テキストの計量語彙論的指標はどのような条件で変化するか

山崎　誠（国立国語研究所言語資源研究系）[1]

Under What Conditions does the Textual Index of Quantitative Lexicology Change?

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

要旨

テキストにおける TTR(Type/Token Ratio)の値は､そこに使われている普通名詞の使用状況に大きな影響を受けているとされる（山崎：2012）。本稿は、その続編として、テキストの特徴を表す計量語彙論的な指標の一つである TTR がテキストの一貫性という観点から、どのような条件で変動するかを調査した。『現代日本語書き言葉均衡コーパス』（BCCWJ）から抽出したテキストを利用して、語順のランダム化、テキストの合成、テキストのｎ分割などの方法を用い、それぞれの場合に TTR がどのような変動を見せるかを調査した。これらの観察結果から、テキストの一貫性と TTR との関係を考察した。

１．はじめに

テキストを成立させる条件として一貫性と結束性という概念が提唱されている。Halliday＆Hassan（1976）によると、結束性は文法的結束性（指示、代用、省略、接続）と語彙的結束性（繰り返し、関連語）とに分かれるとされる。結束性は文法的結束性を中心に言語学や言語処理の分野で研究が行われているが、一貫性についてはまだ十分に研究が進んでいるとは言えない。とくに一貫性を計量的言語学的に把握する研究が少ないようである。

ところで、結束性と一貫性の関係について、Widdowson（1978）では以下のように述べている。

結束性が関係するのは，さまざまな文構造上の操作によって命題を結びつけ，テクストを形成するところまでである。それに対し，一貫性は，こうした命題の発語内的機能，つまり，報告・描写・説明などのさまざまな種類のディスコースを作り出すために命題がどのように用いられるかということに関係している．（邦訳『コミュニケーションのための言語教育』p.66）

また、結束性と一貫性の関係について、Widdwoson（1978）は、以下の例を示して説明している。

1. A:　What are the police doing?
　　（警察は何をしているのですか.）
　B:　They are arresting the demonstrators.

[1] yamazaki [at] ninjal.ac.jp

（デモの参加者を逮捕しています.）

2. A: What are the police doing?
   B: The fascists are arresting the demonstrators.
   （ファシストらはデモの参加者を逮捕している.）
3. A: What are the police doing?
   B: I have just arrived. （今来たばかりです.）

（前掲書 p.34）

発語内行為のいかんにかかわらず，文と文の間の命題関係が統語的にも意味的にもはっきりと形態上で示されていれば，そこには結束性(cohesion)があることがわかる．したがって，結束性とは文を通して表現された命題間の明らかな関係のことである．一方，命題そのもののつながり具合は必ずしもあきらかでないにしても，その命題そのものが行っている発語内行為の間に何らかの関連を見い出すことができれば，その発話には一貫性(coherence)があると言える．上にあげたやりとりを，これらの用語を用いて説明してみると，1 と 2 には結束性と一貫性の両方があり，3 には結束性はないが，一貫性はあるということになる．（前掲書 p.35）

結束性は個々の言語要素間の関係としてとらえられるため、比較的計量的測定が行いやすいが、一貫性はテキスト内のどの要素を測定すればよいのだろうか。そのためには一貫性がテキスト内のどこに存在するのかを把握する必要がある。上述の 3.A、3.B の例で考えると、一貫性は 3.A と 3.B との間、すなわち文と文との意味的な関係としてとらえることができる。また、テキストは文の連続体であるので、当該のテキスト全体にわたる属性としてとらえることもできるだろう。

本稿では、一貫性が生じる条件として言語要素の出現順序という性質に注目してそれを客観的にとらえる方法を考える。例えば、出現順序を操作した結果の指標の測定値を、もとの測定値と比べるという方法である。

２．一貫性のタイプ

一貫性は当該のテキスト全体にわたって、それを統括する働きを有すると考えられるが、その分布のあり方に応じて 2 つのタイプに大別することができるだろう。そのための準備的考え方としてテキストの構造をトピック（話題）の集まりとしてとらえる。トピックは形式的には段落の形で実現することが多いだろうが、意味的なまとまりであるので必ずしも段落と対応するとは限らないと考えられる。このような考え方のもとに、一貫性のあり方は次の 2 つのタイプを認めることができる。

