均衡性と代表性に配慮した『太陽コーパス』の分析法試論

森　秀明（東北大学大学院文学研究科）†

Methodological Consideration on Corpus-Balance and Representativeness of "Taiyo Corpus"

Hideaki Mori (Graduate School of Arts and Letters，Tohoku University)

要旨

『太陽コーパス』は、明治後期～大正期の総合雑誌『太陽』から５年分を抽出した全文コーパスである。近代日本語の成立期に当たるデータが集積されているため、語や文法の経年変化分析に使用されることが多い。ある形式を出版年ごとに比較する場合、それぞれのデータがある程度均質でないと正確な分析はできない。しかし、『太陽コーパス』は出版年ごとのデータに記事数・文字数・ジャンル等の偏りがあるほか、著作権問題により一部の記事が非公開となっているなど、非常に不均衡な状態にある。そこで本稿では「美術・芸術」という語の使用割合を例に、5 種類の経年変化分析の方法を検討した。その結果、これまで多用されてきた粗頻度による分析法では有効に分析できない内容語があり、テキスト平均文字数当たりの調整頻度（PTA）を使用し、ロジスティック回帰分析によってジャンルを統制する分析法（PTA・ジャンル統制法）の方が有効な分析法だと考えられた。

１．はじめに

コーパスの代表性と均衡性は同時に語られることが多いが、これらは基本的に別々の概念である。ごく単純化して定義すれば、代表性とは「あるコーパスが、推定対象の言語を正確に反映していること」であり、均衡性とは「（推定対象の言語を母集団として、そこから均衡にサンプリングした結果）コーパスが均衡な特性をもつこと」であると言えよう。

しかし『現代日本語書き言葉均衡コーパス』（以下 BCCWJ と呼ぶ）のような、均衡に設計されたコーパスに比べ、『太陽コーパス』の代表性に疑問符が付くのは否めない。「コーパスデータの大半が特定の雑誌だけから取られていた場合、得られた結果が言語全般の特徴なのか、当該雑誌の特徴なのかは不明です。」(石川, 2012, p. 22.)という指摘や、新聞記事のような一媒体では日本語の代表としてみなしにくいことを論じた後藤(1995, 1996)の主張は、重く受け止める必要がある。それでは『太陽コーパス』に全く代表性がないのかと言えば、それも程度問題ではあるだろう。雑誌『太陽』が近代日本語を代表する資料であることは、紛れもない事実であると思われる (国立国語研究所(編), 2005 ; 田中 , 2012)。

これに比べより明確な問題点は『太陽コーパス』の不均衡性にある。『太陽コーパス』はデータに様々な偏りを持っているため、そこに何らかの代表性があったとしても、その姿は大きくかき乱されて、単純な粗頻度分析では有効な値を示さない可能性がある。そこで本稿では、どのような分析法を使用すれば、少しでも有効な分析ができるのかについて、5 種類の分析法を比較しながら検討してみる。

２．『太陽コーパス』におけるデータのばらつき

２.１　出版年別テキスト数と文字数

『太陽コーパス』のデータがばらつく大きな原因の一つに著作権問題がある。『太陽コ

† hideaki@moriharuo.com

ーパス』は 5 年分の全記事をコーパス化することを目指して作成されたが、年代が新しくなるにしたがって、著作権上の問題により削除されたデータが増加し、それが 1925 年では文字数にして約 3 割に上っている。図 1 は削除前のデータ（評価版・内部資料）と削除後のデータ（『太陽コーパス』公刊版）の文字数の比較を示したグラフ、図 2 は記事数の比較を示したグラフである（国立国語研究所(編)(2005) , P. 20.の表 5, 6 を元に作成）。

