日本語への近似度を異にする文字系列

愛知教育大学 平 田 賢 一¹

LETTER SEQUENCES VARYING IN ORDER OF APPROXIMATION TO JAPANESE

KENICHI HIRATA

Aichi University of Education

Letter sequences of zero- through second-order approximation to Japanese were computer-generated using tables of single-letter and digram frequencies. The single-letter and digram structures of the sequences were tested by calculating correlation coefficients between observed frequencies in each sequence of 10,000 letters and those of the original tables (with zero-order material, by the goodness-of-fit test to uniformity hypothesis). The results confirmed that the sequences were properly generated. Two sets of tables are presented. The first set, Tables 1-3, gives 8 examples of 25-letter sequences of each of zero to second-order material. The other, Table 4, gives 30 10-letter sequences of each type, selected with the restriction that no letter should appear more than once in the sequence.

Shannon (1948) は、情報理論の分野における彼自身によるオリジナル論文において言語を一つの確率過程 (Stochastic Process) ととらえて記述し、日常言語への近似の程度を異にする人工言語作成の手続きを示した。彼は 1948 年の論文の中で、近似度 (Order of Approximation) の概念を用いて、英語のもつ系列依存性 (Sequential Dependencies) の度合を、さまざまな程度において反映するような近似文字系列を作成してみせた。

英語への 0 次の近似を示す文字系列は、英語アルファベットを無作為に並べた時に作られる系列である。1 次の近似系列は、同様に、無作為な系列であるが、系列内の各文字の出現確率が、英語アルファベットの相対出現頻度を反映するような系列である。2 次の近似系列は、系列内の Digram 出現確率が、英語 Digram の相対出現頻度を反映するような系列である。いいかえれば、系列内のある文字の出現が直前の 1 文字によって確率的に規定されているような文字系列である。同様に、一般に n -次近似系列は、系列内の n -gram 出現確率が、英語 n -gram の相対出現頻度を反映するような系列であり、近似度が高次になるにつれてその系列は次第に日常英語に近づいてくるとされている²。

このような近似度の概念は、英語に限らず他の言語にも適用されうる。それ故、ある言語の n -gram 相対出現頻度表が与えられるならば、その言語への n -次の近似を示す文字系列を作成することが可能である。

Shannon (1948) の論文が公刊された翌年に、すでに Miller 等は情報理論の概念が心理学の分野に適用され得ることを示し (Miller & Frick, 1949)、また近似系列を実際に知覚や記憶の実験の材料として用いた (Miller, Bruner, & Postman, 1954; Miller & Selfridge, 1950)。

以後、心理学のさまざまな分野において、情報理論の概念の適用が試みられ、また近似系列が実験材料として用いられてきた (Attneave, 1959; Garner, 1962; 今栄, 1969; 戸田・高田, 1959)。

ある言語への n -次の近似度を示す文字系列を作成するには、その言語の n -gram の正確な相対出現頻度表を必要とする。しかしながら、これまでそのような表が存在しなかったり、また、たとえ存在しても近似系列の作成には非常な手数を要するため、各研究者はそれぞれ代替的な方法で近似的な近似系列を作成して、実験に用いてきた。そのような方法の一つには、Shannon (1951) や Miller & Selfridge (1950) が用いた推測法 (Guessing Technique) による方法がある。この方法は、正常な成人の、彼の話す言語への熟知性に着目し、彼の自然な言語行動が、あたかもその言語の n -gram 相対出現頻度を正しく反映しているかのように考える。この方法では、 n -次の近似系列を作成するには、まず $(n-1)$ 文字からなる系列をある人に示して、 n 番目に続く文字を推測