A トピック内部の一貫性
B トピックを超えた一貫性

A のトピック内部の一貫性とは、あるトピックの中でその内容に関係するものである。例えば、トピックに合った適切な語を選択することや、ある文の次にその文の内容に関連した文をつなげることなどである。B のトピックを超えた一貫性とは、あるトピック全体をと

らえてそれに関連する別のトピックを次に配置することなど、テキストの構造に関係するものである。一般的には、テキスト全体のテーマに従って適切に構成単位を配列することがトピックを超えた一貫性の表れである。いわば、トピックをメタ的に扱う一貫性と言える。

A のトピック内部の一貫性は、トピックのまとまりということへの関与ということから、語の集合である語彙の計量的な特性、例えば語彙の集中度などに現れるのではないかと推測される。一方、B のトピックを超えた一貫性は、構成単位の順序性を測ることによってその一端が測定できるのではないかと期待できる。

B のトピックを超えた一貫性について 2 つ例を挙げる。

(1)吾輩は猫である。うとうととして目がさめると女はいつのまにか、隣のじいさんと話を始めている。私はその人を常に先生と呼んでいた。こんな夢を見た。

(2)『明鏡国語辞典　第二版』より
みつ - ど【密度】〔名〕❶一定の面積・体積などの中にある量が含まれる割合。「人口の―」
❷内容の充実している度合。「―の濃い議論」❸物質の単位体積あたりの質量。
ミッドナイト［midnight］〔名〕真夜中。深夜。
ミッドフィルダー［midfielder］〔名〕サッカーで、ハーフバックのこと。MF。
（原文は縦書き）

(1)は夏目漱石の小説「我が輩は猫である」「三四郎」「こころ」「夢十夜」の冒頭の文を並べた人工的なテキストである。無関係なトピックが連続するため、一貫性は存在しないと考えられるが、仮に最後の文「こんな夢を見た」をそれ以前の文を統括するものと考えれば、やや牽強付会ではあるがトピックを超えた一貫性があるとも解釈できる[2]。また、(1)の末尾に「これらは夏目漱石の作品の冒頭文をつなげたものである。」を付け加えれば、そのことで、トピックを超えた一貫性があると解釈できる。

(2)は国語辞典の一部であるが、連続する見出しは五十音順に並べられているため、それらの間には一貫性はないのが普通である。ただし、その五十音順に並べるという配列規則がここでは、トピックを超えた一貫性であると考えることができる。(2)のような一定の配列のもとに、並べられたテキストを本稿ではリストタイプのテキストと呼ぶことにする。リストタイプのテキストは、辞書がその典型であるが、箇条書きなども含まれる。例えば、『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）では次の表 1 のような例が挙げられる（山崎 2010）。表 1 は短単位で計った 1 語あたりの平均使用度数（n/k 値）の低いサンプルを挙げたものであるが、それらはリストタイプのサンプルであったことが指摘されている。このことからトピックを超えた一貫性は語彙の計量的指標に反映される可能性があることが示唆される。

[2] 3 文目の「その人」が 2 文目の「隣のじいさん」を指すと解釈すればそこに語彙的結束性が存在するとも考えられる。

表１　1語あたりの平均使用度数（n/k 値）が低いサンプル

n/k 値	サンプル ID	NDC	出典名	著編者	文章のタイプ
1.5198	PB17_00159	7 芸術・美術	淡路人形浄瑠璃　伝統芸能　国宝重要文化財等保存事業		リスト（用語集）
1.5771	PB18_00010	8 言語	漢字・仮名・記号テキスト	佐々木光朗	リスト
1.5906	PB2n_00001	分類なし	日本を伝える	梅澤実（監修）	リスト（図録）
1.6018	LBe2_00037	2 歴史	昭和家庭史年表 1926～1989	家庭総合研究会	リスト
1.6683	LBj8_00006	8 言語	日本語キーワード英語表現辞典 日本語の発想で引けて英語表現が豊かになる辞典 名詞編	三省堂編修所	リスト
1.6814	LBo2_00009	2 歴史	1946-1999 売れたものアルバム	Media View	リスト