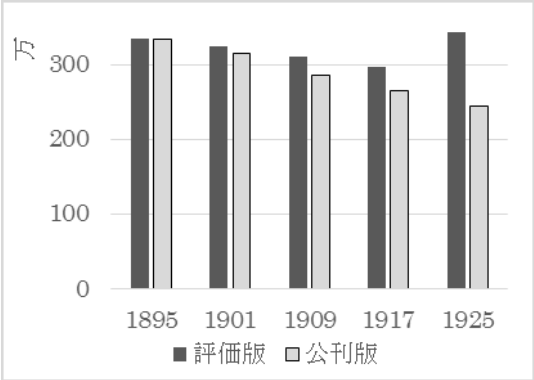

図 1 評価版と公刊版の文字数比較

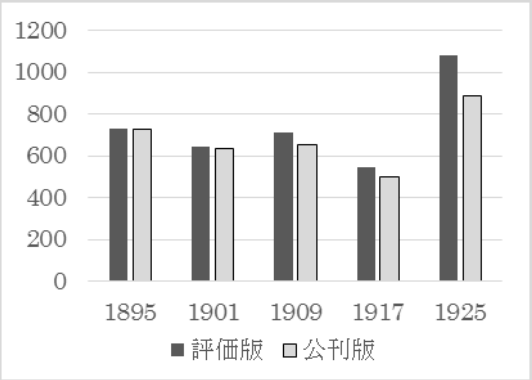

図 2 評価版と公刊版の記事数比較

図 1 で文字数の減少を追うと、1901 年から徐々に減り始め、1925 年では大きく減少しているのが分かる。著作権上の理由により一部のデータが公開できないという問題は、大規模な公開コーパスでは避けられないことであるが、一般利用者にとっては、難しい課題を抱える結果となる。例えば国立国語研究所(編)(2005)は、コーパスを制作した内部者の研究を所収しているため、これらの論文には削除前のデータが使用されている。しかし、公刊版では削除後のデータしか使用できないため、一般利用者がこれらの追試をしようとしても、初めからデータ数が食い違ってしまうのだ。このため、一般利用者は、削除後のデータを使用して、削除前のデータを使用した分析と同レベルの結果が出せるような、何らかの分析法を工夫する必要に迫られるのである。

しかし、『太陽コーパス』におけるデータのばらつきは、この著作権問題によるものだけではない。そもそも評価版の文字数自体が出版年ごとに異なっているのである。雑誌は商業的に作られた一媒体に過ぎないため、様々な要因によって各号ごとの記事数や文字数が異なる。そのような非常にばらつきのある言語資料を使用して近代日本語の平均的な姿を推定するためには、欠損の大きい 1925 年だけでなく、そもそもすべての出版年を均衡化させるような何らかの工夫が必要なのである。

図 2 は、記事数の比較である。これも年代が新しくなるにつれ徐々に公刊版の記事数が減少している。1925 年ではおよそ 2 割の記事が減少していることから、非公開となった記事には比較的長文のものが多かったことが推測される。また、出版年ごとの記事数のばらつきも激しい。特に評価版の段階から 1909 年は他の年より記事数が少なく、逆に 1925 年は非常に多い。この両年を比べると 1925 年は 1909 年のほぼ倍になっている。

２.２　１記事当たりの文字数

表 1 は、『太陽コーパス』全体の統計量である（扉・奥付等を除く）。記事数 3241 件のうち、最も短い記事の文字数は 27 字、最も長いものは 51705 字である。平均は 4442 字だが、中央値が約 3 千字であるため、『太陽コーパス』は多くの短い記事と少数の長い記事

によって構成されていることが推察される。文字数の多い記事上位 5%の文字量は全体の約 20%に及ぶ。これらごく少数の記事が分析対象の統計量を大きく左右している可能性がある。

表 1 記事の文字数統計量

記事数	3241
平均値	4442.17
中央値	2985.00
最頻値	643
標準偏差	4589.748
最小値	27
最大値	51705

表 2 出版年ごとの統計量

出版年	記事数	平均	標準偏差	最小	最大
1895	699	4752.30	4137.125	43	26643
1901	592	5304.35	4647.207	72	38152
1909	616	4624.80	4261.947	140	29343
1917	467	5669.53	6100.026	65	51705
1925	867	2812.55	3643.778	27	22482
合計	3241	4442.17	4589.748	27	51705

