日本語語彙の統計的性質：異なる語彙調査資料を用いて

日本語語彙の統計的性質　－異なる語彙調査資料を用いて－

日本語教育センター　中野 洋
nakano@kokken.go.jp

要旨：

語彙についての統計的法則がある。本発表では、国立国語研究所が行った９つの語彙調査を用いて、異なる内容の調査対象間にも共通に存在する語彙とその使用率の関係について述べる。

使用率の大きい語彙は、そのほとんどが他の調査にもよく用いられる。共通度の分布図によっても「高頻度語彙」と名付けてよい語彙の存在が確認できた。しかし、その中にもその調査対象の特徴語彙とも言える語彙も含まれている。一方、使用率の小さい語彙にも他の調査に用いられる語彙が存在する。これらの語彙は、具体的な内容をともなうものであり、語彙教育の対象となる重要な語彙といえる。

使用率と語彙の関係の解明は、言語教育や辞書作成における語彙の選択法の開発に貢献する。

キーワード： 高頻度語彙； 使用率； 共通度分布； 語彙調査； 語彙の選択

１．目的

語彙について、いくつかの統計的法則が発見されています。

よく知られているジップの法則は、使用度数(f)とその度数を持つ見出し語数(k)についての分布についてですが、精度がよくありません。同じ系統のものに水谷の法則があり、こちらの方がずっと精度がよいのです。また、文章における品詞別の語彙分布について、延べレベルで樺島の法則が、また異なりレベルで大野の法則があります。これらの詳細については文献を参照してください。

さて、本発表では、国立国語研究所が行った９つの語彙調査を用いて、異なる内容の調査対象間にも共通に存在する語彙とその使用率の関係について報告します。これは、ひとつの文章における語彙ではなく、雑誌やテレビなど情報伝達媒体を同じくする多くの集合における語彙について統計的関係を調べたものです。残念ながら法則には程遠いものですが、ここでの報告が、法則発見のきっかけにでもなれば幸いです。

２．意義

学問的には、計量語彙論での研究の一つです。われわれ人間が恣意的に用いた語彙にも科学的法則が存在するのです。もともと語彙調査は、語彙教育に役立てるという実用目的のために行われました。また、文章作成・解読のためやワープロや言語処理のための辞書に載せる語彙選択にも役立てることができるでしょう。しかし、私は、純粋に学問的関心によってこの研究を行っています。もし、実用にも役立つのであればすばらしいことだと思います。

3 方法

どのような調査でも共通の高頻度語彙は存在するのでしょうか。また、それほど多くは用いられないけれど共通に使用される語彙が存在するのでしょうか。

3．1 調査対象 電子化されている７つの語彙調査の語彙表を調査対象とします。

調 査 対 象	略 称
「テレビ放送の語彙調査」（音声・画面）	テレビ音声・テレビ画面
「雑誌 90 種の用語用字」調査	雑誌
「中学教科書の語彙調査」（M 単位・W 単位）	中学 M 単位・中学 W 単位
「高校教科書の語彙調査」（M 単位・W 単位）	高校 M 単位・高校 W 単位

3．2 使用率による分類

それぞれの調査で得られた語を度数順に並べ、その累積使用率によって次の５つのクラス（集合）に分類します。

クラス１： ０%～２０%以下の語　　クラス２： ２０%～４０%以下の語
クラス３： ４０%～６０%以下の語　　クラス４： ６０%～８０%以下の語
クラス５： ８０%～１００%以下の語

クラス分けに使用率を用いたのは、異なる調査規模を比較するためです。

それぞれのクラスの異なり語数を次に示します。

クラス	ﾃﾚﾋﾞ音声	ﾃﾚﾋﾞ画面	雑誌	中学M	中学W	高校M	高校W
1	13	37	26	20	36	20	43
2	66	384	168	100	213	135	330
3	399	1,250	785	278	766	369	1,389
4	2,489	2,689	3,495	810	2,806	1,131	5,995
5	14,455	3,316	35,522	6,780	13,770	13,844	32,994
合計	17,422	7,676	39,995	7,988	17,591	15,499	40,751
延べ語数	102,856	16,579	438,758	131,774	100,709	321,058	233,855

