ID	KK280601
種別	調査報告
タイトル	文章中における名詞の反復の量的様相 —Type-Token Ratioを利用した分析—
Title	The Quantitative Condition of Repeated Nouns in Writing: A Type-Token Ratio Analysis
著者	鯨井　綾希
Author	KUJIRAI Ayaki
掲載号	28巻6号
発行日	2012年9月26日
開始ページ	211
終了ページ	225 （英文要旨：p.230）
著作権者	計量国語学会

計量国語学第二十八巻第六号〔Mathematical Linguistics, Vol.28 No.6〕2012 年

調査報告

文章中における名詞の反復の量的様相

—Type-Token Ratio を利用した分析—

鯨井　綾希（東北大学大学院文学研究科）

キーワード：反復語，新出語，Type-Token Ratio，BCCWJ

１．はじめに

文章を文章たらしめる言語的要因の一つとして，同じ情報が繰り返し提出されることによって生じる意味の連鎖とまとまりの形成を挙げることができる．日本語を対象とした文章論・談話論においても繰り返された語句に注目した研究があり，馬場（1986; 2006）や砂川（2000; 2005），高崎ほか（2007）などが代表的な研究として挙げられる．しかし，馬場（1986; 2006）や砂川（2000; 2005）は，文章・談話の内容上の中心的な命題である「主題」の実態解明が主たる目的である．重要性の高い語句の反復が「主題」の形成過程に関与していることから，「主題」の分析手段として語句の反復が注目されたのであって，語句の反復そのものが問題とされているわけではない．高崎ほか（2007）では語句の反復そのものに注目して，それらが文章内における展開にどのように関わるかを考察しているが，反復された語句の量的側面は分析されておらず，その多寡がどのような形で文章上の特徴の変化と関わっているのかについては明らかにされていない．

しかしながら，語句の反復は既に述べていることを改めて提示するという点で冗長さを助長する存在であるため，その使用には量的な幅と限度が生じているはずである．その点で，語句の反復の実態を把握するためには，その量的な様相を明らかにすることが重要であると考えられる．加えて，値の変化が具体的にどのような文体と関係しているのかを明らかにすることで，反復語の文章上における位置づけも明確化されると考えられる．

そこで本稿では，文章中における同一語，特に同一名詞の反復（以後，反復語と表記．名詞に限った理由は後述）を取り上げ，その多寡を定量化し，既に述べた語句を改めて述べるということが文章中においてどの程度まで許容されるのかを明らかにする．また，その多寡が具体的にどのような文体的な特徴と関わるのかという点について考察し，反復語の文章中での利用のされ方の一端を明らかにする．さらに，以上の目的を達成するために，本稿では Type-Token Ratio（TTR）という計算法を利用して反復語の多寡を分析する．

２．分析上の手続き

２．１　反復語の量的様相を分析するための TTR の利用法

TTR とは文章中の異なり語数を延べ語数で割った値のことであり，以下の式によって

KUJIRAI Ayaki (Graduate School of Arts and Letters, Tohoku University) — The Quantitative Condition of Repeated Nouns in Writing: A Type-Token Ratio Analysis —

表すことができる．

$$TTR = V（N） / N$$

［V（N）：異なり語数，N：述べ語数］

TTR は，先行研究では主に語彙の豊富さの指標として捉えられている（小野ほか 2007，金 2009，田島ほか 2009）．

一方で，Youmans（1991）や Stubbs（2002）では TTR を語彙的多様性の尺度としてだけでなく，文章中における新出語と反復語の割合を示すものとして解釈し，文章・談話単位での分析に利用している．TTR が新出語と反復語の使用状況を量的に表す手段となりうるのは，異なり語数に用いられる見出し語の追加がその語の最初の使用によってのみ行われ，二回目以降の使用は延べ語数に用いられる単位語として追加されることにより，TTR の値が文章全体に対する新出語の使用率となるためである．新出語ではない二回目以降に使用された語は反復語であるため，TTR は新出語の使用率であると同時に，反復語の使用率を把握する指標としても利用できる．

Stubbs（2002）では TTR を新出語と反復語の選択の問題と関係づけ，文章・談話分析に利用することの有効性を指摘しているのみであるが，Youmans（1991）では TTR を利用した具体的な分析が行われている．ただし，Youmans（1991）では新出語と反復語の使用が一つの作品の展開においてどのような変動を見せるのかという点が主たる関心であり，分析に利用している文章の数が少ない．そのため，大量の文章を取り上げた上で，一般的に新出語と反復語の使用の割合が文章構成上においてどのような範囲に収まるのかという点については明らかになっていない．

したがって本稿では，TTR の結果を反復語の使用実態を把握する手段として利用するとともに，『『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）』（以下 BCCWJ と表記）という大規模なコーパス[1]を用いることで，多くの文章において反復語がどのように利用されているのかを分析する．

２．２　分析対象とする語の選定

本稿では，新出語と反復語の割合を分析する際に，文や文章における具体的な内容を意味する名詞[2]を取り上げた．文と文を結びつける意味的な依存関係を「結束性」と呼んで分析した Halliday & Hasan（1976）では，語彙的な結束性を形作る「再叙」，すなわち反復も取り上げており，その分析対象として名詞を多く扱っている．Halliday & Hasan（1976）においては，「指示」や「代用」，「省略」によって形作られる結束性も名

