日本語における音素配列論上の制約

高　橋　　渉

1　目　　的

本稿の目的は国立国語研究所（1989）をデータとして，小学生の作文使用語彙における様々な音素配列論的制約（phonotactic constraints）について，パーソナルコンピュータを援用した計量分析を行い，ついでそのような制約がなぜ生じるのかを考察することである。

本稿で用いる資料は高橋（近刊）で用いているものと同一のものである。そこでは同一資料における日本語のCVという音節構造の計量分析を，他言語のデータと比較しつつ行なったのであるが，今回はCVという音節単位に限定せず，従来から指摘されている様々な音素配列論に関わる制限がどの様に資料の中で作用しているかを，特に日本語語彙における出現頻度順表のうちから，高出現度の語彙群と，低出現度の語彙群との間に見られる音素配列論的制約の差異に焦点を当てて分析を試みるものである。

2　資料とその分析方法

本章では国立国語研究所（*ibid.*）の概略を紹介し，そこに提示されたデータを本研究のためにデータベース化した際に行なったいくつかの処理ならびにデータの分析方法について述べることにする。

同書はその書名の通り，児童の作文を調査対象とする計量語彙調査である。そこでは，児童の語彙使用過程およびその発達過程に関する具体的な資料が語彙表の形で提供されている。研究対象となった作文は全国各地から収集された10年間にわたって書かれた文集で，1年から4年までは各学年400編，5，6年はそれぞれ360編で合計2,320編であるが，これらは当初は25,000編に及ぶ作文からジャンル的に予備選抜を行なったものである。(1)

これらの作文の原文章からα単位（文節から助詞，助動詞を除いたものに近い）を基に単位語を取り出し，さらにそれらの語から，地名，人名，方言等一般性の低い語彙を除いたものがその資料総体である。これは，異なりとしては総数20,849語，また延べ使用回数としては474,243語にのぼる。これら語彙の五十音順の一覧表が第一表，さらに同一コーパスを出現度数順に並べたものが本研究で用いる第二表である。

本稿ではこの第二表を使用頻度の高い群（以下H群）と使用頻度の低い群（L群）とに分ける。L群は1年から6年までそれぞれ一回ずつ使用されたことになる使用回数6回の語彙であり，その総数は544語である。(2)

この様にしてL群の544語を選んだあと，対応するH群は使用回数延べ17,392回で第一位の語，「いる」から544位「通り」（使用回数は132回）までを選ぶことにする。

さてこれらのHL各群の語彙を計量分析しやすいようにデータベース化は次のような手順で行なわれた。

まずH群を上位から70語ずつ区切ってグループ分けし，HIGH1からHIGH7と名付ける。最後のファイルは残り54語を含むファイルHIGH8である。H群は使用頻度順のリストであるから，HIGH1＞HIGH8の順に内部の語彙の使用頻度は減少していく。他方L群は全て使用頻度6回語彙のみであるので，このサブグルーピングにはとくに配慮はせず，単に原資料のリスト順に入力しLOW1からLOW9までの9ファイルに分けた。

同資料は発音記号では表記されていないので，本稿ではそれぞれの語彙を以下の特別な表記法を該当する箇所に使用しつつ発音表記に直して入力した。

(1)

/ʃ/ ＞ /S/
/tʃ/ ＞ /T/
/dʒ/ ＞ /D/
/ŋ/ ＞/G/
syllabic nasal＞/N/
/ʔ/ /＞ /Q/

(1)において示した発音記号の改変は主としてパーソナルコンピュータに入力する際の半角文字種の制限によるものである。さらに後の検索作業において余計な検索条件を付加しなくてもすむように破擦音/tʃ/，/dʒ/は単一のセグメントとみなされるようアドホックではあるが上記のように変更表記した。またnonsyllabic nasalとsyllabic nasalとを明確に区別するため/one:saN/のように/n/と/N/で区別することとした。最後に長音母音のマークは/:/を用い，これも1セグメントとみなすこととした。

