

# 畳語の分布調査

中渡瀬 秀一<sup>†</sup> 大山 敬三<sup>†</sup>

<sup>†</sup> 国立情報学研究所

## はじめに

本稿は前報である畳語の出現分布に関する調査 [1] の続報である。前報までに畳語の出現頻度が変動する際の傾向を分析することを目的に、新聞記事中の平仮名 4 字畳語についての調査を行った。本稿では前回までの調査に加えて新たに片仮名畳語の追加調査を行ったので、これらの調査結果をまとめて頻度分布に関する分析を行うことを目的とする。

近年、語の出現頻度変化の側面から言語変化を論じる研究が行われている [2]。この中では語彙の頻度分布の平衡性（安定性）と個々の語の頻度変化の安定性が問題として取り上げられた。この研究では Twitter の発言を資料に 1 年弱の期間に渡り、語（形態素）の出現頻度変化を調査した。その結果、語彙の頻度分布形状は安定しており、個々の語については高頻度語の方が頻度順位変化が少ないとの結果を得ている。

本稿では、より長期の経年変化を経た頻度分布に対する上のような統計的性質の有無を確認する。そのため比較する期間は 10 年以上を対象とする。また荒牧らの調査では語彙を形態素解析器の出力する形態素で代用していたのに対し、我々は正確に語であるものだけを調査の対象とする。その代わりに語彙を小規模に限定する。この際に語彙として用いるのが畳語である。畳語とは「われわれ」のように同じパタン（「われ」）を 2 回反復して造られ、擬態語・擬音語に多く用いられる語である。

畳語を対象とする利点としてはその名前の範囲が狭いことが挙げられる。例えば平仮名 4 字畳語の場合、語基は 2 文字であるため最大でも平仮名の種類の 2 乗である数千種類の語しか造語することができない<sup>1</sup>。このため正確な調査を行うのに必要な人手検査の負担を大幅に軽減することが可能となる。畳語は名前空間の大きさが制限されるため語彙は小規模であるが、しかしその使用範囲が狭いわけではない。畳語が含まれる品詞の範囲は広く、名詞・代名詞・動詞・形容詞・副

詞・感動詞など多岐にわたる。機能語になる品詞以外はほぼカバーしているといえる。それに加え、小規模であるがゆえに全名詞のような巨大な語彙と異なり、全体の頻度分布やその経年変化の概観・俯瞰が容易になることも期待される。

## 調査内容と方法

本調査の内容について述べる。調査資料には新聞記事（毎日新聞記事データの 1992 年と 2006 年分）を用いた。前回の調査で任意の長さの平仮名畳語について出現数をカウントしたところ全体の 95 % 以上が 4 字畳語であったことから、4 字畳語を対象に調査を行った。今回も同様に 4 字片仮名畳語を対象にして、次の 2 項目について調査した。ひとつは片仮名 4 字畳語の出現頻度分布で、もうひとつはその出現頻度分布の経年変化である。

### 片仮名 4 字畳語出現頻度分布の調査

この調査では 1992 年、2006 年の記事に含まれる全片仮名 4 字畳語の出現頻度分布について調査した。手順は、まず新聞記事の中から片仮名 2 文字パタンの反復となっている文字列を全て機械的に抽出する。しかしこの段階の文字列にはまだ畳語でないものが含まれているので、次にこれを人手でクレンジングする。畳語にならない片仮名反復型の文字列は次のようなものである。

（不適切な 2 文字反復パタンの例）

「グアジャジャラ族（ブラジルの先住民集団）」

「エスエス製薬（Science & Society の頭文字から）」

最後にクレンジング後の文字列を畳語としてカウントし、各畳語の出現頻度、頻度上位からの累積頻度を計算した。

<sup>1</sup> 実際には音韻的制約でさらに名前空間は小さい

## 出現頻度分布の経年変化の調査

この調査では片仮名 4 字畳語の出現頻度分布の経年変化を見るために 1992 年の分布と 2006 年の分布とを比較した．比較には順位の相関係数<sup>2</sup>を用いた．これによって両年の畳語出現順位の類似度を指標化する．相関係数は頻度順位上位  $N$  件の畳語について計算することで高頻度語と低頻度語の違いを比較する．