[1] 2012 年 5 月現在は既に正式な『現代日本語書き言葉均衡コーパス』が公開されており，モニター公開データに含まれないデータを加えたコーパス全体像の内実は国立国語研究所コーパス開発センター（2011）に詳細が記されている．

[2] 名詞のうち，形式名詞として使われそれ単独では具体的意味を持たない「物（もの）・事（こと）・頃（ころ）・為（ため）・間（あいだ）・内（うち）・時（とき）・程（ほど）・由（よし）・後（あと）・項・まま・せい・はず・かた・ふし・ところ・ゆえ」と，数詞，ならびに名詞に位置づけられるが単独で用いられず関係節を構成するような用いられ方をする「以上」・「以前」・「場合」・「辺り（あたり）」を除いている．同様に，文法的な使われ方をして具体的な意味をそれ自体で持たない代名詞も対象から除いた．

詞との関わりが強い．その他の品詞もそれらの結束性と関わることはあるが，どの結束性にも広く見られるというわけではない．このことから，文と文の意味的関係性を作る上で，名詞が汎用的な役目を担っていると考えることができる．

名詞による反復が文と文を意味的に関連づけていることは，次のような用例によっても見て取れる．

（1）しかし，手順を踏むと成績が上昇するというこの事実はまた，コンピュータの記憶とはまったく異なります．コンピュータは，それがたとえ多段階で複雑な手順でも，試行錯誤することなく一回の記憶で完全に習得することができます．（池谷裕二 2002『高校生の勉強法-最新脳科学が教える-』ナガセ，サンプル ID:PB2n_00075）

用例（1）において，下線部の「コンピュータ」を含む文は，それ以前の文に波線部「コンピュータ」があるということを通して，お互いの意味上の関係性が見出される．

反復そのものは名詞だけでなく，述語部分の内容語にも存在する．例えば，次の用例では「伸ばす」という動詞が連続して用いられている．

（2）足を肩幅より広めに開いてまっすぐに立ちます．両手を組んで頭上に伸ばし，手首を回転させて手のひらを上に向けます．そのまま両腕と背骨をぐーっと伸ばします．（鈴木正成（監修）2001『ダンベル体脂肪ダイエット』日本文芸社，サンプル ID:PB15_00060）

ただ，用例（2）の波線部と下線部における「伸ばし」は，それぞれ異なった対象における一回的な事態の記述である．「伸ばす」という動詞を含む文によって表された二つの事態の間には関係性を見出すことができるが，二つの「伸ばす」という語自体には意味上の関係性を想定することができない．

以上のことから，本稿では，様々な結束性のあり方と関わる点で汎用性があり，それ自体で意味上の関係性を見出せる名詞を分析対象として扱い，名詞の使用における TTR を計って，その反復の量的様相を分析した．

２．３　分析資料

本稿では，分析に利用する資料として BCCWJ のうち，「出版（生産実態）サブコーパス」と「図書館（流通実態）サブコーパス」が含まれる「書籍」のデータを選んだ[3]．BCCWJ そのものには，モニター版であっても「書籍」データ以外に Web 上の質問サイトである「Yahoo!知恵袋」や，話し言葉の書記記録である「国会会議録」，政府の報告書である「白書」といったデータが存在する．しかし，それらはコーパス収録上の区分としては「特定目的サブコーパス」に位置づけられ，「出版（生産実態）サブコーパス」「図書館（流通実態）サブコーパス」という，書き言葉の社会的な流通上の「均衡性」を目的に集められたデータとは別個に集められたデータである．「特定目的サブコーパス」は，流通上大きな比率を占めなかったとしても，分析可能な量までデータを大きくしたものであるため，それらを含めると書き言葉の均衡性が崩れる可能性がある．そのため，本稿では分析対象として「特定目的サブコーパス」に含まれるデータを除き，書籍データのみを選んだ．

[3] 利用したデータは，ディスク内のフォルダ構成においては BCCWJ2009_DATA/Plain/BK/以下にある．

２．４　書籍データの絞り込み

BCCWJ に含まれる書籍データはサンプル ID ごとに分けた場合 10423 ファイルあるが，その中にも分析対象として適切なものであるとみなしにくいものが存在する．それらについては，ファイルの観察を通じて取り除いた．本稿で分析対象としなかったものは，以下の表 1 に示したような特徴が全体の半分程度に見られたファイルである．

表 1：文章ファイルから取り除く条件

(a)	語句や文の羅列・箇条書きであり，その間に連続的なつながりがない
(b)	人名解説や事項解説等，独立した短い情報が繰り返し提示される
(c)	引用文の過剰な脱落や注釈・図表タイトルの過剰な割り込みがある
(d)	写真の解説等，ファイル内の文章のみでは理解を完結できない
(e)	アルファベット等の外国語表記や漢文・擬古文・方言等によって書かれている
(f)	対話形式で書かれ，話者の名前が注釈的に繰り返し現れる
(g)	散文ではなく，詩や歌である
(h)	使用された名詞が50語未満である

条件（a）や（b）のような，箇条書き的で文と文の連続性に欠けるものは，文と文の意味の連続性の中で全体の内容を一つにまとめていくことが行われないものであるため，自然な文章とは認めにくい．また，条件（g）で取り上げた詩や歌も，内容的な連続性や全体としての意味的なまとまりよりも技巧性が優先されるため，自然な文章とは異なると考えられる．