分析の方法の最後に本資料を分析する際に用いた検索ソフトmifindについて簡単に触れておきたい。同ソフトはメガソフト社製のエディターMIFESに附属する文字列検索ソフトであり，指定した文字列を極めて高速に検索することができる。さらに同ソフトはUNIXの正規表現をサポートしているので，様々な検索パターンを簡単に指定できる。一例を挙げよう。

(2)

a. mifind -r -c2 -d "^r[aiueo]" filenamel ＞ filename2
b. mifind -r -c2 -d "[aiueo]/"$"^r[aiueo] filenamel ＞ filename2

(2a)はfilenamelのデータの中から行の先頭がラ行で始まる語を抜き出し，その結果をfilename2というファイルにリダイレクトしている。

また同検索ソフトは論理否定，論理積，論理和などの論理演算子を用いて複合条件による

検索も行える。(2b)がそれで，これはfilenamelというデータファイルの中から語末が母音で終わるかあるいは語頭がラ行で始まるかのどちらかの条件を満たす語をfilename2に書き込むものである。[3]

本章の最後にデータのほんの一部を参考のために掲げておく。[4]

(3)　サンプルデータ

H群			
iru	17,392	1	
suru	10,624	2	
iu	9,626	0	
wataSi	9,524	3	
naru	7,948	2	
kuru	6,376	2	
L群			
SiNsatu	6	3	
Ditubutu	6	4	
kokoroGamae	6	5	XKo
taikeNsuru	6	4	
hjoQto	6	1	
haNdaNsuru	6	4	XN

3　分　　析

3.1　語彙の長さ

以上2で述べた分析方法を用い3では，いくつかの項目についてHL両群別に特色を検討することとする。

まず語彙の長さについて分析してみよう。本来本稿で扱うデータは高橋（*ibid.*）の中で用いられたものと同一である。そこでもH群とL群との間に存在するCVという音節の数の相違については指摘しておいたが，本稿では改めてCVの数ではなく個別のセグメントを基盤に計量をしなおしてみた。[5]

(4)

	語彙数	総セグメント数	一語平均
H群	544	2,819	5.18
L群	544	4,001	7.35

平均値でみると，H群よりL群のほうが約42%も一語当りのセグメントは多いことがわかる。次に，各群それぞれのセグメント数別の語彙の度数分布は(5)のようになる。

(5)

セグメント数	H群	L群
1	1	0
2	17	0
3	49	12
4	139	42
5	111	49
6	136	118
7	41	87
8	38	90
9	6	48
10	2	45
11	3	22
12	1	18
13	0	9
14	0	2
15	0	2

H群の1セグメントの語は/a/，2セグメント語は/hi/，/te/，/me/，/ki/等。11セグメント語は/biQkurisuru/始め3語。H群最長の12セグメント語は/dokidokisuru/であった。一方L群の語群において最長の語は15セグメントあり，それは/kaNto:daiSinsai/，/ko:kaGakuko:Gai/の2語であった。

(5)に明らかなようにH群は最短の語は1セグメントの語から最長12セグメントの語までであるのに対し，L群は2セグメント以下の語はなく，反対にH群には存在しない13から15セグメント語が総計13語存在する。

H群では4セグメント語（大抵はCVCVの2音節語）以下の語は37.8%も占めるのに，L群では4セグメント語以下の語は9.9%にすぎず，逆にH群で10セグメント以上の語は1.1%のみであるのにL群では18%にものぼる。

このように，様々な観点から見て，H群の語に比してL群の語はセグメントを多く含んだ語が多い事が実証される。

前述の通りH群のデータは頻度順にHIGH1からHIGH8までのファイルになっている。そこでHIGH1からHIGH8までのファイル別にセグメント数の平均値をとると(6)のようになる。

(6)

