783, 436	49190, 8355, 4586

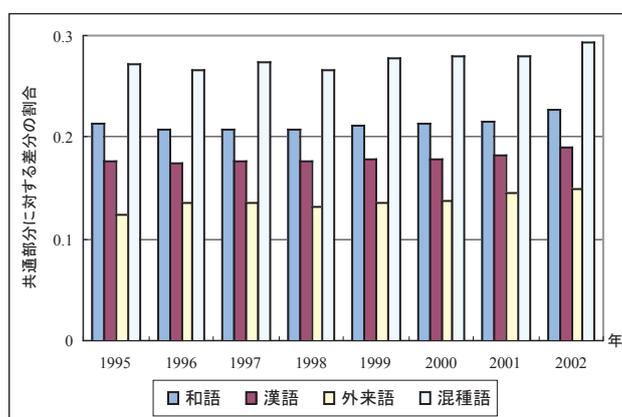


図7 使用語彙における差分の割合

4 おわりに

本論文では、新聞を対象に、出現頻度の時間的変化と語種との関係を分析し、次のことを明らかにした。

- ・語種比率の時間的変化に関して、延べ形態素数，異なり形態素数，ともに大きな変化は認められなかった。
- ・出現頻度の変動係数
 - どの語種においても，変動係数の平均，分散は，高頻度の場合ほど小さくなる。
 - 変動係数の平均値は，集合2,3,4（低・中・高頻度）のどの場合でも，外来語／漢語／混種語／和語（降順）となった。
 - 変動係数の分散は，集合2,3,4（低・中・高頻度）のどの場合でも，和語が最も小さい。特に，集合4（高頻度）で他の語種との差が大きく，他の語種の分散が $4.52 \times 10^{-2} \sim 5.47 \times 10^{-2}$ なのに対して，和語の分散が 1.81×10^{-2} となった。

・使用語彙の時間的変化

-1994年と1995～2002年の使用語彙を比較したところ、変化は認められなかった。

-ただし、未知語と解析された部分で語彙が変化している可能性があり、かつ、語彙量自体は変化していないことを推論した。

なお、本論文は第13回社会言語科学会での予稿資料（山口ら（2004））に加筆、修正したものである。

注

- 1 『かたりぐさ』については、本報告書第3部第8章「語種辞書『かたりぐさ』の開発」を参照のこと。また、『かたりぐさ』自体は、国立国語研究所のWebページで一般に無償公開されている。
- 2 帯グラフの下から、「和語」、「漢語」と続く。
- 3 調査期間の9年間、年平均で2回出現するということを基準に、全年の総出現頻度が18以上とする条件を設定した。

参考文献

- 国立国語研究所(1964)『現代雑誌九十種の用字用語 {I} 総記および語彙表』秀英出版
橋本和佳(2003)「戦後の新聞社説に見る外来語の増加」国語学会2003年度秋季大会予稿集
佐竹秀雄(2002)「新聞の生活家庭面における外来語」『日本語学と言語学』199-207, 明治書院
山田俊雄ら(1995)『現代語・古語 新潮国語辞典第二版』新潮社
山田俊雄ら(2000)『新潮 現代国語辞典第二版』新潮社
小学館(2002)『日本国語大辞典[第二版]』小学館
松本裕治ら(2002)『形態素解析システム『茶釜』version 2.2.9 使用説明書』
日外アソシエーツ(1995-2003)「CD-毎日新聞データ集」日外アソシエーツ
山口昌也, 茂木俊伸, 桐生りか, 田中牧郎(2004)「語種との関係に基づいた新聞記事における語彙の時間的変化分析」
第13回社会言語科学会研究大会予稿集, 113-116

謝辞

分析対象のデータの形態素解析にあたっては、形態素解析システム『茶釜』を利用させていただいた。開発者の方々に深く感謝する。