なお、数字や助詞・助動詞などの扱いのため、文献とは数値が異なることがあります。

3．3 共通度の計算

異なる調査の各クラスとの語彙の異同を調べ、それらの共通度を計算します。

調査によっては、調査単位や同語異語の判別の基準が異なりますが、共通する語も少なくないためまとめて扱います。そのため、この報告は全体の分布を概観するにとどめなければなりません。見出し語形や語の判別のための情報も異りますが、これらは統一しました。

共通度は、次の式で求めました。

$$K(AB) = \Sigma i\, pi(A) \times 5$$

K（AB）：語彙Aの語彙Bとの共通度

pi(A)：語彙Aにおける、語彙Bと一致した語 i の使用率

これは、あるクラスの語彙全体の使用率、すなわちその調査における使用率の２０%のうち何パーセントが他のクラスの語彙と一致するのかを示した値です。正規化のために５倍します。

3．4 結果と分析

（1）クラス１の共通度の分布

最初に、使用率の最も大きいクラス１の語彙の共通度の分布を示します。次の表は、内容の似ている「中学Ｍ単位・中学Ｗ単位・高校Ｍ単位・高校Ｗ単位」の共通度の平均値です。

ｸﾗｽ	1	2	3	4	5	合計
中学M	0.88	0.03	0	0	0.01	0.92
高校M	0.86	0.04	0	0	0.01	0.9

ｸﾗｽ	1	2	3	4	5	合計
中学W	0.73	0.15	0	0	0	0.88
高校W	0.72	0.14	0.01	0	0	0.86

各調査のクラス１の語彙の共通度は他と比べて明らかに大きいことがわかります。また、そのほとんどがいずれかのクラスに出現していることが、共通度の合計の大きさによってわかります。

次は、内容の異なる「雑誌・中学Ｍ単位・テレビ音声・テレビ画面」の共通度の平均値です。

ｸﾗｽ	1	2	3	4	5	合計
雑誌	0.46	0.27	0.05	0.01	0.05	0.84
テ音	0.34	0.1	0.16	0.08	0.07	0.76

ｸﾗｽ	1	2	3	4	5	合計
中学M	0.52	0.33	0.01	0.01	0	0.87
テ画	0.14	0.07	0.12	0.15	0.18	0.65

異なる内容の調査間の類似度は、似た内容の調査と比べて小さいのですが、「テレビ画面」を除き、クラス１の類似度が他のクラスより明らかに大きいことがわかります。「テレビ画面」のクラス１の共通度は、他のクラスと比べて小さく、その語彙の特殊性を示しています。画面では、テロップやフリップ、看板など語句が多く、文の形をなさないものが多く、文を成り立たせるための語彙、いわゆる形式名詞、基本動詞、指示詞などが少ないためだと考えられます。また、クラス１の語彙の多くが他の調査の各クラスに出現していることも先と同様です。

（2） すべてのクラスでの共通度の平均値の分布

各クラスの語彙の異同を概観するために、共通度の平均値のグラフを示します。

1） いずれも同じクラスの共通度が大きい。クラスが近いほど共通度が大きく、したがってそのクラスを頂点に山型の分布を示す。
2） 調査対象の内容が似た調査の共通度が大きい。また、山型の分布がよりはっきりしている。さらに、似た内容の対象の調査結果の方が異なる内容の対象の調査結果より共通度が大きい。
4） テレビ、特に文字の共通度が低く、また分布が他と異なる。
5） 使用率の小さいクラスでは、共通度が小さい。しかし、ここでも同じクラスの語彙との共通度が高い。

図の縦軸は共通度、横軸はクラス。図中の記号は、図１から５までは、菱形が「中学Ｍ単位」、三角が「高校Ｍ単位」、四角が「中学Ｗ単位」、×印が「高校Ｗ単位」。また、図６から１０までは、四角が「雑誌」、菱形が「中学Ｍ単位」、三角が「テレビ音声」、×印が「テレビ画面」です。