以上については内容的な問題であるが，その他，構成上の問題や形態素解析上の問題として以下のようなものも除外した．まず，条件（c）や（d）のように，データの電子化の過程で元の文章の構成が明らかに崩れているものは，電子化の前であれば文章として認められるものであっても，意図せず構成が変化しているものであるため，除外対象とした．条件（f）に示したような注釈的な情報の反復もまた，その場で話されて形成されたであろう文章のまとまりとは別個の役割として用いられるものであるため，除外対象とした．

また，アルファベットや擬古文・漢文などで書かれた条件（e）に当てはまる文章は，形態素解析上の誤解析が目立ち，それ自体は文章として自然であっても分析上適切な結果を得られないため，除外した．

以上の条件（a）から（g）は，当てはまるファイルの文章の一部を以下に抜粋した．なお，条件（h）は，条件（a）から（g）に比べると，より操作的なものである．反復語の使用実態を文章という枠組みの中で把握するためには，文章としての話のまとまりを形成するだけの語数が必要であると考えられる．本稿では，そうしたファイルの大きさの最低単位を分析対象とする名詞 50 語以上と定めた．名詞 50 語は，全体の文字数としてはおおよそ 400 字詰め原稿用紙 1 枚分に相当する．よって，ファイル全体の名詞数が 50 語に満たないものは，本稿では扱わなかった．なお，名詞に注目したのは，2.2 節で取り上げた分析手続き上の理由による．ファイルを一定の単位で区切る理由は 3 節で述べる．

・（a）の例

（3）エビ豆，シジミの山椒煮，シジミのショウガ煮，ガンゾの煮付け，イシガイの

佃煮，イサザの佃煮，
（滋賀県教育委員会文化財保護課（編）『滋賀の食文化財-滋賀県選択無形民俗文化財記録作成（平成 11 年度-平成 12 年度）-』滋賀県教育委員会，サンプル ID:PB13_00617）

用例（3）は佃煮の名前が羅列されるだけであり，文の形式とその連続という文章の形が成り立っていない．

（4）これはわたしのほんです．
それはあなたのほんです．
これはたなかさんのかさです．
（片桐ユズル 1990『はじめてのにほんご』大修館書店，サンプル ID:LBe8_00003）

用例（4）は同一の文型を語を入れ替えて表現しているだけであり，それぞれの文に内容的なつながりがあるわけではない．

・（b）の例

（5）安部公房【あべ・こうぼう】
１９２４年東京生まれ．１９５１年『壁 - カルマ氏の犯罪』で芥川賞受賞．不条理やシュールリアリスムな作風が話題となり，一躍人気作家となった．１９７３年，演劇集団「安部公房スタジオ」を結成，主宰．他の代表作は『砂の女』『緑色のストッキング』など．１９９３年没．
大江健三郎【おおえ・けんざぶろう】
１９３５年愛媛生まれ．東京大学文学部在学中の１９５７年に書いた『奇妙な仕事』でデビュー．１９５８年『飼育』で芥川賞受賞．
（堀越正光 2005『東京「探検」-現役高校教師が案内する東京文学散歩-』宝島社，サンプル ID:PB59_00422）

用例（5）は，「安部公房」と「大江健三郎」という別個の人間の解説が短い間隔で羅列的に行われている点で，その間の必然的な内容的連続性が欠けている．

・（c）の例

（6）近所に幽居している，もとは渡辺氏といった得船入道がいたので，その人と茶を飲み菓子を食べ，あれこれと話をした．
d/
いつの時代も人々は同じことをいって嘆くものらしい．ああだこうだと話をしてきたのだが，結局，「がぜぼ谷」についてはわからない．
d/
と当人も記している．
（群ようこ 2003『浮世道場』講談社，サンプル ID:PB30_00040）

用例（6）は，本来書かれている部分の脱落が記号「d/」によって示されている．（6）であれば「d/」部分に挿入されていたであろう「話」の具体的な内容が脱落することによって，内容の連続性が失われている[4]．

[4] この「d/」については，ディスク内の BCCWJ2009_DATA/Plain/README.pdf において「文字入力対象外で削除された要素の存在を表す(例えば，図表)」とされている．この説明では，「d/」によっ

（7）今日の代表的な理論家としてはＪ・ベアード・キャリコットやホームズ・ロールストン三世らが挙げられる．キャリコットは，自分の思想を単に環境倫理と呼んでいる(（9）).

（9） j ・ベアード・キャリコット「動物解放論争―三極対立構造」『環境思想の系譜三―環境思想と多様な展開』，五九ページ.

「自然の権利」の中でも，一九七〇年代は動物解放が主流であったが，八〇年代には生態系保存論が「自然の権利」を代表する思想となっているようである.

（海上知明 2005『環境思想-歴史と体系-』NTT 出版，サンプル ID: PB55_00083）

用例(7)は文と文との間に注釈が挟まり，文章としての連続性を遮っている.

・（d）の例

（8） 5　張るのに失敗して，弦がはねて歯に当たって怪我をしたようです（左）.

この人は足を怪我したようです（右）.

（東京新聞大阪本社(編)1992『騎馬民族の謎』学生社，サンプル ID:LBg2_00005）

用例（8）は，「エレクトラム製の壷」とそこに描かれた絵が読者にとっても視覚的に把握できることを前提に書かれている．その証拠に，「この人」の「この」は，元々の文章では視覚的に示されているであろう図内の写真を指しており，文章内の言語的要素を意味しない現場指示的な用法であると考えられる.