表 2 は出版年ごとの統計量だが、この平均の値を見ると予想通り 1917 年の平均は 5669 字、1925 年の平均は 2812 字であり、両年の記事の長さがかなり異なっている。これをさらに詳しく分析するため、記事の文字数を縦軸にして描いた箱ひげ図で検討する。図 3 は出版年ごとの全体図、図 4 は中心部を拡大した箱ひげ図、図 5 は参照のために掲げた BCCWJ「図書館書籍」サブコーパスの箱ひげ図である。

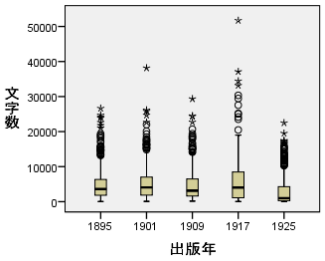

図 3 『太陽コーパス』全体

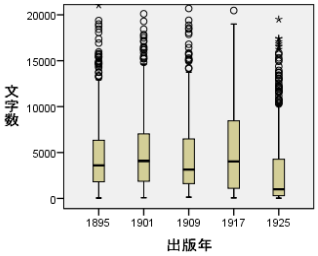

図 4 『太陽コーパス』中心部

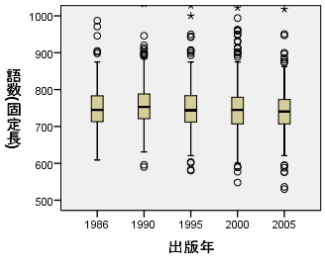

図 5 BCCWJ「図書館書籍」

図 3 を見ると各年とも多数の短い記事と、少数の長い記事によって構成されていることが確認される。特に 1917 年は長い記事が多い。図 4 で各年の箱の状態を見ると、初めの 3 年はほぼ均質だが、1917 年は箱の長さも長く、95 パーセンタイルの位置も高い。一方 1925 年は箱の長さも短く、中央値が極端に低くなっている。この年だけ極端に中央値が異なるのは、そもそもの記事数が多かったためと、著作権上の問題で長めの記事が非公開になったためと考えられる。一方、図 5 は、参考のために BCCWJ の「図書館書籍」サブコーパス（固定長）を約 5 年ごとに取り出し、その語数を描いたものである。これを見るとすべての出版年のデータが非常に均質である。BCCWJ と比較すると『太陽コーパス』がいかにばらつきを持ったコーパスであるかが良く分かる。

それではこれらのばらつきによって、具体的にどのような問題が生じるのであろうか。最も問題と思われるのが、出版年ごとに語の出現のしやすさが異なってしまう可能性である。一般にテキストの文字数が少なく制限されるほど、名詞比率は大きくなる。また 1 テキスト当たりの名詞比率が大きくなるほど、形容詞や副詞などが直線的に減少する（これを「樺島の法則」という。 樺島, 2009）。単純化していうなら、1917 年では長い記事の中で同じ語が何回も使われ、1925 年では短い記事の中で異なった語が次々と使われるため、

あるトピックに現れやすい語の頻度は、1917 年では多く、1925 年では少ない結果になりやすい。ただし Stubbs (2006) によれば助詞、助動詞などの機能語は、1 テキスト当たりの文字数にはそれほど影響されないため、これは主に内容語に生じる問題であると考えられる。

『太陽コーパス』のデータのばらつきが、単なる総文字数だけであれば、各出版年ごとの PMW などを求め、それで各年の比較を行えば問題ない。しかし、出版年ごとに 1 記事当たりの文字数に大きなばらつきがあるということは、PMW などでは平準化できない問題をはらんでいると考えられる。