¹ 本系列作成にあたって、愛知教育大学今栄国晴助教授より、終始有益な示唆と助言をいただいた。また、京都大学学術良二教授からは本稿作成にあたり御教示をいただいた。

² この点について、言語を確率過程としては記述できないとする生成文法の立場からの批判がある (Miller, 1967)。

させたり、あるいは、その系列を含むことばを自由に推測させる。はじめに示した系列に、その人が加えた最初の文字は系列に加えられ、はじめの系列の最初の文字を除いた新しい ($n-1$) 文字系列が次の人に示され、同様にしてその系列に続く文字を推測させる。この手続きをくり返すことによって、次々と人々を経て、 n -次の近似文字系列が作成されるわけである。この方法は、簡便で実用的であるが、成人の言語行動は、必ずしも客観的な相対出現頻度を正しく反映しない (Attneave, 1953) ことから、妥当性に問題があると考えられる。

Miller, Bruner, & Postman (1954) は、基本的には Shannon (1951) や Miller & Selfridge (1950) の用いた手続きにしたがったが、系列作成に人間を使用するかわりに書物を用いた。この方法では、 n -次近似系列を作成するには、($n-1$) 字からなる文字列と同一系列を含むことばを、書物のページをめくってさがし、そのようなことばがみつければ、その系列の直後に続く文字をその系列に加え、はじめの系列の先頭文字を除いた新しい ($n-1$) 字系列について同一手続きをくり返す。この手続きをくり返すことによって、必要な長さの n -次近似系列が作成される。しかし、この方法では、特に高次の近似系列を作成するには非常な手数と膨大な量のデータを要すると思われる。

いずれにせよ、これらの方法は、あくまでも近似的な方法であって、用いられる人間や書物によって種々のバイアスが系列に混入する。今柴 (1960) も指摘するように、情報理論的観点に立って近似文字系列を使用した研究が、その当初の期待をうらぎって必ずしも順調な発展をみせなかったのも、一つの理由には、実験材料として用いられた近似系列自体の不明確さに帰因する実験データの信頼性に問題があったことが指摘されなければならない。

本研究は、かかる問題点を解決するため、日本語 n -gram 相対出現頻度表に基づく日本語近似文字系列を作成することを目的とした。英語においては、Hirata & Bryden (1971) によって英語 n -gram 相対頻度表に基づく 0 次から 4 次までの近似文字系列表が作成されたが、日本語については、まだ一般に入手しうる形で公表された近似系列表は存在しない。

現在、日本語については、Trigram, Tetragram 等多文字綴の相対出現頻度表の公表されたものが存在しない。現在までに作成されたもので入手しうる最大の表は、今柴 (1960) の 40,000 字サンプルによる日本語 Digram 相対頻度表であろう。それ故、本研究では、今柴による Digram 相対頻度表、及びそれに基づく日本語各文字の相対出現頻度表を利用して、日本語への 0 次、1 次、2 次の近似を示す 3 種の近似文字系列の作成にその目的が限られた。

方 法

日本語近似文字系列作成において、その単位として用いられた文字は、次の 72 文字とスペースである。

アイウエオ カキクケコ サシスセソ タチツテト
ナニヌネノ ハヒフヘホ マミムメモ ヤユヨ
ラリルレロ ワン ガギグゲゴ ザジズゼゾ
ダデド
バビブベボ パピプペポ ャュョ (拗音字)
ッ (促音字)

今柴 (1960) により作成された日本語 Digram 相対頻度表は、その単位を音節にしている。それ故、本研究においては、今柴の Digram 頻度表を基にして、上記の 72 文字 (スペースを 1 文字と数える時 73 文字 = 以下同様) を単位とした新しい文字単位の Digram 頻度表を作成しなおし、それに基づいて近似文字系列を作成した。

一般に知覚や記憶・学習等の領域における心理学実験では、スペースを含まない近似語一単語への近似文字列が使用されるが、研究領域や研究目的によっては、スペースを含む文字系列、すなわち、文字を単位とした近似文を必要とすることが考えられるため、本研究では、スペースを含む系列と含まない系列の両方を作成することにした。