## 調査結果

各調査の結果について述べる．

### 片仮名 4 字畳語出現頻度分布の調査

1992 年の片仮名 4 字畳語出現頻度の上位 10 語を表 1 に示す．またそれらの頻度分布とその上位からの累積頻度のグラフを図 1 示す．この調査の結果，出現した畳語は 318 種類であった．そのうち頻度が上位の 30 % で全体の 80 % 以上，上位 50 % で約 90 % をカバーしていることが確認された（2006 年についても概ね同様である）．また語彙の頻度分布を近似するジップ則の冪指数を算出したところおよそ  $-1.12$  (決定係数  $R^2 = 0.95$ ) であった．したがって順位と頻度の関係は反比例に近く，累積頻度は対数関数で近似される．この場合の決定係数は  $R^2 = 0.99$  であり，回帰成分の高い寄与率を示している．

表 1: 片仮名 4 字畳語出現頻度 (1992 年)

出現頻度	畳語	累積出現頻度
0.073	バラバラ	0.073
0.070	ギリギリ	0.143
0.049	イライラ	0.192
0.023	トントン	0.215
0.020	ニコニコ	0.235
0.020	ゴタゴタ	0.255
0.020	ハラハラ	0.275
0.020	ハタハタ	0.295
0.019	ピリピリ	0.314
0.016	ズルズル	0.330

<sup>2</sup>相関係数  $r_{xy} = \frac{S_{xy}}{S_x S_y}$ ， $S_x, S_y$  は 1992 年と 2006 年の分布の分散， $S_{xy}$  はそれらの共分散

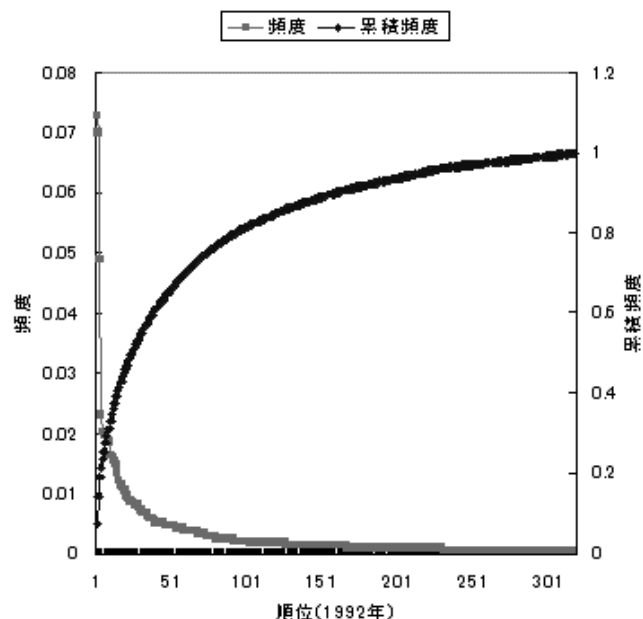


図 1: 片仮名 4 字畳語の頻度分布 (1992 年)

## 出現頻度分布の経年変化の調査

1992 年と 2006 年の片仮名 4 字畳語出現頻度上位 20 語を表 2 に示す．この結果より 1992 年の上位 3 語はすべて 2006 年の上位 3 語に含まれていることが分かる．この中で \* が付けられた語は他方の上位 20 語内に現れない片仮名畳語である．2006 年において頻度が高い語の多くが 1992 年にも高頻度であったことが分かる．

1992 年，2006 年双方の記事に現れる片仮名 4 字畳語について，出現順位上位  $N$  件内 (1992 年の順位) での順位相関係数のグラフを図 2 に示す．これにより出現頻度の高い語で相関係数も高いことが分かる．全体としての相関係数は約 0.6 となった．これは 1992 年に多く使用された畳語が 2006 年にも多く使用されているという傾向を示している．

## 考察

### 出現頻度分布

単語の頻度分布についてはジップ則が有名である．日本語単語の場合の冪指数については亀山によって雑誌を対象に調査されている [3]．この報告によると対象雑誌 [4, 5] に対して，べき指数はそれぞれ  $-1.031$  から  $-1.065$  と  $-1.009$  から  $-1.056$  に分布している．一方，本調査による片仮名 4 字畳語と片仮名 4 字畳語 (1992

表 2: 片仮名 4 字畳語出現順位の比較 (1992 年と 2006 年)