1- 70	4.29
71-140	4.93
141-210	5.23
211-280	5.89

281-350	5.47
351-420	5.29
421-490	5.32
491-544	5.46

ここでは，順位が下がるにつれてセグメントの長さも増加傾向にあることが期待されるのだが211位から280位のグループと281位から350位のグループにその前後のグループに比してセグメント長の増大が見られる。この211位から350位の140語に平均値を普通より高くするなんらかの要因が存在すると思われるが内容を検討してもその要因を特定できなかった。但し前述のグループの平均値を除けば，期待通りリストの順位が下がるにつれてセグメント数も増大する傾向は明らかに存在する。

なお，L群はもともと全てが6回使用語のみからなるファイルであるから下位集合ごとの平均値をとる意味はない。

3.2　個々のセグメントに課される制約

前節では語彙の長さに関するHL両群の差異について検討したが，本節では個々のセグメントに課される音素配列論上の制約がHLそれぞれにおいていかに異なった振舞いを見せるかを検討することとする。(6)

3．2．1　/Q/について

日本語の促音音素/Q/は普通母音と/p，t，k，c(＝ts)，s/の間に生じると築島（*ibid.*）は指摘するが，我々のデータの中に生じる語彙には(7)のように/tʃ/と/ʃ/も生じている。

(7)　H　/iQʃo/（一緒）
　　/iQʃo : keNmei/（一生懸命）
　　L　/neQtʃu : suru/（熱中する）
　　/piQtʃa : goro/（ピッチャーゴロ）
　　/keQʃo : seN/（決勝戦）

小泉（1989：16）はこの促音音素/Q/はその後に/h，w，y，m，n/を従える事はないと述べている。我々のデータには，たまたま存在しないが，外来語起源の「マッハ」(＝/maQha/)は日本語としては，極めて異例の語彙と言えよう。

我々のデータに関する限りは築島（*ibid.*）の記述より，小泉（*ibid.*）のネガティブな記述のほうが，より狭く促音音素/Q/の生じる環境を指定していると思われる。但し/Q/の左に生じるセグメントは母音に限られるという築島の記述は妥当である。

/Q/自体のHL両群における生起数は，それぞれ20回と33回で全セグメント数に対する割合は0.71％と0.82％となり，HL間に際だった差異は存在しない。

3．2．2．　ハ行の制約

築島（*ibid.*）は和語では/ha，hi，hu(＝fu)，he，ho/は語頭にくることが多く，語中語

尾にくることは少ないと指摘している。我々のデータの中でこの制約はどの様に現われているか検討してみよう。

/ha/ はHで27回，Lで30回生起する。Hに生じる27の語彙のうち25は語頭に生じ，語中に生じているのは/gohaN/（ご飯）のみ，語尾に生じるのは/haha/（母）のみである。/gohaN/ における接頭語 /go-/ を取り去れば実質 /-haN/ は形態素の最初に存在することになるが，本稿での分析の単位はα単位なので，語中の /ha/ として扱うことにする。/haha/ は全くの例外である。

L群に生起する30の /ha/ のうち20は語頭に生じているが残り10は語中に生じている。

(8)

/miharu/（見張る）
/ko : hakurire:/（紅白リレー）
/arumihaku/（アルミ箔）
/seNko : hanabi/（線香花火）他

L群には語末に生起する例はない。

/hi/ はHに11回，Lに12回生じるがHの /nihiki/（二匹）のみが語中に生じる例で，他は全て語頭である。

/hu/ はHに9回生じるが全て語頭である。Lに16回生じるがそのうち10回は語頭，5回は語中（/seihuku/（制服），/amehuri/（雨降り），/udetasehuse/（腕立て伏せ）他），語末には外来語 /gurahu/（グラフ）が一回生じている。

/he/ はHに3回生じるが（/heja/（部屋），/taihen/（大変），/heN/（変））語末に生起する例は見られない。Lに生起したのは /heitai/（兵隊と /heikaiSiki/（閉会式））の2語でいずれも語頭に現われる。