図1 クラス1平均 中M 高M 中W 高W

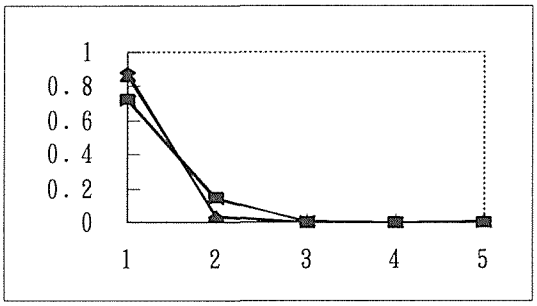


図2 クラス2平均 中M 高M 中W 高W

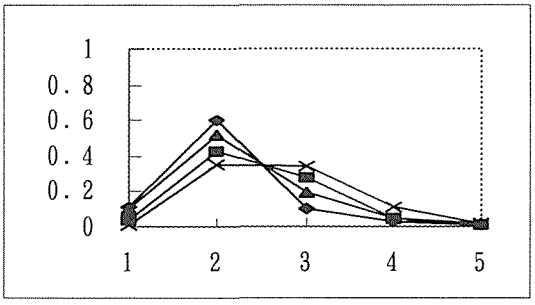


図3 クラス3平均 中M 高M 中W 高W

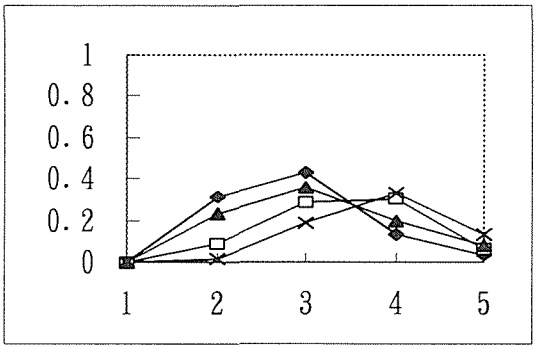


図4 クラス4平均 中M 高M 中W 高W

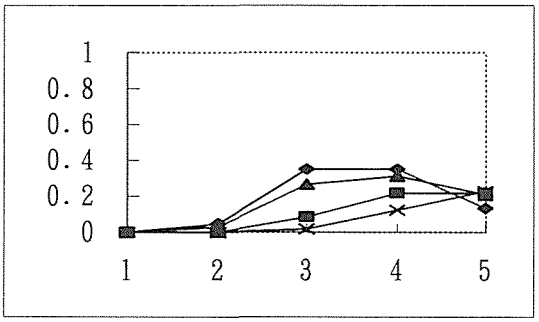


図5 クラス5平均 中M 高M 中W 高W

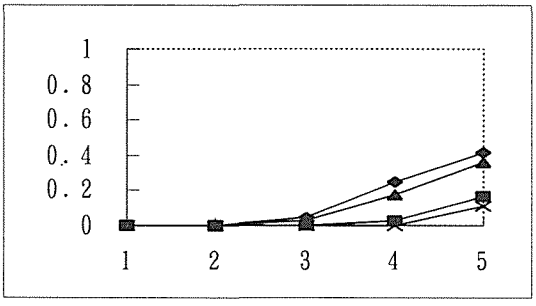


図6 クラス1平均 雑誌 中M ﾃﾚﾋﾞ(音声・画面)

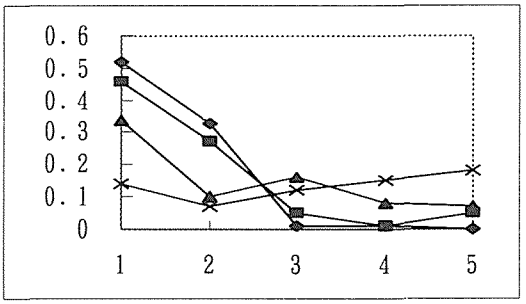


図7 クラス2平均 雑誌 中M ﾃﾚﾋﾞ(音声・画面)