２.３　ジャンル割合

『太陽』は雑誌であるため、各号のジャンルは時事的な出来事に大きく左右される。図 6 を見ると 1917 年では他の年より「社会科学」(点々)の割合が高い。この理由は 1914 年から始まった第一次世界大戦や、1917 年に勃発したロシア革命に関する記事が多く所収されたためだ。戦争のように社会的な影響力が強い出来事の場合、ジャンルの変化を言語の変化と見なす立場もあるかもしれない。しかし同年の「歴史」(横縞)などは、単に紙面上の制約によって減らされたと考える方が妥当だと思われる。

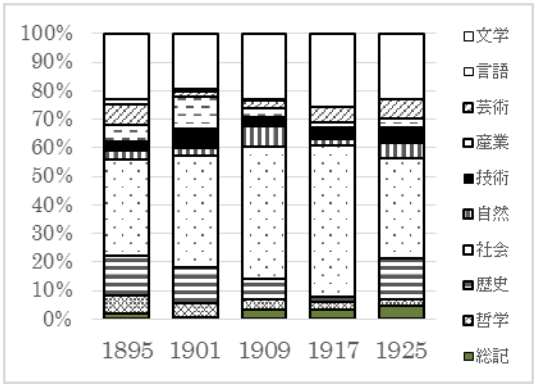

図 6 『太陽コーパス』のジャンルの割合

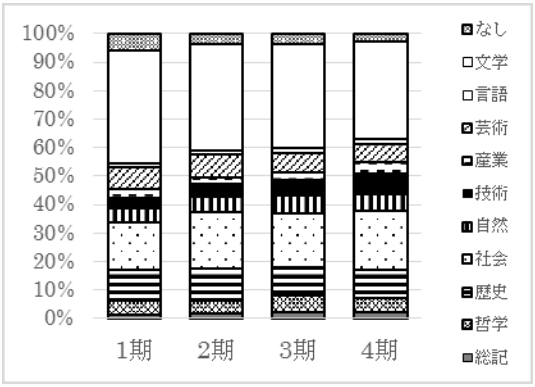

図 7 BCCWJ のジャンル割合

図 7 は、比較のため BCCWJ「図書館書籍」サブコーパスのジャンルを 5 年ごと 4 期に分けて示したものである。かなり均質な分布になっているが、古い時代ほど「技術」(黒)の割合が低く、「文学」(白)の割合が高いなどの変化も見える。これは技術的な内容の場合、技術の更新に伴って古い書籍が廃棄されやすく、文学は時代を経ても古い書籍がそのまま読み継がれるといった図書館特有の理由によるものだろう。

これに比べ、『太陽コーパス』では時代ごとに様々なジャンルのばらつきがあり、そこに一定の性質は見出せない。このようなばらつきは、その時々の編集方針によって生じたものと考えられる。ジャンルの偏りは、特定ジャンルに使用されやすい語の出現頻度に影響を与える。歴史ジャンルで使用されやすい語であれば、1917 年には歴史に関する記事そのものが極端に少ないため、その語の頻度も当然少なくなることが予想される。

３．均衡性に配慮した分析法の検討

第 2.2 節で述べたように、文字数やジャンルの影響を受けやすいのは内容語だと思われる。そこでここでは「美術・芸術」という内容語を取り上げ、データの均衡性に配慮した分析を様々に検討してみる。「美術・芸術」を取り上げるのはその語史がある程度解明されて

いるためである。なお、表計算ソフトは Excel、統計ソフトは SPSS を使用した。

３.１．「芸術・美術」の経年変化の予想

『日本国語大辞典第二版』では「芸術」の語史が以下のように記されている。

> 近世まではもっぱら「学芸・技術」の意で用いられたが、明治期に西洋文化の摂取が盛んになるに及んで、英語の art その他、美の表現・創造を共通の概念とするヨーロッパ各国語の訳語としての（中略）意味が出現した。ただし、明治初期にはむしろ同じ訳語に「美術」を用いることがより一般的であり、（中略）芸術が新しい意義で定着するのは、ほぼ明治三〇年（一八九七）前後である。（第四巻，p.1247.）