出現順位	1992 年	2006 年
1	バラバラ	バラバラ
2	ギリギリ	ギリギリ
3	イライラ	イライラ
4	トントン *	ワクワク *
5	ニコニコ	ニコニコ
6	ゴタゴタ *	ドキドキ
7	ハラハラ	コツコツ
8	ハタハタ *	ハラハラ
9	ピリピリ	キラキラ
10	ズルズル *	ボロボロ
11	ガタガタ	ピリピリ
12	コツコツ	ドロドロ *
13	ボンボン	バタバタ *
14	ボロボロ	ボンボン
15	ジワジワ *	サラサラ *
16	カンカン *	バリバリ *
17	ドキドキ	ポレポレ *
18	ガラガラ *	ピカピカ *
19	ジリジリ *	ガタガタ
20	キラキラ	パラパラ *

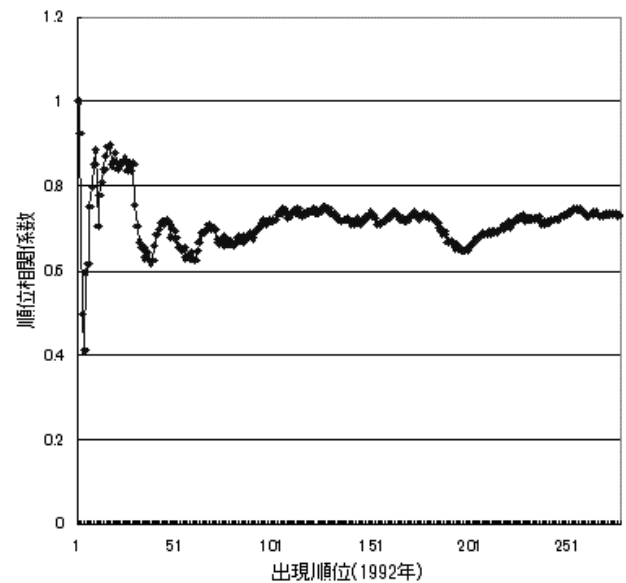


図 3: 平仮名 4 字畳語順位の相関係数

表 3: 4 字畳語頻度分布のべき指数 (括弧内は  $R^2$ )

	1992 年	2006 年
平仮名	-1.521(0.98)	-1.406(0.98)
片仮名	-1.121(0.95)	-1.045(0.95)

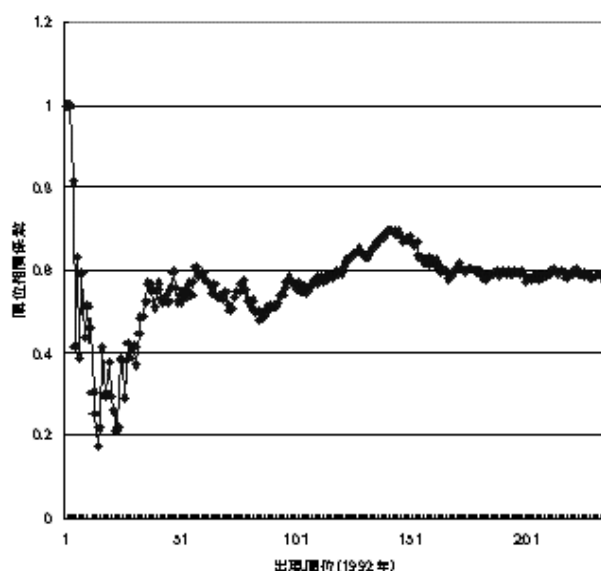


図 2: 片仮名 4 字畳語順位の相関係数

年と 2006 年) の頻度分布もこの分布で近似される (決定係数 95 % 以上)。べき指数については表 3 に示すように雑誌の単語の場合より低い結果となった。年別では平仮名・片仮名ともに 2006 年の方が大きい。また字種別では 1992 年・2006 年ともに片仮名の方が大きい傾向にある。片仮名畳語 2006 年のみが雑誌単語の水準に近い値となった。これを語の出現度数の側面からみると、字種別では平仮名畳語が片仮名畳語の 4 倍以上の度数であった。また年別では 1992 年が 2006 年の 1.1 倍以上であった。これにより出現度数の高さがべき指数の低下に影響している可能性が考えられる。今回の調査では資料の各 1 年分全体を用いたが、今後はその使用量を制限して両年の度数を揃えることで、度数の影響を排除した評価を行う予定である。

## 出現頻度分布の経年変化

畳語の頻度分布の経年変化については 1) 分布形状変化と 2) 個々の畳語の頻度変化とを考える。これまでツイート中の形態素を対象にした調査では約 1 年以

内の範囲で分布形状は安定しており、個々の語（形態素）においては高頻度語の順位変化が小さいことが観察されていた [2]。

#### 4.2.1 分布形状の経年変化

分布形状は頻度順位の冪乗に比例する形で近似されている。この形状は約 15 年離れた 1992 年と 2006 年で変化がない。べき指数についてはこの間に若干上昇が見られるが前述した通り、出現度数変化の影響の可能性が考えられる。