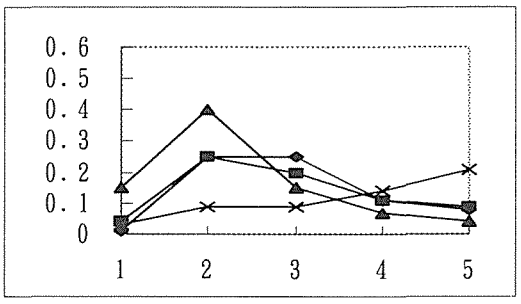


図8 クラス3平均 雑誌 中M ﾃﾚﾋﾞ(音声・画面)

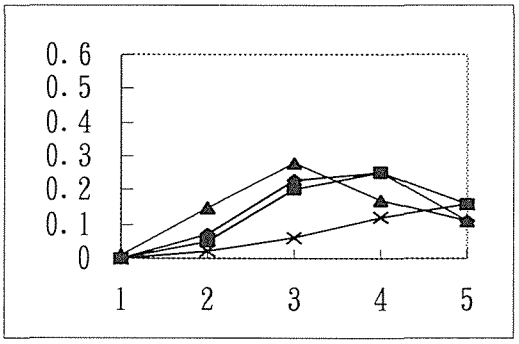


図9 クラス4平均 雑誌 中M ﾃﾚﾋﾞ(音声・画面)

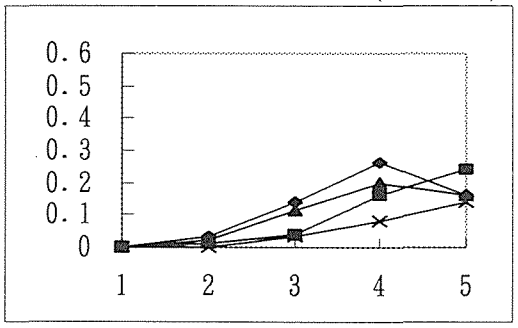
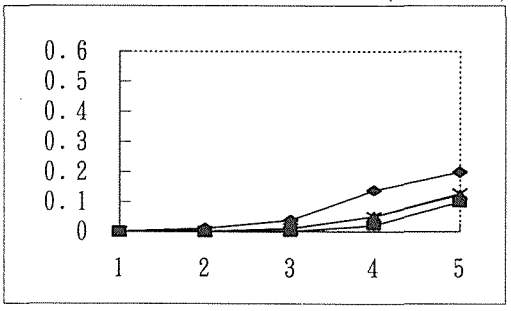


図10 クラス5平均 雑誌 中M ﾃﾚﾋﾞ(音声・画面)



4．語彙の分析

4．1 目的

前章で、どのような調査でも共通する高頻度語彙が存在し、それが共通度を高めているらしいことがわかりました。それがどのような語彙なのか、またそれ以外のそれぞれの調査に固有の高頻度語彙が何なのかを分析し、語彙の特徴を明らかにします。

4．2 クラス１の語彙

（1）対象

「新聞」（文献２）と「知識人」（文献６）を加えてクラス１の語彙をここでの対象とします。異なり語数は、以下の通りです。

	ﾃﾚﾋﾞ音声	ﾃﾚﾋﾞ画面	雑誌	中学M	新聞	知識人	中学W	高校M	高校W
ｸﾗｽ1	13	37	26	20	50	14	36	20	43

（2）分析

「テレビ音声・テレビ画面・雑誌・中学M・新聞・知識人」のすべてにクラス１で出現した語

範囲数	語数	見出し
6	1	する
5	3	こと、なる、この
4	4	これ、いう、いる、ある
3	8	その、それ、一、二、三、五、十、二十
2	17	もの、よい、ない、零、四、六、七、八、九、年、さん、お はい、うん、え・ええ、よう、れる・られる