0 次の近似系列は、上述の 72 文字 (73 文字) から、乱数を利用してランダムに 1 文字ずつ引いては記録することによって作成される。各文字の出現確率は $1/72(1/73)$ で等しい。

1 次の近似系列は、0 次の近似系列の場合と同様に、72 文字 (73 文字) からランダムに 1 文字ずつ引かれて記録されるが、各文字は日本語における相対出現頻度に比例した出現確率をもつ。

2 次の近似系列作成においては、まず Digram 表から任意に一つの Digram を選び出す。例えばその Digram は“サタ”であったとする。次に、その Digram の後の文字“タ”ではじまる全ての Digram のリストと、その相対出現頻度を調べる。そして 1 次の近似系列作成の場合と同様にして、相対出現頻度に比例した出現確率をもつ“タ”ではじまる Digram の中から乱数を用いて一つの Digram をとり出す。例えばその Digram は“タカ”であったとしよう。選ばれた Digram “タカ”の後の文字“カ”が記録されて、系列“サタカ”が作成される。次は、“カ”ではじまる Digram について上述の手続きをくり返して一つの Digram をとり出し、その Digram の後の文字を記録する。このような手続きをくり返すことによって、一連の 2 次近似文字系列が作成される。

本研究における近似文字系列の作成は、名古屋大学大型計算機センターを利用して行なわれ、乱数の発生には FACOM 230-60 基本外部関数 RANDOM が使用された。

結 果 と 検 討

“方法”の項で述べた手続きにしたがって日本語への

TABLE 1

日本語への0次の近似度を示す25文字系列例

スペースを含む系列

バラサソスチドセテイテコボナルジナアハタヨバペワケ
 ブコヘネセラドユサテロモビカタヨシホズコネテドヌヒ
 ネケワベツドネヨボヨトルグウヤギザヨワカヤヘビド
 リテノユナタブユケベトササコビゼクテバホヘネザヨヨ
 リキヅワツトミマメババドフデヤバペヤンモネキュイブ
 ハボレトトラバデブミイフデビシグリバデブウザアズゼ
 フズテムビミダカチソダカネヒニサムペアロルモシノツ
 ニゲボジチウレデダマウペイワデチゾユブヘスヨビダゴ

スペースを含まない系列

ボヨダユガツゾヌデヤゾルシザベモシブゾホドドノズ
 ゴブプチャダノセゾヤヤコゲボボソエヒロスニヤヘムヤ
 ヨヒツカクヘガレドテヤヨバペデシテゾノハムケムワロ
 デレイメサドノモトブボガビイドレサニリワケナヘロネ
 リャブドギサルヌワイセヘスッバダゾクケドフエメモザ
 コバミュトズビキイミゼロレダムンジセコボパゾラヤケ
 ベエケデクユベノビオバジニトデサミサダヤゼベゴテヒ
 ソヨリキンズケニスグネウヨキサブヨッタチモヤダパル

TABLE 3

日本語への2次の近似度を示す25文字系列例

スペースを含む系列

アタマタ タイナクミヤスル ヘエザイ オヨウナガ
 ダサセイキモノ オ オキョウ テ カタ ヒトル シ
 キンサ ユウ ヨウコロコウサイコ モクシヤク ヤボ
 エ ナイサイ モ オ ト ジンシツナ イ ワ イカ
 ク オケイ パシヨウハラク コウジサケイキユウキケ
 ツカエソウ モロ カ シ ド ト シツカレ ニ ア
 ゴカゼンタエト ノベ オ トコ リコリ オオ パア
 ル トコウイ タ イイ アソクサンセンゴ ノ ナ

スペースを含まない系列

ダカラコショウジンギンドウナリョウセルデスルセンノ
 ジュウシゲンテキズケンシンセキョクゾウシテンシホドコ
 シラメキラシキョウツマデニシコノボクトウシヤクカノ
 ブンテキョウケナゲンシジッシュウボレツゲンスリョウ
 ウソノウネルイガイルクッポノキンシュツムチラレルコウ
 ガイチダレンセイイシヨウヨビトコクテモンエテキタイ
 シンジメセンボシヨウダシユウケイテモノムサンシイダ
 ロシイワタクジブンナルクリタイッコリハジョウサンセ