・（e）の例

（9）因是観之，我桑野村殖民所ノ如キ，之ヲ北海道殖民所ニ比シテ大小異ナル所アリト雖共，殖民所タル所以ニ至リテハ則チナリ.

（立岩寧 2004『大久保利通と安積開拓-開拓者の群像-』青史出版，サンプル ID: PB46_00005）

用例（9）のような例は，形態素解析した結果に誤解析が多くなる.

・（f）の例

（10） 平岩

それは絶対に揺れないほうがいいですもの．私は「ロッテルダム」以来，少し揺れたなと思うと用心してプロムナードデッキを歩くようにしています.

阿川

僕が早起きして歩いていると，もうデッキを歩いている人がいて，見ると平岩さんなんだな.

（平岩弓枝 2001『幸福の船』新潮社，サンプル ID:PB19_00029）

用例（10）では，発言者を明示化するための注釈的な情報として話者の名前が繰り返し提示されている.

・（g）の例

（11）わたしのごたいの日毎なる衰え

て示される部分に文字要素が含まれないかのようにも思えてしまう．実際には，図表の省略だけでなく，引用文のような文章の一部の省略が「d/」によって示されていることがある．例えば用例（6）の「d/」は図表があったわけではなく，本来であれば『紫の一本（ひともと）』という，江戸の地誌に関する本の一節が引用されている部分である.

それも黄ばむだこのころのうすれ日のように
久かたわたしは
老づるのようなやつれをまとふて　うら山に降りたつ
（木村林吉 2001『眼のない自画像-画家幸徳幸衛の生涯-』三好企画，サンプル ID:PB17_00051）

用例（11）のような詩や歌においては，技巧的な表現が目立ち，文章による情報伝達そのものに必ずしも重きを置いていない．

条件（a）から（g）の特徴を持つファイルは 1475 ファイルあり，それらを取り除いた結果，書籍データ内の 10423 ファイル中 8948 ファイルを分析対象とすることができた．

３．分析結果

2 節の手続きによって絞り込んだ書籍データの中で対象となる名詞の TTR を計り，新出語・反復語の構成状況を検討した．始めに，分析対象となる各ファイルの大きさの概略を文字数換算で表 2 に示す[5]．

表 2：ファイルの大きさ

	文字数
最小値	430
第1四分位	2537
中央値	4392
平均値	5189
第3四分位	7432
最大値	16276

各ファイルの分析に際しては，本稿で対象となる名詞が 50 語用いられるごとに文章を区切り，その中での TTR を計った上で，それらの平均をファイル内の文章における新出語と反復語の比率として認定した．

これは，TTR に文章が長くなるほど値が小さくなるという傾向があり（Stubbs 2002），そのまま計算するとファイルごとの大きさの差の影響を受けてしまうためである．また，ある語と再び用いられた同一語との文章中での距離があまりにも離れている場合，後者を反復語と捉えて良いのかは疑問であり，一定の距離ができた時点で反復している語とそれ以前の同一語との関係づけを解消した方が良いと考えたためでもある．なお，50 語ごとに計算した際には余剰分の語が出てしまうが，それらは計量に利用できないため切り捨てた．

上述の手続きに基づいて TTR を計算した結果を表 3 に示す[6]．

---

[5] 集計は Perl の length 関数を用いて行った．

[6] なお，各語の認定や品詞分解には，形態素解析エンジン MeCab（Ver.0.98）と，形態素解析用辞書 UniDic（Ver.1.3.12）を用いた．したがって，本稿は「語」単位ではなく，UniDic の認定単位である「短単位」と呼ばれる分割基準に則って計算されたものである．分割単位として「短単位」を採用

表 3：TTR の値

最小値	0.39
第1四分位	0.727
中央値	0.776
平均値	0.767
第3四分位	0.82
最大値	0.96

表 3 を用いて TTR の値の幅を見ると，最大値が 0.96，最小値が 0.39 となっていることが分かる．TTR は新出の語の提示率であるから，新出語は文章内において 39%〜96%という比率を占める．言い換えれば，文章中において二回目以降の使用となる反復語は，全体の 4%〜61%程度を占めることとなる．

また，第 1 四分位の値が 0.727，第 3 四分位の値が 0.82 であることから，この範囲に収まる文章ファイルが全体の半数を占めることが分かる．さらに，中央値は 0.776，平均値は 0.767 である．

この結果をより詳細に見るため，TTR の値に対する全体の分布状況を調べたものが表 4 である．なお，TTR の値の境界は x 以上 y 未満で割り振っている．

表 4：TTR の値の分布

TTR	ファイル数	割合(%)
0.35–0.40	1	0.01
0.40–0.45	2	0.02
0.45–0.50	14	0.16
0.50–0.55	54	0.6
0.55–0.60	162	1.81
0.60–0.65	364	4.07
0.65–0.70	867	9.69
0.70–0.75	1729	19.32
0.75–0.80	2486	27.78
0.80–0.85	2305	25.76
0.85–0.90	863	9.64
0.90–0.95	96	1.07
0.95–1.00	5	0.06
合計	8948	99.99

表 4 によれば，0.70 以上 0.85 未満の範囲のそれぞれの区間に含まれるファイルの数が多く，その範囲に対象とするファイル全体の 72.86%が含まれており，多数を占める．この表 4 の分布状況から，文章は 75%〜80%程度の新出語と 20%〜25%程度の反復語とい

した理由は注 7 で述べる．また，計算はプログラミング言語 Perl（Ver.5.10.1）を用いた自作のプログラムで行った．基本統計量の確認には統計解析言語 R（Ver.2.10.1）を利用している．