３．方法とデータ

前節で一貫性は 2 つのタイプに分けることができ、その特徴を利用して一貫性の測定の方法が考えられることを示した。そのことを実現するために、一貫性のないテキストを 2 種類の方法で人工的に作り、それと元のテキストを比べるという方法をとる。その際の比較のための指標は異なり語数の延べ語数に対する比である TTR（Type/Token Ratio）を用いる。TTR は 1 語あたりに平均使用度数の逆数であり、語彙の多様性の指標とされ、コーパス言語学では TTR がよく用いられる。具体的な方法は次の 2 つである。

(3)トピック内部の一貫性については、語をランダムに入れ替え、n-gram による組み合わせを比べる。

(4)トピックを超えた一貫性については、テキストの前半と後半とをそれぞれ別のテキストから選び、トピックを合成して人工的に一貫性を低下させたテキストの TTR 値を元のテキストの TTR 値と比較する。

データは BCCWJ の図書館サブコーパス（LB）から無作為に選んだ 22 テキストである。ただし、TTR 値は延べ語数に影響を受けるため、本発表では短単位・可変長部分が延べ語数で 2,000～2,100 語の範囲に限定している。なお、選択の際は、分野を考慮して各 NDC（図書分類）と分類なしとから 2 テキストずつを選んでいる。

４．考察１

４．１　語順のランダム化

テキスト内に現れる語が一定の順序で現れる通常のテキストと、語順をランダムに並べ替えて一貫性を低下させたテキストとについて、2-gram（＝2 語の連続。但し記号は除外する）の TTR 値を比較した。語順のランダム化の例を(5)(6)に挙げる。(5)のテキストをランダム化したのが(6)である。

(5)吾輩は猫である。名前はまだ無い。どこで生れたかとんと見当が付かぬ。

(6)。見当吾輩。はである生れ名前かどこたぬ付かが。は無いとんとまだで猫

結果を図 1 に示す。ランダム化したテキストでは，元のテキストに比べて 2-gram の TTR

値が有意に高くなることが確認された（t=-20.93,df=21,p<0.001）。

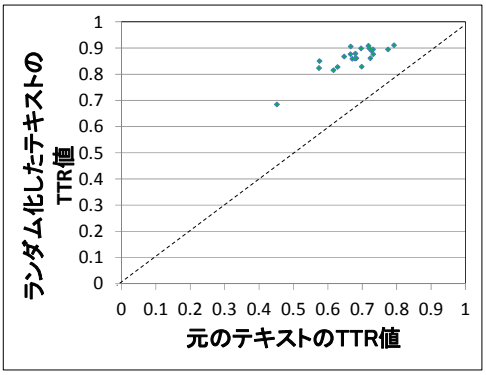

図 1 ランダム化したテキストの TTR 値の増減

４．２　テキストの合成

22 サンプルについて、それぞれのサンプルの前半と別のサンプルの後半を合併した人工的なテキストを作り、その TTR を計測した。全部で 462 のテキストが作成される[3]が、そのテキストの TTR を元となった 2 つのサンプルの TTR の平均値と比較する。そうすると、全 462 テキスト中、元となった 2 つのそれぞれの TTR の値と比べると値が増加しているものが多いが、減少しているものも見られた。ただし、元となったテキストの TTR の平均値と比べると 462 テキスト中 461 テキストで人工的に作成したテキストの TTR の値が増加していることが分かった（平均で 0.028 増加）。結果を表 2 に示す。

表 2　合成テキストの TTR 値

比較する対象	TTR 値が増加	TTR 値が減少
テキスト 1 の TTR 値	359	103
テキスト 2 の TTR 値	370	92
上記 2 つの平均	461	1[4]

実際の分布の様子を図 2 に示す。図 2 の横軸は、1 つめ（前半）のファイルにおける、元の TTR の値と合併したファイルの TTR の値との差であり、縦軸は、2 つめ（後半）のファイルにおける、元の TTR の値と合併したファイルの TTR の値との差である。元のテキストと□□的に作成したテキストの TTR との差には負の相関があることが分かる。

なお、テキストを 3 分割した場合は全 9,241 例の合成テキストのすべてにおいて人工的に合成したテキストの TTR 値がそれを構成する 3 つのテキストの TTR 値の平均を上回った。（平均 0.043 増加）。