0次, 1次, 2次の近似を示す文字系列が, それぞれスペースを含む場合と含まない場合の合計6系列作成された。各文字系列は, その系列の長さがそれぞれ10,000字に達するまでその手続きがくり返された。

心理学実験での使用のために, 2種類の表が用意された。一つは, 作成された10,000字からなる各近似文字

TABLE 2

日本語への1次の近似度を示す25文字系列例

スペースを含む系列

ゼガ カトカア ダ ダヨ ツナ タニシ モカチ
 ニ ハクオツヨギン ウシイ ズ チュ ナガ
 カヘウワ ウシボツナ ダカテ ウワコ ウタ ブト
 ニヨシ ナエノ テリガヘイオ ヨ シバジ ヤサ
 オデレフガネ ザリアレ モ チクノクンタイマヒア
 ヨレビンルクシイシタチマツマウデヤ ニオヨクヨ
 コマッゴニトザツキクサデナノズシマコデ ヨシデドサ
 ヤシ シト デ ヤ ノ シシウワヤレ シャ

スペースを含まない系列

スラフウオトアノニテマリテケレテナウケソズジラホチ
 イカアノレニユテリウナレサヨタンナラグニョヘゲハデ
 ノゲオノシカコノヘイシシゾニウクキタヤタシアデテノ
 シミルザイタウンコウキミアクカオユアンガタエノワード
 ミワケレドエツイケデバラフルシリノンエヒシインノル
 ボノタクノウナユリダアニアデナタニシタゲギツカジ
 ルガウトウヒョイケアノミニツニシノシッイノッノウ
 ネユノトウミアレナワスエフフンテダコカカトノデタウ

系列の中から, 25文字単位で, ランダムに選り出したもので, その代表的な例が Table 1, Table 2, Table 3 に示される³。スペースを含まない文字列(近似語)については, 実際上25文字からなるような長いことばは日本語にはほとんど存在しないわけであるが, 利用目的によって適当な長さの系列を25文字系列からサンプリングすればよいわけである。

知覚や記憶・学習等の実験においては, 系列位置効果(Serial Position Effect)が問題とされることが多い。そのような場合に, 同一文字がくり返して出現する系列は, 実験材料として使用できない。このような実験のために, スペースを含まない0次, 1次, 2次の各近似文字系列作成結果の中から, 一系列内には同一文字のくり返しをもたない, 10文字からなる系列を, それぞれ30系列選り出して示したものが Table 4 である。ただし, 拗音字及び促音字を含む系列は除外された。

本研究において作成された, 日本語各近似文字系列の特性は, 各系列作成に使用された, 日本語72文字(73文字)相対出現頻度, 及びDigram相対出現頻度との相関を調べることによって吟味された。すなわち, 0次の近似系列については, 作成された10,000字系列内における各文字の出現頻度を調べ, 72文字(73文字)の分布の一樣性の仮説を, 適合度の検定によつて検定した。1次の近似系列については, 10,000字内での各文