う割合で構成されるのが一般的であると考えられる．

以上から，文章中における反復語の割合はおおよそ次のような範囲となる．

（12）・反復語の割合の最大範囲

4 % ~ 61 %（0.39 ≦ TTR ≦ 0.96）

・反復語の割合の一般的範囲

15 % ~ 30 %（0.70 ≦ TTR ≦ 0.85）

また，表 4 より，TTR が 0.5 未満，すなわち反復語が文章内の名詞の半分以上を占めるものは 0.19%と非常に少ない．また，TTR が 0.9 以上，すなわち新出語が文章内の名詞の 90%以上を占めるものも 1.13%と少ない．

反復語の使用は話題のまとまりを作る重要な要素ともなっているが（高崎ほか 2007），以上の結果から，話題のまとまりを作り出す際の文章中における語彙的性格として，新出語の提示量の増加はかなりの程度まで許容できる反面，新出語のみで全体の 90%以上を構成することは難しく，また，反復語の提示量を過度に増やすことには強い制限が加えられており，反復語が文章内で半分以上を占めることはほとんどないと言える．

4．分析結果の考察

4．1　上限・下限が異常値かどうかの確認

始めに，本稿の分析結果で上限・下限の値を示したファイルの内容を確認するために，上限・下限に位置づけられる文章を抜粋し，その具体的な中身を観察する．

4．1．1　反復語が極端に少ない文章（TTR の値が 0.95 以上）の実例

表 4 において 0.95 を上回り TTR の値が極端に大きい 5 ファイルを以下に挙げる．これらは新出語の使用が最も多く，反復語をほとんど用いていない文章である．以下にその文章の一部を抜粋する．なお，引用中では新出であってもファイルとしては反復語であることもあるため，ここでは反復語と同一の語は全て下線によって示している．また，出典末尾には TTR の値を示した．

（13）　並んでいる平野さんに，つつかれるまでもなく，白いウエディングドレスを着けた辺見さんは，形容のしようもなく，目を見張るばかりでした．ほとんど素顔かと思われるその顔が，緊張して澄みきり，白い冠を受けて，犯し難い気品に満ちています．

カメラを預けられていた僕は，ファインダーが曇り，シャッターを押す指が震えました．

（米長保 2003『アマンドの粒』鳥影社，サンプル ID:PB39_00489, TTR:0.96）

用例（13）は，語り手の「僕」が片思いをしていた「辺見さん」の結婚式に出席している場面である．反復語がなく，新出語が次々と提示される．

（14）　競争の気持ちが高まるといずれは攻撃に出る．しかし露骨な攻撃は得策ではない．かえって損を招きやすい．第一に当方の精力をみだりに浪費して疲れる．第二に世間から嫌われる．世の人は常に物見高く騒動を好むくせに，一方また判官贔屓で弱い者に声援をおくるから，勢い盛んに突き進んでいる方を憎む．あからさまな攻めこみは避けねばならぬ．

（谷沢永一 1995『人間通』新潮社，サンプル ID:OB4X_00100, TTR:0.96）

用例（14）は人との争いに勝つ方法を書いた文章である．引用中では「攻撃」が反復語として一度用いられるのみで，基本的に新出語を用いて文章を作っている．

（15） 郷里へ戻り，私は，父が毎朝私のために近くの神社へ参拝に行っていることを知りました．母は，食欲が低下している私に一粒でも多くの栄養を摂らせたいと，心を尽くして私の好物を作ってくれるのです．友人たちが次々とやって来ては，何か俺たちにできることはないか，と尋ねてくれます．
（井村和清 1980『飛鳥へ，そしてまだ見ぬ息子へ-若き医師が死の直前まで綴った愛の手記-』祥伝社，サンプル ID:OB1X_00183, TTR:0.96）

用例（15）は病気の「私」が病院を退院し，帰郷した場面の一節である．「私」を基準にした出来事が新出語の多用の中で語られている点で用例（13）に文体的に近い．

（16） まっさおなインド洋に面したダルエスアラームの浜辺で仲よくなった鋼鉄のように黒光りした精悍な青年が「お守りに」と，さめの奥歯を私の手のひらにのせてくれた．
「ありがとう．ついでにもう一つおねだりしていい？私，アフリカの食べ物を知りたいの．ホテルでは英国かインドの料理だけなんですもの．あなたのお家で昔から普通に食べてるものは，どんなもの？」
（桐島洋子 1976『聡明な女は料理がうまい-女ひとりの優雅な食卓からーパーティのひらき方まで-』主婦と生活社，サンプル ID:OB1X_00068, TTR:0.96）

用例（16）は「私」がタンザニアを旅行した際の経験が書かれた一節である．地の文と台詞文が使い分けられており，「私」を基準として文章が作られている点で用例（13）や（15）に近い文体でもある．

（17） なにしろ頼れるのは日本語のガイドブックだけ．「きょろきょろしてはいけない，目的地に一目散に向かうこと」という教えにきちんと従ったのだ．
だから五十七丁目のにぎやかな景色は，目のはしっこでデュフィの絵のごとく流れるように過ぎてゆき，そこがシャネルやティファニーなどの店が連なる華やかな通りと知ったのはずっと後のこと．
（牧かほり 1995『NY アートストリート 18 番地』アリアドネ企画，サンプル ID:LBj3_00109, TTR:0.953）