これからすると、「美術」の出現頻度は明治初期には高く、時代が新しくなるにつれ低くなる（「芸術」の頻度は逆の結果となる）ことが予想される。BCCWJ で「芸術割合」（芸術の頻度と美術の頻度の合計における芸術の頻度の割合）を調べると、「図書館書籍」サブコーパスの固定長で 46.4%、「出版書籍」サブコーパスの固定長で 53.6%、「特定目的」サブコーパスも含めた全体で 48.7%となっている。これらが最終的な使用割合であると仮定すれば、「芸術割合」はおよそ 50%前後で頭打ちになることが予想される。『太陽コーパス』を使用した分析法でこのような予想に合致する分析結果が得られるなら、その分析法はある程度有効な分析法だと考えられる。

ただし、両者の拮抗する時期が本当に 1897 年前後なのかは疑問である。この記述は同年に書かれた正岡子規の文章[1]を根拠にしているが、このような指摘は言語が変化し始めた初期に、言語変化を言葉の乱れとみなして表明されることが多い。『太陽コーパス』で正確な分析ができるなら、両者が拮抗する時期の推定もある程度正確に行えるはずである。

なお、検索ソフトは『ひまわり』を使用し、一般的な記事とは見なしにくい扉、奥付等のデータは分析から除く。また、旧漢字は新漢字に直して表記する。

３.２　粗頻度分析と出版年ごとの PMW 分析

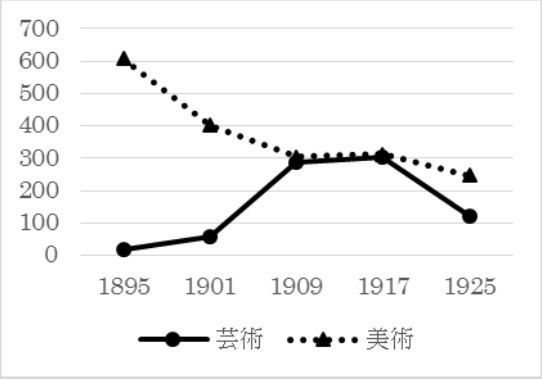

図 8 「芸術・美術」の粗頻度分析

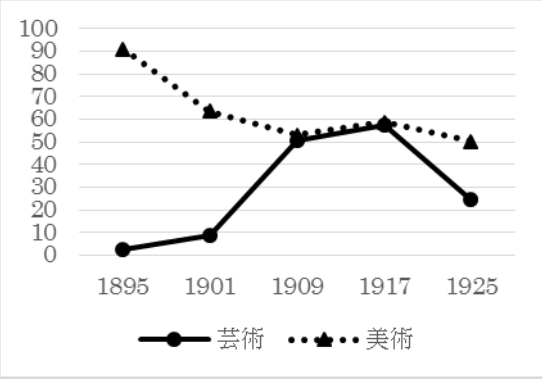

図 9 「芸術・美術」の PMW 分析

[1] 「博覧会などにて美術工芸品といふ語を用うるはやがて世人をして絵も彫刻も織物も陶器も同一美術品なりと誤想せしめ、従って美術品は工芸品なりと誤想せしむるの一大原因にやあらん」[閒人閒語]

図 8 は粗頻度（縦軸は語数）、図 9 は PMW（縦軸は PMW）のグラフである。2 つのグラフは、ほぼ似た形となっている。PMW 分析の場合、年代が新しくなるにつれて語の頻度が高くなるのは、著作権問題で削除された文字数分が補正されているからである。予想からすると「芸術」と「美術」の頻度は徐々に接近し、真ん中付近で一本の線になるはずであるから、1917 年まではある程度正確な分析になっているようにも見える。

問題は 1925 年である。この年は、もともと記事数が特別多かった上に、著作権問題のため、長めの記事が文字数で 3 割ほど非公開になっているなど、非常に偏りが大きい年であった。この年では懸念した通り両者の頻度が少なくなっている。『太陽コーパス』の場合、粗頻度分析と PMW 分析では有効に分析できない内容語があると言えるだろう。