これらは、いずれもどの調査対象にもよく用いられる語です。しかし、個々に見れば、調査対象の影響もみられます。すなわち、数字の「一」から「十」、「二十・零」、助数字の「年」が複数の調査でクラス１に入るのは、新聞やテレビ、あるいは教科書など情報を扱う対象のためでしょう。また、接辞の「お・さん」が入るのは、意見や心情を表す文章が多い「雑誌」や、音声言語の「テレビ」「知識人」が対象だからでしょう。「はい、うん、え・ええ」は、音声言語の専用語彙といえます。

4．3 他のクラスの語彙

（1）対象

内容の異なる「雑誌・中学Ｍ・テレビ音声」の全クラスの語彙を対象とします。

（2）分析

各クラスの共通度分布はそのクラスを頂点とする山型を描きます。この３つの調査のすべて同じクラスに現れる語彙数は、クラス１は６語、クラス２は10語、クラス３は35語、クラス４は114語、クラス５は496語です。この一部を下に示します。

クラス　　語　例

1　こと　この　これ　する　それ　なる

2　方　しかし　ため　つくる　とき　ところ　とる　なか　もつ　ゆく

3　色　数　形　体　現在　下　時間　時代　全国　人間　町　右
　当たる　動く　生れる　変わる　進む　送る　変える　住む　違う　続く　続ける
　始める　引く　開く　増える　求める
　新らしい　全て　たくさん　例えば　一方　どんな　長い

4　イタリア　沖縄　関東　跡　安全　意志　石　一定　改革　海外　環境　管理　ガラス
　当てる　動かす　移る　奪う　売る　生かす　遅れる　及ぶ　関わる　限る　囲む　重なる
　およそ　固い　厳しい　極めて　苦しい　濃い　互い　近く　細い　貧しい　豊か　（以下略）

5　会津　阿蘇　伊豆　伊勢　アルプス　明智　新井　家康　上がり　悪習　あつまり　網
　育児　池　意　向　遺産　泉　板　市　一文　一連　一旦　祈り　衣服　医療　引力　円形
　改まる　荒れる　祝う　老いる　かなえる　枯れる　乾かす　帰する　競う　（以下略）

内容が異なる全ての調査に現れる語彙には、各調査の特徴語彙は含まれません。クラス１や２にはいわゆる基本語彙が現れています。また、クラスが下がるにつれて具体的な内容を表す語彙や意味の狭い語彙が増えます。しかし、使用率の小さい語彙にも異なる内容の調査間での共通語彙があり、かつそれらは一般に知られている語彙も少なくありません。

５．まとめ

使用率の大きい語彙は、そのほとんどが他の調査にもよく用いられます。共通度の分布図にも現れているように「高頻度語彙」と名付けてよい語彙が存在することが確認できました。しかし、その中にもその調査対象の特徴語彙とも言える語彙も含まれています。使用率の小さい語彙にも異なる内容の調査に現れる語彙があります。これらは、国語教育や日本語教育にとっても重要な語彙であり、語彙選択の新たな方法を生む可能性があります。

６．文献

1. 国立国語研究所(1962)『現代雑誌九十種の用語用字　第一分冊　総記および語彙表』（秀英出版）
2. 国立国語研究所(1970)『電子計算機による新聞の語彙調査』（秀英出版）
3. 水谷静夫(1974)「国語学五つの発見再発見)(創文社)
4. 水谷静夫(1982)「数理言語学」（現代数学レクチャーズＤ－３、培風館）
5. 国立国語研究所(1983)『高校教科書の語彙調査』（秀英出版）
6. 志部　昭平(1980)『日本人の知識階層における話しことばの実態－語彙表－』（文部省科学研究費特定研究「言語」、「日本語教育のための言語能力の測定」資料集２、国立国語研究所）
7. 国立国語研究所(1997)『テレビ放送の語彙調査（Ⅱ）』（大日本図書）
8. 石綿敏雄(1989)「雑誌・新聞語彙と教科書語彙」（『高校・中学校教科書の語彙調査　分析編』、秀英出版）