用例（17）は書き手がニューヨークの「アートスクール」に通った際の経験談の一節であり，全体として反復語があまり用いられないファイルである．

４．１．２　反復語が極端に多い文章の実例

TTR の値の極端に小さいと思われるファイルは，0.45 未満の 3 ファイルであると言えるため，それらに含まれる文章の一部も以下で取り上げる．これらは反復語の使用が最も多かった文章の抜粋である．下線部が反復語として用いられる語である．

（18） 「カメラの付いた小さな携帯電話」という表現には何ら問題はありません．しかし，「小さなカメラの付いた携帯電話」という表現には，ひとつ問題があります．「小さな」が「カメラ」の修飾語に見えてしまうのです．このままでは，「小さなカメラ」が「携帯電話」に付いていると受け取られても仕方ありません．
（後藤禎典 2005『「後藤式」文章の技術-わかりやすい文が書ける明快ルール-』PHP 研究所，サンプル ID:PB53_00574, TTR:0.39）

用例（18）では，文構造によって「小さな」と「カメラ」と「携帯電話」の文内での関係性が変わることを指摘している一節である．それらの関係性の変化を具体的に示すために「カメラ」や「携帯」「電話」[7]が反復語として多く使われる[8]．

（19）エッチを１年ぐらいして，何か違うからダメだと気づくのは早い遅い以前の問題です．
そんなことはごはんを食べている間に気づくべきです．
エッチで彼がダメだと思っても，君がエッチ[9]な女性ということにはなりません．
エッチの時には，彼の品性や感性，ライフスタイルがすべて出ます．
エッチのことをあまり口に出せない女性がいますが，それは「言っていい」という男性と出会ったことがないだけです．
（中谷彰宏 2001『尊敬できる男と，しよう-女を上げる 43 の方法-』大和書房，サンプル ID:PB11_00048, TTR:0.44）

用例（19）は，特に「エッチ」という語が何度も用いられており，その反復を中心として文章がまとまっていることが見て取れる．

（20） 第１項の規定は，確定申告書に同項の規定による控除を受けるべき金額及びその計算に関する明細の記載があり，かつ，控除対象外国法人税の額を課されたことを証する書類その他財務省令で定める書類の添付がある場合に限り，適用する．この場合において，同項の規定による控除をされるべき金額は，当該金額として記載された金額を限度とする．
（川田剛 2004『国際課税の基礎知識』税務経理協会，サンプル ID:PB43_00651, TTR:0.449）

用例（20）は法律を始めとする規則の明示化を図る文章に位置づけられ，誤読を避けるために同一の情報は同一の情報として明示的に記すという文体を取っていると考えられる．

４．２　反復語の多寡が見せる文体的特徴

文章の構成要素自体は多様であるため，それらの複合によって作られる文体も，必ずしも反復語の多寡のみで決定されるわけではない．しかし，表 5 によれば，地の文において「僕」や「私」といった一人称表現がファイル中で一度以上用いられる文章は，TTR が 0.9 以上で反復語が相対的に非常に少ない 101 ファイルの中の 63.37%に見られ，

---

7 「携帯電話」は複合名詞として一語とみなすこともできるが，本稿では「携帯」と「電話」の二つに分離して取り扱う．これは，例えば「携帯電話」と「携帯用の電話」といった表現が同じ文章内で使われた場合，「携帯電話」を一語として捉えると両表現が別語となり，コンピュータによる集計上は「携帯」や「電話」が反復語に含まれなくなってしまうためである．複合語の一部であっても同じ語が出ているならば，両者に関係を認める方が自然であると考えられるため，本稿では「短単位」を採用した集計を行った．

8 この一節が含まれるファイルは日本語表現に関する問題を日本語で説明している点でメタ言語的な文章であり，言語外的な事態を言語を用いて描く文章とは異なった性質を持つとも考えられる．しかし，ある言語表現について説明することと，ある物事について説明することは，結果的にそれが説明手段として用いる言語と同一であるかどうかの差以外には違いがない．したがって，このようなメタ言語的な形式を持つ文章も，本稿では分析対象として含める．

9 注 7 と同様に，「エッチな」は一語の形容動詞と捉えることもできるが，ここではそれぞれを分離して取り扱った．

多くを占める一方，TTR が 0.55 未満で反復語が相対的に非常に多い 71 ファイルでは 19.72%と，TTR が 0.9 以上のときに比べ割合が小さくなる．また，法律文書や法律に関する手引き書は，TTR が 0.9 以上の場合には一例も見られないが，TTR が 0.55 未満の 71 ファイルでは 38.03%と，多くを占める．

表 5：一人称の文章と法律関連の文章の数

TTRの値	地の文:一人称	法律文書・手引書
0.55未満	14/71(19.72%)	27/71(38.03%)
0.90以上	64/101(63.37%)	0/101(0%)

このことから，反復語をほとんど用いない場合，書き手・語り手が自らの考えや経験を具体的に語っていることを明言する一人称表現の使用によって，文章における内容の連続性やまとまりを保つことが多くなることが窺える．反復語を非常に多く用いる場合は必ずしも書き手の一人称が必要ではなくなり，法律文書のような客観的な規定を表現する文章がそこに位置づけられるようになると考えられる[10]．

地の文で一人称表現を用い，反復語が少ない文章は用例（13）や（15）でも観察できるが，ここではそれらより反復語がやや多く，表 4 において TTR の値が 0.90 以上 0.95 未満となったファイルの一部を抜粋する．