#### 4.2.2 個々の畳語の頻度変化

1992 年と 2006 年の分布の間には、平仮名・片仮名ともに正の順位相関が見られる (図 2,3)。約 15 年の経過後に平仮名で約 70 %、片仮名で約 60 %程度の頻度傾向が保たれている。新聞記事中の畳語分布は全体として安定的であるが、出現頻度別に相関係数を観察すると、出現頻度の高い語の相関係数が大きい。つまり高頻度の語は高い頻度を保つ傾向があり、低頻度の語は経年変化の度合いが高いと言える。[2] の調査で短期間の場合に観察されていた現象が長期間の場合でもみられることが確認された。

平仮名畳語と片仮名畳語とを比較すると、平仮名の方が相関係数の低下が緩慢であり、全体としての値も高くなっている。実際、平仮名の場合には 1992 年の上位 15 語までの語が全て 2006 年の上位 15 語に含まれるのに対して、片仮名で同様な語は上位の 3 語だけであった。もし出現度数の絶対数が相関係数の大きさに影響するならば、絶対度数が片仮名の約 4 倍高い平仮名畳語で相関係数が高くなる可能性がある。今後はそれを確認するために調査資料の分量を調整して観測度数を変えて調査する予定である。

## 関連研究

本研究の調査対象である畳語についての研究には、畳語を資料から広く採集し、分類や計量を行うものと畳語の特徴や特有の用法について論じるものがある。本研究は前者に近い。この種の先行研究には松林の畳語分類がある [6]。松林は上代から近世までの作品<sup>3</sup>を資料に用いて、畳語を収集しそれらに形態的な分類を行った。また資料ごとに出現度数の統計も行っている。その他、熊谷は俳句・狂歌・川柳・民謡・童謡を資料

にした畳語の収集と分類を行った [7]。その中では畳語の種類を擬音語・擬態語・対応語・命令型畳語・継続表現などとしている。また感覚・感情・所作・強意・動作・形容詞・副詞への分類も行った。これらに対し本研究では分類でなく、頻度統計を行い頻度分布の形状やその経年変化について論じた。本研究と同じく日本語の頻度分布形状に関する研究には雑誌中の単語についてジップ則を検証した [3] があり、頻度変化に関する研究にはツイートを対象に調査した [2] がある。

## おわりに

本稿では、新聞記事中における 4 字畳語（平仮名・片仮名）の出現頻度分布などの調査結果について報告した。この調査の結果、分布形状については一般の単語同様にジップ則に従うことが確認された。また 1992 年と 2006 年との間の経年変化を分布形状と個々の畳語の頻度変化について調べたところ、分布形状は安定しており、個別の畳語の頻度変化は高頻度語であるほど順位相関の変化が少ないことが確認された。これは数ヶ月の短期間で見られた頻度変化の性質が 10 年以上にわたる経年変化においても同様であることを意味する。今後は絶対出現度数の影響を調整した調査や調査対象年数の拡大を行い、連濁タイプの畳語（「ときどき」等）についても調査対象にする予定である。

## 参考文献

- [1] 中渡瀬秀一, 大山敬三: 畳語の頻度分布調査, 信学技報 TL 思考と言語, TL2011-56, pp. 13-16, 2012.
- [2] 荒牧英治, 増川佐知子: 微小時間における日本語の変化とその法則, 言語処理学会第 17 回年次大会発表論文集, 2011.
- [3] 亀山寛: 日本語単語頻度数におけるジップ則, 計量言語学, 26 巻, 4 号, pp. 123-138, 2008.
- [4] 国立国語研究所: 現代雑誌九十種の用語・用字第一巻, 秀英出版, 東京, 1962.
- [5] 国立国語研究所: 国立国語研究所報告 121 『現代雑誌の語彙調査 - 1994 年発行 70 誌 - 』, 2005.
- [6] 松林 睦枝: 畳語の研究, 日本文学, Vol. 32, pp. 59-78, 1969.
- [7] 熊谷 忠三郎: 畳語の研究, 創文社, 東京, 1973.

<sup>3</sup>岩波日本古典文学大系の和文作品から約 90 冊