３.３　割合による分析

前節で観察したように、粗頻度を使用した場合、有効に分析できない内容語がある。しかし、出版年ごとの割合であれば、同じばらつきの影響を受けたもの同士を割り算するため、それらの影響が相殺されて有効な分析となるようにも思われる。そこでここでは出版年別の「芸術割合」で分析してみる。

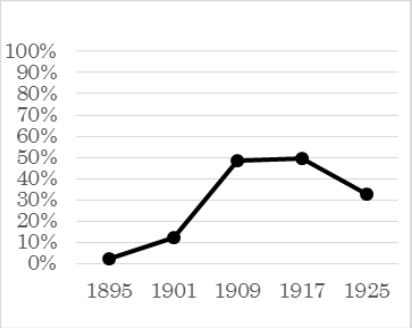

図 10 粗頻度を使用した芸術割合

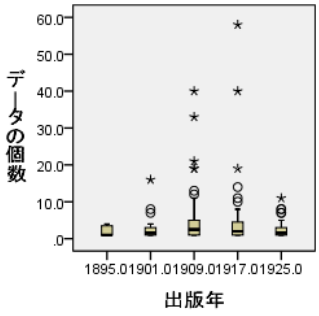

図 11 芸術の箱ひげ図

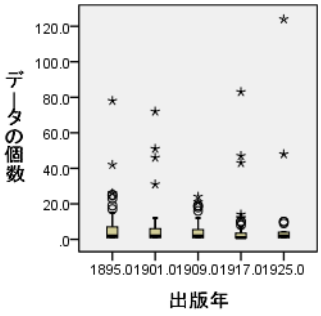

図 12 美術の箱ひげ図

図 10 は粗頻度を使用して算出した「芸術割合」である。1925 年では若干割合が高くなるが、1917 年よりまだ 20%程度も低い。割合分析であっても粗頻度を使用する限り、1925 年のばらつきが解消できない内容語があることが確認される。

この原因を探るため、1 記事当たり頻度の箱ひげ図で観察すると、図 11「芸術」の 1925 年ではすべての記事が十数頻度以下であるのに対し、図 12「美術」の 1925 年では 1 つだけ頻度 120 を超える記事がある。「美術」の 1925 年の総頻度は 73 記事合計で頻度 246 であるから、この記事だけで全頻度の約半数が出現していることになる。統計では、あまりに異質なデータは外れ値として分析に含めないことがある。この記事を外れ値として除外すれば 1925 年の「芸術割合」は 49.5%となる。これは予想に合致した値であるため、この記事は除外した方がよさそうにも思える。しかし、その判断は妥当だろうか。

『太陽コーパス』の記事は、最少文字数が 27 字、最大文字数が 51705 字というばらつきを持つ。そのためこのような現象は分析のたびに観察される可能性がある。そのたびに都合の悪いデータを除いていたのでは、データの改ざんにつながりかねない。問題の本質はこの記事だけにあるのではなく、『太陽コーパス』の多くのデータがばらついていることにあると考えるべきであろう。

３.４　テキスト平均文字数当たりの頻度（PTA）による割合分析

記事ごとの文字数のばらつきが分析を阻害する要因であるなら、記事の文字数がすべて平均文字数に近い 4500 字などであるとみなして、その分、頻度を調整した値にすればよい。例えば 450 字の記事に 1 回出現するなら、その頻度は 10（1÷450×4500）、45000 字の記事に 1 回出現するなら、その頻度は 0.1（1÷45000×4500）とカウントするのである。このようにテキスト平均文字数当たりの頻度に換算した調整頻度を PTA（per number of the text average letters）と命名する。ただし、「芸術」や「美術」は二字熟語なので、計算は（1÷検索語が出現したテキストの文字数×4500÷2）で行う。