[3] 同じテキスト同士の合成は除外したので、22×21 ファイルが対象となる。

[4] NDC8(LBs8_00014)と NDC6(LBb6_00012) の組み合わせである。

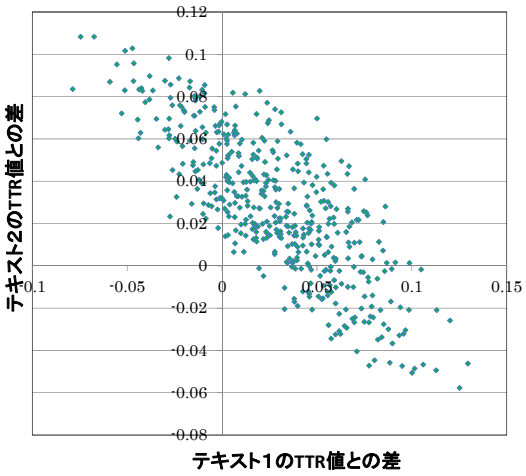

図 2 合成テキストの TTR 値の増減の分布

以上 2 つの事例の結果により、一貫性が低くなると語彙的指標である TTR の値にその影響が現れる場合があることが確認された。しかし、その逆である TTR の値が低くなれば、一貫性が低くなるかこの方法では把握できない。

５．考察２

本節では､テキストをいくつかの区間に分割した場合の TTR 値の変化の様子を観察する。単純に n 分割したもの、n の剰余系により分割[5]したもの、ランダムに n 分割したものの 3 つの人工的テキストについて TTR を計測する。

データは、図書館サブコーパス（LB）の可変長部分の延べ語数（空白・補助記号・記号を除く）が 5,000～5,100 語である 252 ファイルである[6]。

分割数に応じた TTR の値の変化を図 3 に示す。図 3 から、単純に分割した場合よりも

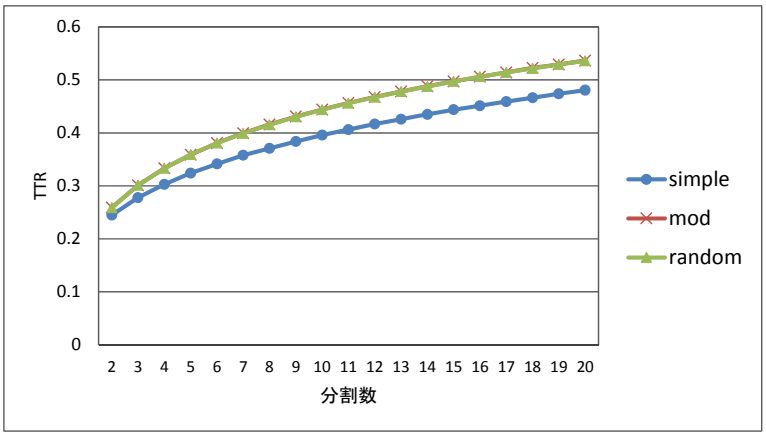

図 3　分割数に応じた TTR の値

[5] テキストを構成する語に先頭から番号を付け、それらを n で割った余りが同じものを一つの語彙として分割したもの。たとえば、2 分割の場合は、偶数番目の語の集合と奇数番目の語の集合とに分かれる。

[6] 各レジスターの内訳は、LB93 個、OB17 個、OL7 個、OP4 個、OT1 個、OW19 個、PB99 個、PM11 個、PN1 個である。

剰余系による分割およびランダムに分割した場合のほうが TTR が高いことが分かる。また、剰余系による分割とランダム分割とには差がないことも見て取れる。図からは、単純な分割と剰余系・ランダム分割との TTR の差[7]は 0.05 くらいに収束しているように見える。