（21） 私は特別，神経質にならざるをえない体質であるけれど，これから生まれてくる子たちの親となる人には，私の悩みとあなたの子のそれは同じかもしれない．
古代中国には，地に生えている植物を枝でたたいただけで薬草を見分けたという伝説の神農さんという神人がいた．
私が痛いめを繰り返して自分なりに分類するのとはえらい違いだが，そんな人がいたら，どんなに世間の人の目が開かれることか，と思う．
（柏木亜希 2004『気まぐれ雑記帳』日本随筆家協会，サンプル ID:PB40_00037, TTR:0.92）

用例（21）は食品や薬品に対して過敏な反応を示す体質である「私」によって語られている文章だが，読者の「子」の話と古代中国の「神農」の話という別個の話が，「私」の体質に関わる話であるということが自明であることによって内容的まとまりを形成している．

また，法律文書で反復語が非常に多いものとしては用例（20）が挙げられるが，ここでは法律に関する解説を行った文章で，表 4 において TTR の値が 0.50 以上 0.55 未満の中に位置づけられるファイルの一部を取り上げる．

---

10 一人称が用いられる文章と，法律文書のような文章との関係をどう位置づけるかには問題が残る．一人称（=書き手）に関わる文章に対置されるべきものは二人称（=読み手）・三人称（=書き手・読み手以外）に関わる文章であると言え，二人称の事態・概念について書くことは考えにくいことから，実際には三人称の文章が一人称に対置される存在であると考えられる．法律文書は三人称の文章であるが，小説にも三人称で書かれたものがあり，それらは必ずしも反復語が多いわけではない．この点について本稿では，法律文書やその手引書が実用的な目的で書かれた文章であり，書き手の特殊な経験を問題としていないものであるのに対し，一人称を用いた文章が書き手の特殊な経験・論理を重視し，実用性を問題としないものである点で，対比的に位置づけられると考えている．

（22） なお，株式会社であれ，有限会社であれ，会社には社長がいるのがふつうです．社長は会社のトップであり，社長は当然に会社の代表機関であると思っている人も多いでしょう．しかし社長は，会社の内部規定（定款）で会社が自主的に定めた職制上の定めにすぎず，法律上の代表機関ではありません．
（小林英明 2001『契約書のすべて-契約の基本から有利な契約書の作成方法まで消費者契約法に完全対応！！決定版-』PHP 研究所，サンプル ID:LBp3_00026, TTR:0.52）

用例（22）でも（20）と同様に，同一の名詞は反復語としてできるだけ明示的に用いられており，それによって相互の文の関係を読み取れるとともに内容の正確な読み取りが可能となっている．

以上で観察してきたように，反復語が多く用いられた場合，それが文章内の意味の連続性とまとまりの形成に強く作用し，法律文書やそれに関する手引書のような文体を形作るようになると言える．一方で，反復語を用いない場合であれば，書き手の一人称表現を明示的に用いることで，文章の意味の連続性やまとまりを作るための反復語の代わりとすることが可能になり，その典型として，「私」の経験談や物語であることを明示するような文体が存在すると考えられる．

５．まとめ

本稿では，Type-Token Ratio（TTR）という計算方法を，新出語と反復語の比率を表すものとして利用し，文章における名詞の反復の多寡の幅を定量的に明らかにした．

分析の結果，TTR の値と，そこから考えられる反復語の比率は以下のような範囲を取ることが分かった．

（23） TTR の最小値　　：0.390
TTR の第 1 四分位：0.722
TTR の中央値　　：0.774
TTR の平均値　　：0.765
TTR の第 3 四分位：0.819
TTR の最大値　　：0.960
反復語の割合の最大範囲　：　4 % ~ 61 %
反復語の割合の一般的範囲：　15 % ~ 30 %

この結果においては TTR の値の上限が比較的高く，反復語をほとんど用いないということがありえる反面，反復語は用いることが多くなっても文章中で 60%程度までしか占めることができていない．したがって，新出の名詞を最低でも 40%程度まで提示すること，実際には表 4 で示したように TTR の値が 0.5 を切った文章が全体の 0.19%に過ぎないことから，少なくとも半分程度を新出語によって表現するということが，文章構成上の一つの特徴をなしていると考えられる．

さらに，反復語の多寡が具体的にどのような文体的特徴と関係するかを分析するために，TTR が 0.55 未満で反復語が相対的に非常に多い 71 ファイルと，TTR が 0.9 以上で反復語が相対的に非常に少ない 101 ファイルを観察した結果，それぞれに次のような傾向が見られることが分かった．

（24）　・反復語が非常に多い文章
　　一様な文体ではないが，法律文書やその手引書が目立つ
　　・反復語が非常に少ない文章
　　法律文書やその手引き書が見られず，一人称を用いて書かれた文体が多い

名詞による反復語の使用は内容の意味的連続性とまとまりを明示的に形成できるために，法律文書やその手引書といった明確な情報伝達がなされなければならない文体で使用されると考えられる．その反面，反復語が少ない場合は別の言語手段によって意味的連続性とまとまりを保証しなければならず，その手段として書き手自身の語りであることを明示的に示す一人称が用いられることが多くなると言える．

本稿では，反復語が文章中において占める量的様相と，その多寡と文体的特徴との関係の一端を明らかにした．ただし，そもそも本稿で資料の絞り込みに用いた条件がそのまま「文章」の成立条件となっているわけではなく，選定上の条件の網羅性に欠けている点に問題が残る．この点については，今後詳細に検討したい．また，文体上の分析も，一人称の文体と法律文書の文体は必ずしも相補的ではないため，対比的に取り上げるためにより詳細な文体的特徴の把握が重要であると考えられる．内容上のまとまりについても，本稿では操作的に名詞 50 語としたが，具体的な内容展開に即した計測を行い，より正確な反復語の量的様相を明らかにしたい．

文献

石川慎一郎（2004）日韓の大学入学試験英語問題に見る構成語彙の特徴–英文テキスト・コーパスの解析に基づく考察–，『アジアの英語と英語教育』7, pp.1-15.