しかし、実際にこの方法で分析しようとすると、ひとつ困難な課題に直面する。『太陽コーパス』では、記事ごとの文字数がタグ付けされていないため、PTA を求めたくとも、簡単には求められないのである。利用者各自が約 3400 記事について、文字数をタグ付けすることも考えられるが、誰しもが実施しやすい方法とは言えないだろう。

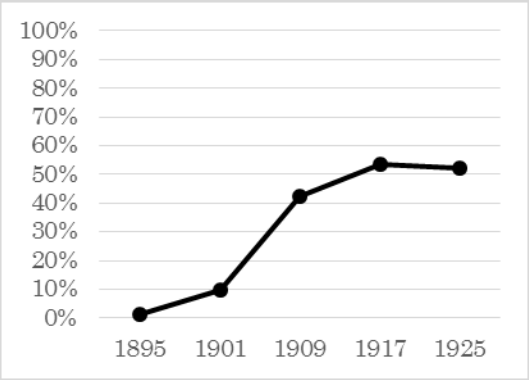

図 13 芸術 PTA 割合

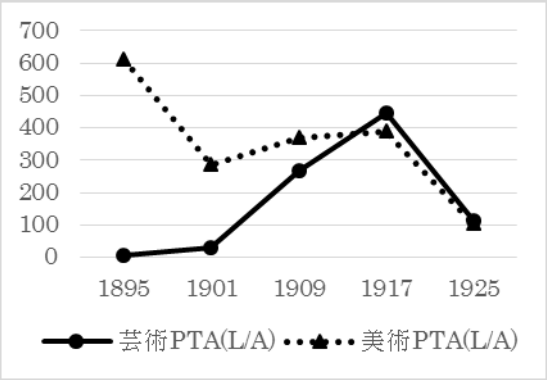

図 14 芸術と美術の PTA(L/A)の経年変化

そこで今回は検索語が出現した記事ごとの頻度リスト[2]と『太陽コーパス』全記事の文字数リスト[3]を作成し、同名の記事をできるだけ区別できるようにした上で、リストを利用した置換[4]を行って各記事の文字数を特定した。これによって「芸術 PTA 割合」を求めたのが図 13 である。図 13 の「芸術 PTA 割合」は、ほぼ予想と合致する結果である。このためテキスト平均文字数当たりの調整頻度（PTA）による割合分析は、ある程度有効であると思われる。図 14 は、各々の PTA をさらに出版年ごとの平均記事数で平準化させた調整頻度（こ

[2] 記事ごとの頻度リストの作成法：①『太陽コーパス』で検索した語の用例を Excel で開く。② ①のシートをコピーし、データ→重複の削除で、年・号・題名にチェックを入れ、重複を削除する。③元のシートで年・号・題名のデータを選択してピボットテーブルを挿入し、年・号・題名を列に、題名を値に入れて集計する。④ピボットテーブルの中身をコピー&ペーストし、ピボットテーブルは削除した上でデータに通し番号を付ける。⑤行ラベルでソートし、題名以外のデータを削除する。⑥通し番号でソートし、元の順番に直す。⑦題名とデータの個数をコピーし、②のシートの題名の横に挿入する。⑧題名が一致していることを確認したら、コピーした題名の列を削除する（同名記事で判別不能なものは平均値を使用）。

[3] 文字数リストの作成法：太陽コーパス付属のツール・プリズムを立ち上げ、①「入力 XML ファイル」ウインドウの下にある「別フォルダを指定」のボタンで、太陽コーパスの ZASSI フォルダ（Himawari_ﾊﾞｰｼﾞｮﾝ番号→Corpora→Zassi →Taiyo）を開き、corpus.xml を指定する。②「適用するスタイル」の中から csv.kiji.xsl を選び、「変換（ファイルへ出力）」を押すと、デスクトップにテキストファイルが出力される。これを Excel などで読み込む。