次に n 分割した n 番目の区間の TTR の特徴を見よう。

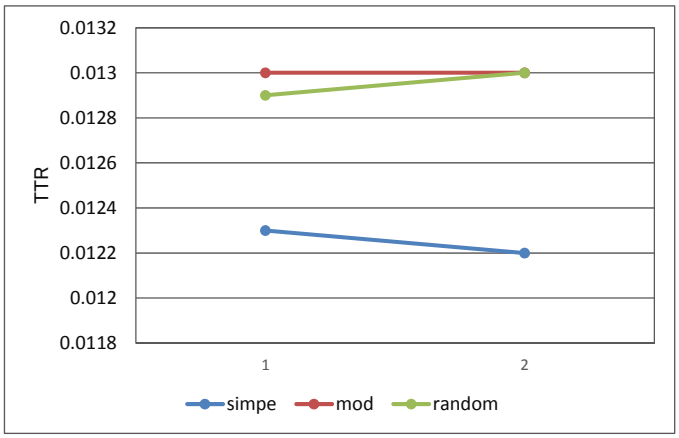

図 4　分割区間ごとの TTR の値（2 分割）

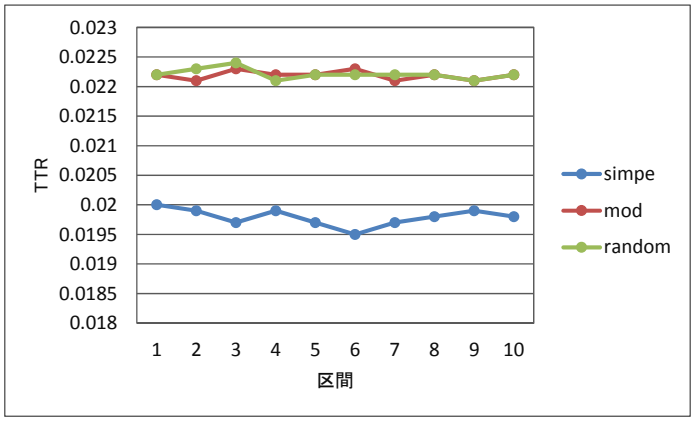

図 5　分割区間ごとの TTR の値（20 分割）

図 4 は 2 分割、図 5 は 20 分割の例である。ここでも、単純な n 分割の場合と剰余系による n 分割、ランダムな n 分割との関係は図 3 と同様である。各区間の TTR の値はランダムに上下しているようであり、特定の傾向は見出しにくい。ただし、区間 1 と区間 2 との関係だけを見てみると、単純な n 分割は、2～20 分割のすべての例において、区間 1 よりも区間 2 の TTR の値が低かったのに対して、そのような傾向を見せるのは、剰余系による n 分割では 9 個、ランダムな n 分割では 11 個であった。このことは、文脈が維持されている場合、冒頭部分から一定の分量の区間は、語の繰り返しが多いことを示唆しているものと思われる。

7 シンプルな分割の TTR から、剰余系による分割の TTR＋ランダムに分割による TTR÷2 を引いた値。

６．まとめと今後の課題

本稿ではテキストの計量語彙論的指標である TTR の値がどのような条件で変化するかを考察した。とくにテキストの一貫性という観点から、文脈がそのまま維持されている場合と文脈が破壊されている場合を比較するという手法で TTR の値を観察した。その結果、文脈を維持せずに人工的に合成したテキストは総じて TTR の値が高くなることが確認された。今回の考察では、剰余系による分割とランダムな分割との間には TTR の差が見いだされなかった（見込みでは幾分かの差があると想定した）。今後の課題としては、文脈がどの程度維持されていれば、TTR の値が維持されるのか、新たな条件を模索することが挙げられる。

謝　辞

本稿は 2013 年 7 月 21 日に行われた、国立国語研究所基幹型プロジェクト「コーパス日本語学の創成」の共同研究発表会で行った発表「テキストの一貫性と計量語彙論的属性との関係」および山崎（印刷中）に加筆・修正したものである。

本研究は国立国語研究所の共同研究プロジェクト「コーパス日本語学の創成」による研究成果の一部である。データとして利用した BCCWJ は、国立国語研究所のプロジェクト及び文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度、領域代表者：前川喜久雄）による補助を得て構築したものである。

参考文献

Halliday, M.A.K. and Hasan, R.（1976）Cohesion in English. London:Longman.（邦訳『テクストはどのように構成されるか』、大修館書店、1997 年刊）

Widdowson, H. G.（1978）Teaching Language as Communication. Oxford:Oxford University Press（邦訳『コミュニケーションのための言語教育』研究社出版、1991 年刊）

山崎誠（2010）語の平均使用頻度に現れるテキストの特徴、特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ（研究成果発表会）予稿集 pp.5-14.

山崎誠（2012）Type/Token Ratio と品詞との相関,修剛（編）『新時代的世界日語教育研究』pp.59-64、北京：高等教育出版社

山崎誠（印刷中）テキストの一貫性を表す語彙的指標について、『日語研究』10、北京：商務印書館