/ho/ はHに7回全て語頭に生じる。Lに12回生じるが，内訳は語頭に7回，語中に5回であって語末には生起していない。ただしこれは，我々の資料に関する限りであって，語末の位置に /ho/ を持つ語も「候補」，「橋頭堡」などいくつか存在するが明らかにこれらは，漢語である。

以上築島に従って日本語のハ行の音素配列論的制約を見てきたが，明らかにこの制約は和語，漢語の区別さらには，外来語か否かの区別にも依存しているようである。我々のデータはこの区別を明確にして入力していないので，今後の課題としては，データの中にこれらの区別を容易にできる仕組みを備えておくことが必要となろう。

3．2．3．　/r, g, z, d, b/ で始まる語

築島（*ibid.*）の最後の見解として，/r, g, z, d, b/ で始まるモーラを持つ語は和語には少なく多くは漢語，あるいは外国語であるという主張をしているので，データの中でどのように反映されているのか見てみよう。

ラ行で始まる語は和語には少ないことは従来からもしばしば指摘されるところである。H群ではラ行で始まる語は（9a）に挙げた3語のみである。

(9)

a　/renSu : suru/（練習する）
　　/renSu : /（練習）
　　/rinGo/（りんご）全体の0.55%

b　/reDi/（レジ）
　　/reibo : /（冷房）
　　/rikaisuru/（理解する）等18語3.3%

（9a）の「練習する」と「練習」はもちろん別語ではあるが，「練習する」は名詞を動詞化する形態素 /-suru/ がついたのみであるから，実質2語のみの生起といっても差し支えない。ところがL群にはその9倍にあたる18語の例が見られる。築島の言う他の4つの音素を語頭に持つ語の割合を以下（10）にまとめよう。

(10)

/g/ で始まる語
　H　5　0.9%
　L　35　6.4%
/z/ で始まる語
　H　6　1.1%
　L　6　1.1%
/d/ で始まる語
　Hで始まる語
　H　23　4.2%
　L　22　4.0%
/b/ で始まる語
　H　9　1.65%
　L　16　2.9%
/p/ で始まる語
　H　1　0.18%
　L　8　1.4%

(9)と(10)で見る限り，HとLに2倍以上の生起率の開きがあるセグメントは /r, g, p/ であり，他のセグメントは特にその差はほとんど同じと言ってよいものである。ここでは，/r, g, p/

についてのみ，HLに差が存在すると言える。

但しこれら5つのセグメントが語頭には生起することが少ないというのは事実であって，上記5つのセグメント以外で語頭に比較的少ないのはナ行がHで38回（6.9%），Lで45回（8%）であること，またヤ行（/ja, ju, jo/）がそれぞれ3%前後であること位で，他のセグメントはほぼ10%以上となっている。

特に母音で始まる語彙はHで130語と23%の高率を示し，Lでも75語（13%）と比較的高い割合である。（註7を参照のこと）

3．2．4　濁音，拗音，鼻音について

最後に築島が指摘した事実以外で我々のデータが示す事実を述べておこう。濁音，拗音，鼻音は次のような生起回数を示す。

(11)

		生起回数	全セグメント数に対する%
/N/	H	72	2.55%
	L	143	3.5%
/CjV/	H	7	0.25%
	L	32	0.78%
/G/	H	31	1.1%
	L	42	1.05%

(11)に示すように実際の生起回数はいずれもLのほうがHに比べてかなり高い回数を示している。しかし全セグメント数に対する生起の割合は拗音を除きいずれも大差はない。このことは，濁音，鼻音の出現率はHLとも変わりはなく，むしろHL間のもともとの総セグメント数の差が生起回数に反映しているだけであるといえよう。

4　結　　語

以上3節では日本語の音韻構造に現われる音素配列論上の制約を概観してきた。最後に本節において，それらの制約がどの様な原因から生じ，またどんな意味を持っているのかを考察してみる。