³ より包括的な表は別に作成される予定であるので希望される方は筆者あて申込まれた。

TABLE 4

同一文字のくり返しをもたない 10 文字系列

0 次 近 似 系 列

ハザペルムニノアミチ	ワチゲキノズレブデリ
ムビボゾモキヘギカガ	ネルザチバヌジヘメマ
ヌツナゾトヒカロパネ	ケホリチメサゲガヒン
メウゼアヨボスゲマバ	リヒゼラビブベボシゴ
ニベナヌビスパチビロ	ソイフヘネチヨゲビノ
ゾゲクエトポツウマム	イヤユレアムンビブセ
オラビヒドハエメヌキ	ユマガグヌナドチジビ
ニカペネエダグキサヒ	ナクチブヒトユザパケ
ブバギダネロリントレ	ケロハゴタホニオガヒ
ウノリシベエルトゴレ	レタグケノツモコニズ
シダタロヒキブテヤメ	フレスルノカパゼパワ
ビハバモエヘシフコサ	セツフゾガデゼムバ
スダイウベユバサテ	マニチケガベブビドワ
マホピチアデレロバワ	フブドハコンデバロモ
モウルダリラデヒゼ	ムゾネンチラスハコヒ

1 次 近 似 系 列

カコイサニラワセソハ	アオリルトカヒノワケ
イチユンクトマウカラ	カナンノタテクオウエ
ガチトキンオシタボミ	サテイトクモチコマウ
ラドソガイトギレサン	レウゴイサンルガクド
エウセタキインドレシ	コチルタシダテアグカ
ニチビリノウサオソゴ	ノトカガメタクウヒセ
テランクチトワノユイ	イレリテブセモオクデ
アラマネキイリンテダ	トニダメゴウイコフア
ソリシイノキスエカウ	キゴタセニミジヘイム
バコウゴナキジワケン	ルキインソテタマセク
クネンニノビウイタト	セコテノミワヨタンシ
ヒルエバアマウテゲト	チカシノウンクイヨブ
エナウイステロシトニ	ルイワニリウナタノホ
ヘエカクガウナソニイ	ボムミジイコリエフウ
ウンシノコニチルトモ	イコトデリノシハタル

2 次 近 似 系 列

ルカケイシマスノチロ	タンチカラユウナアデ
タキナオクリツケルホ	テンドウノシナイルコ
ウゼンドレモノリニネ	ヤカニンケイツギヒマ
ルイコトキナガワシラ	イタラスメシンコトウ
サストタシカラケイキ	ブキシツイテワカラタ
タマカンヘエダサセツ	クルビキノゾンセイロ
サンキヒロアルホカラ	モウメルマンカシズイ
クシノリトカルイキミ	メテキタナリダスルシ
センテキフタカケレラ	アタメナムキイカラク
ノシヤタベクセツヨバ	テキナジツタバカンシ
ハマレワルクライテキ	ロイキナクルブンノウ
コトツサエルクワカタ	クムヨサキセンゴソラ
トンガクソノボウセイ	ツマエリトウフクテラ
サエツカケイコナジン	イルマタチオコウヨク
キエソレカオトモノデ	シノウケナワレカンモ

TABLE 5

各近似系列 10,000 字内における単一文字, Digram 出現頻度と, 各系列作成のもとになった原表出現頻度との相関(r) (0次系列については, 単一文字の一様性分布に関する χ^2 検定)

		0 次 系 列	1 次 系 列	2 次 系 列
単一文字	スペースを含まない	—	.997***	.987***
	スペースを含む	—	.999***	.998***
Digram	スペースを含まない	—	—	.899***
	スペースを含む	—	—	.992***

一様性分布に関する適合度の検定 $\chi^2=59.03^{n.s.}$ $df=71$ 適合度の検定 $\chi^2=75.12^{n.s.}$ $df=72$

*** 0.1% 以下の危険率で有意

n. s. not significant

字の出現頻度が調べられ, 系列作成に用いられた 72 文字 (73 文字) の相対出現頻度との偏差積率相関係数が求められた。また, 2 次の近似系列については, 10,000 字系列内における単一文字 72 文字 (73 文字) の出現頻度, 及び Digram の出現頻度が調べられ, それぞれ系列作成に使用された原表の相対出現頻度との偏差積率相関係数が求められた。結果は Table 5 にまとめられている。0 次近似系列に関する適合度の検定は有意を示さず, すなわち各文字の一様性分布の仮説は棄却されず, 他方, 1 次及び 2 次の近似系列に関する各相関係数は, いずれも有意な高い正の値を示している。かかる結果は, 本研究において作成された各近似文字系列は, 系列作成に用いられた日本語単一文字及び Digram 相対出現頻度をよく反映していることを示しており, “方法” の項に述べられた手続きに従って正しく作成されていることをうらづけている。