小野望・田中省作・持尾弘司（2007）母語学習者コーパスの基礎調査，『筑紫女学園大学・短期大学部人間文化研究所年報』18, pp.27-36.

影浦峡（2000）『計量情報学–図書館/言語研究への応用–』丸善.

金明哲（2009）『テキストデータの統計科学入門』岩波書店.

国立国語研究所コーパス開発センター（2011）『『現代日本語書き言葉均衡コーパス』利用の手引』.（『現代日本語書き言葉均衡コーパス』DVD 版に収録）.

鈴木崇史・影浦峡（2011）名詞の分布特徴量を用いた政治テキスト分析，『行動計量学』38.1, pp.83-92.

砂川有里子（2000）談話主題の階層性と表現形式,『文藝言語研究（言語編）38』.（砂川有里子（2005）『文法と談話の接点-日本語の談話における主題展開機能の研究-』くろしお出版に再録）.

高崎みどり・新屋映子・立川和美（2007）『日本語随筆テクストの諸相』ひつじ書房.

田島ますみ・深田淳・佐藤尚子・玉岡賀津雄（2009）語彙指標数値と文章主観評価の関係–日本人大学生による 2 種類の書き言葉コーパスを使った実証研究–，『中央学院大学人文・自然論叢』29, pp.57-77.

馬場俊臣（1986）「主要語句の連鎖」と「反覆語句」との交渉,永野賢（編）『文章論と国語教育』朝倉書店.（馬場俊臣（2006）『日本語の文連接表現-指示・接続・反復-』おうふうに再録）.

水元篤（2008）自由英作文における語彙の統計指標と評定社の総合的評価の関係，『統計数

理研究所共同リポート』215, pp.15-28.
安江佐和子（1981）行文を追った異なり語数の動き，『東京女子大学日本文学』56,pp.32-45.
山崎誠（1983）文章の話題の展開を計る尺度–用語類似度 D の 1 利用法，『計量国語学』13.8, pp.346-360.
山崎誠（2010）語の平均使用度数に現れるテキストの特徴,『特定領域研究「日本語コーパス」公開ワークショップ（研究成果報告会）予稿集』.
Halliday, M. A. K. & Hasan, R（1976）*Cohesion in English*. London: Longman.（M. A. K. ハリデイ・ルカイヤ ハサン（著）安藤貞雄・多田保行・永田龍男・中川憲・高口圭轉（訳）『テクストはどのように構成されるか』ひつじ書房）.
Stubbs, Michael（2002）*Words and Phrases: Corpus Studies of Lexical Semantics*. London: Blackwell Publishing Ltd.（マイケル・スタッブズ（著）南出康世・石川慎一郎（監訳）（2006）『コーパス語彙意味論–語から句へ–』研究社）.
Youmans, Gilbert（1991）A New Tool for Discourse Analysis: The Vocabulary-Management Profile, *Language*, 67(4), pp.763-789.

調査資料

国立国語研究所（2009）『『現代日本語書き言葉均衡コーパス』モニター公開データ（2009年度版）』.

使用したソフトウェア

「MeCab:Yet Another Part-of-Speech and Morphological Analyzer」http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html　2012/5/30 時点.
「形態素解析辞書 UniDic」http://www.tokuteicorpus.jp/dist/　2012/5/30 時点.
「The Perl Programming Language - www.perl.org」http://www.perl.org/　2012/5/30 時点.
「The R Project for Statistical Computing」http://www.r-project.org/　2012/5/30 時点.

（2011 年 11 月 8 日受付，2012 年 3 月 18 日再受付）

*** Keywords and Abstracts***

Report

The Quantitative Condition of Repeated Nouns in Writing: A Type-Token Ratio Analysis

KUJIRAI Ayaki (Graduate School of Arts and Letters, Tohoku University)

Keywords: repeated words, new words, Type-Token Ratio, BCCWJ

Abstract:

Repeated nouns act as a means to guarantee continuity and unity of meaning in a text. The purpose of this report is to clarify the quantitative condition of such repeated nouns in writing. A Type Token Ratio (TTR) analysis is used in order to accomplish this goal.

The data used in this analysis is the “Publication Sub-corpus” and the “Library Sub-corpus” from the “Balanced Corpus of Contemporary Written Japanese” (Monitor Exhibition Data 2009).

As a result of the analysis, it is found that the percentage of repeated nouns in the data occupies between four and sixty one percent of a text, and the average of repeated nouns is approximately twenty five percent. These results show that most of the repeated nouns may not be used in a text. At the same time, even if a lot of repeated nouns are used in a text, nearly half of all nouns are new.

In addition, the majority of the texts with a low quantity of repeated nouns are written in first person. Likewise, texts with a large quantity of repeated nouns, mainly consist of legal documents and manuals concerning law.