3．1で検討した語彙の長はに関しては高橋（*ibid.*）でHL群間での音節数の相違について言及した際にも指摘しているように，「Zipfの法則」に合致した現象と考えられる。Zipfは単語の長さとその単語の使用頻度との間には反比例の関係があることを示した。これは，英語のみならずドイツ語のようにその語彙に極めて多数の多音節語を含む言語でさえ当てはまるという。（Crystal（1987：86－87参照）

なぜこの様な分布になるのかについては次のように説明される。長い語は調音の際あるいはその語を記憶する際に我々に様々な負担を強いるのであるから，その語を使用する際にはかなりのコストがかかるものである。そのコストの高さを補うものは長い語が

短い語に比べて持つ情報量の多さである。

一般に出現の確率が低いほど情報量は多いという。しばしば出現する語は仮に1回見逃したり聞きのがしたりしてもまたすぐに生じるのであるからその語がもつ情報量は少ないと考えられる。以上のように考えると，我々のデータ中H群には比較的短い語が，そしてL群にはセグメント数の多い語が生じているものまさに「Zipf の法則」の現われであるといえよう。（新英語学辞典参照）

前述のように国立国語研究所（*ibid.*）の語彙量は異なりとして20,849語，延べ使用数は474,243語にのぼるが，本稿でH群の語彙として使用した使用頻度一位から544位までの語の総使用回数は326,640回にのぼり，異なりでは全語彙20,849語中2.61%に過ぎないH群の544語が延べ使用数の68.8%を占めている。そしてL群はH群より42%も一語当りのセグメント数が多いのであるから，我々の日常使用する語は比較的短い語がほとんどであると言えよう。

このことは英語にも当てはまることである。渡辺（1990）は（12）に示す調査を報告している。

(12)

音節数	M1	M2	B1	B2	B3	B4
1	786	761	769	735	607	575
2	181	193	161	191	229	213
3	29	35	49	59	124	143
4以上	4	11	21	15	40	69

M1: Fall in love (movie)
M2：Kramer vs. Kramer (movie)
B1 ：New Prince English Course
B2 ：The Exploits of Moomingpapa
B3 ：Reader's Digest 1979 : Oct.
B4 ：Conditions on Rules of Grammar

彼は上記6つの資料の中から1000語を無作為に抽出しその音節数を計量したのだが，チョムスキーによる専門的な文献B4で，やや3音節以上の語が増加しているものの，他は6割以上が単音節語に占められていることが明らかになった。全体の平均値も1音節語が70%を占め，一方4音節語以上は2.6%強を占めるのみである。

渡辺の調査は頻度については何も言っていないが，我々の日常使用する語彙のいかに多くを短い音節を持つ語が占めているかを示しているものであろう。我々のデータの示すHL間のセグメント数の差異もきわめて自然な現象と言えよう。

最後に3.2で検討した個々のセグメントの生起に関する制約はどのように解釈したらよいのであろうか。促音音素 / Q / についてはHL間の出現率に大差はなかった。ハ行の制約に

ついては，本来語頭に生じることが多いというハ行にしては/ha, hu, ho/の3つが，Lにおいて語中語末に多く生起した。但しL群はα単位として分析する際に，どうしても複合語などの長い語として取り出されてしまうので，本来語頭だったセグメントが結果として語中に生じることになるものがあることに留意する必要があろう。(/seihuku/（制服），/amehuri/(雨降り）など参照のこと）

/r, g, z, d, b/のうち，HL間の生起率が明らかに異なっているのは/r, g, p/の3つであった。そしてこれら3つはいずれもLにおいてその出現率が高いことは注目に値する。濁音，拗音，鼻音においても出現率の違いが著しい拗音はL群において高い割合を示すのである。

以上HL群でその生起率の大きく異なるセグメントを列挙してみたがすべてL群に高い生起率を示すのは，どのような要因が考えられようか。使用頻度の低い語彙群に共通して，いろいろな意味でその生起に制限が課されている語が多いという事実は生成音韻論でしばしば論じられるセグメントのMarked-Unmarkedの概念と使用頻度の高低がなんらかの関連を持っているものと考えられる。(7)