引用 文 献

- Attneave, F. 1953 Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology*, **46**, 81—86.
- Attneave, F. 1959 *Applications of information theory to psychology*. New York: Holt.
- Garner, W.R. 1962 *Uncertainty and structure as psychological concepts*. New York: Wiley.
- Hirata, K. & Bryden, M.P. 1971 Tables of letter sequences varying in order of approximation to English. *Psychonomic Science*, **25**, 322—324.
- 今榮国晴 1960 日本語 Digram の相対頻度とその特性 心理学評論, **4**, 85—100.

- 今栄国晴 1969 記憶研究法 梅本堯夫(編) 講座心理学 7 記憶 第2章 東京大学出版会 Pp.19—66.
- Miller, G. A. 1967 *The psychology of communication*. New York: Basic Books.
- Miller, G. A. & Frick, F. C. 1949 Statistical behavioristics and sequences of responses. *Psychological Review*, **56**, 311—324.
- Miller, G. A. & Selfridge, J. A. 1950 Verbal context and the recall of meaningful material. *American Journal of Psychology*, **63**, 176—185.
- Miller, G. A., Bruner, J. S., & Postman, L. 1954 Familiarity of letter sequences and tachistoscopic

- identification. *Journal of General Psychology* **50**, 129—139.
- Shannon, C. E. 1948 A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379—423, 623—656
- Shannon, C. E. 1951 Prediction and entropy of printed English. *Bell System Technical Journal*, **30**, 50—64.
- 戸田正直・高田洋一郎 1959 心理学における情報理論適用の現況 心理学研究, **30**, 40—65.

—1973. 11. 5. 受稿—

The Japanese Journal of Psychology
1974, Vol. 45, No. 1, 37—45

資 料

文 字 認 知 に 関 す る 発 達 ¹

大阪教育大学 田 中 敏 隆・岩 崎 純 子
大阪社会事業短期大学 三 木 千 勢

DEVELOPMENT OF THE COGNITION OF LETTERS I

TOSHITAKA TANAKA AND JUNKO IWASAKI

Osaka University of Education

CHISE MIKI

Osaka College of Social Work

This paper aims at examining the following three questions; (a) which is easier to recognize, Hiragana or Kanji, (b) which sex is superior, and (c) which composition of letters is easier to recognize, horizontal or vertical? The method employed is to ask the S to cancel one kind of particular letters in one case and two kinds of particular letters in the other in both of Hiragana and Kanji. The Ss are from the age of 5 to the age of 14 years. (a) Hiragana is easier to recognize than Kanji from the 5 through 7 years old, and vice versa for the 8 years old and above. (b) Females are superior in performance to males. (c) The horizontal composition is easier to recognize than the vertical.

文字の読みに関する学習指導の順序性として、平仮名、漢字、片仮名のいずれが適切であるか、また、これらの文字を横組と縦組のいずれに提示した方が読書行動にとって容易であるかの論争は、国語教育研究者によって従来からしばしば繰返されてきている。最近、文字の読みに関して、幼児が文字に関心を示し始める時期にお

ける学習の順序性としては、平仮名よりも漢字の方が習得しやすいという論さえ主張されている(石井, 1970)。しかしこれは、平仮名と漢字との学習過程を厳密に比較した研究成果からの結果ではない。文字の読みに関する学習の順序性とか、文字の横組と縦組のいずれが読書にとってより容易であるかの問題を吟味するには、大別して①文字の認知、②読書における了解時間と理解度からのアプローチが考えられる。

¹ 本論文は、昭和 46 年度文部省科研費(一般研究 D81533)によるものである。