以上本稿においては児童の作文使用語彙をデータに用い，日本語の音韻構造に課された様々な制約が，特に高頻度に用いられる語彙群と，使用頻度の低い語彙群それぞれにどのような現われ方をするのかを計量分析によってあきらかにしたものである。(8)

註

(1) 予備的選択によって除いたものの中には，読書感想文，詩，創作，手紙などがある。それぞれ児童の日常使用する語彙の実態にあわないタイプの語彙がデータとして収集されてしまうのを防ぐためと思われる。

(2) ただし実際の使用学年分布で各学年1回ずつ使用した結果6回になった語彙はひとつも存在しない。すなわち6回使用語をL群のデータの基準としたのは，L群を選定する際に，全学年通じて1回のみ使用された語彙を選べば，異なりとして総数8,914語にのぼるきわめて周辺的な語彙の集合になってしまう恐れがあるためだけで，単純に，平均すれば各学年とも1回使用した語という意味で6回使用語を選んだにすぎない。

(3) 以上の記述で明らかなように，検索ソフトを用いる際は様々な条件の組み合せをMS-DOSのコマンドラインから入力するので，必然的にその入力文字はほんの一部分異なるだけの長大な文字列を繰り返し入力することが多くなる。この様な状況で効率的に作業するのに有用なソフトとして，KSH.COMという常駐PDSソフトがある。これは，コマンドラインでの編集機能を大幅に高めるばかりでなく，それまで実行したコマンドをデフォルト値で500バイト程記憶しており（ヒストリー機能），いつでもコマンドを呼び出し適宜変更を加えて実行できるものである。

(4) サンプルデータ中，中の縦列はその語の総使用回数である。右の縦列はその語に含まれる音節数であるが，今回の研究では使用していない。Xko, XNなどの記号は，当該語彙にそのセグメントが2つ生起していることを示す計量のためのマークである。

(5) 以下3．2．1．から3．2．3．までの制約は築島（1964 : 11-14）による。

(6) 本データ以外には多くの/hu/で終わる語がある。（豆腐，毛布，豊富，等）これらをはすべて漢語であって，和語で語末に/hu/を持つ語は存在しないようである。

(7) HL間の出現率の相違で，今までの論議とは逆にH群の方が高頻度に出現したセグメントは語頭に現われる母音である。H群では，23%にものぼる語彙が母音で始まるのに対し，L群ではそ

のような語彙はほぼ半減して13%にすぎない。Marked-Unmarkedの概念に従えば，日本語において母音で始まる語というものは無標であることが考えられる。

(8) 本文中でも指摘したように本データにおいては，和語漢語あるいは外来語の区別を即座に集計する手段を講じなかった。これらの区別は高頻度語に和語が多く，低頻度語に漢語外来語が多いことが容易に想像できるから，当然しておくべき区別であった。本データベースの将来の改良の際の課題としておきたい。

参考文献

Crystal, David. 1987. The Cambridge encyclopedia of language. Cambridge : Cambridge U. P.

上村幸雄　1989.「五十音図の音声学」杉藤（1989）41-63

大塚高信　他（編）新英語学辞典　東京：研究社

小泉　保　1989.「音声と音韻」杉藤（1989）1-20

国立国語研究所　1989.「児童の作文使用語彙」（国立国語研究所報告98）東京：東京書籍

杉藤美代子（編）1989.「講座日本語と日本語教育」第2巻「日本語の音声と音韻（上）」東京：明治書院

高橋　渉　近刊「日本語のCV構造」字賀治正朋教授還暦記念論集　東京：開拓社（1991年6月発行予定）

築島　裕　1964. 国語学　東京：東京大学出版会

渡辺雅仁　1990.「推移と英語強勢規則」第8回日本英語学会口頭発表およびConference Handbook　65-68

（1990年11月30日　受理）