[4] リストを利用した置換の方法は http://stabucky.com/wp/archives/3259 で紹介されている（2014.07.22 閲覧）。

れを PTA(L/A)と記述する）でグラフ化したものである。これを見ると、どんなに文字数や記事数を平準化させても 1925 年の頻度は他の出版年と同程度にはならないことが分かる。これは 1925 年の偏りが単純な数だけの問題ではなく、「樺島の法則」に見られるような質的な問題によって偏りを持っていることを示唆している。このため、PTA(L/A)であっても頻度で分析することは難しく、出版年ごとの割合で分析するしかないと考えられる。

３.５　PTA・ジャンル統制法による分析

PTA 割合分析により、一定の成果は得られたが、『太陽コーパス』にはまだジャンルのばらつきという問題が残っている。そこで「芸術・美術」という語がどのようなジャンルで使われているのか、出版年の経過によってどのように変化していくのかを観察してみる。

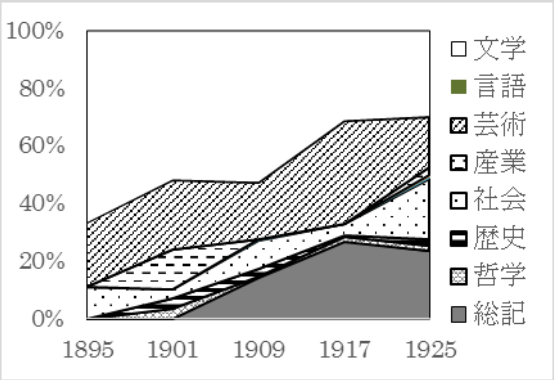

図 15 「芸術」が出現するジャンルの経年変化

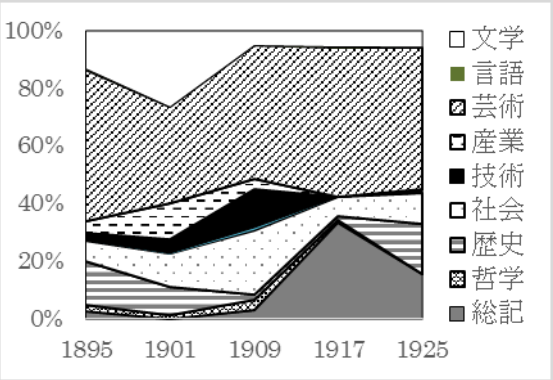

図 16 「美術」が出現するジャンルの経年変化

図 15 を見ると、「芸術」という語は、当初「文学」や「芸術」といった一部のジャンルでしか使用されていなかった。それが、年代を経るに従って多様なジャンルにも使用され、その使用割合も変容させていった様子が伺える。一方、図 16 の「美術」という語は、初めから多様なジャンルに使用され、基本的に同様の割合で使用され続けたように見える。

ただし、このグラフ自体は様々なジャンルが入り乱れている。例えば「産業」（横の鎖点）は、1901 年に多く現れるが、これは 1900 年にパリで行われた万国博覧会などの記事が所収されているからである。また、1909 年には「技術」（黒）が多く現れるが、これは黒田鵬心が 11 号と 13 号に仏教建築に関する長文の記事を書いているからである。『太陽コーパス』は時事的影響や編集方針がそのままデータに反映されるため、このような乱れが現れる。

しかし、近代日本語がこのように乱れていたわけではないだろう。近代日本語の平均な姿は、ジャンルの割合が一定か、緩やかな変化で推移していたと思われる。そこでここでは近代日本語において「美術」が使用されたジャンルの割合は一定であったと仮定し、全出版年のジャンル平均をすべての年代にあてはめた統制を行う。

分析は、ケースを PTA で重み付けし、「芸術・美術」を従属変数、出版年とジャンルを説明変数としたロジスティック回帰分析で行った[5]。変数は出版年も名義尺度とみなした。

[5] 「美術」における各偏回帰係数と標準誤差（カッコ内）は次の通り。P 値は** P<.01，* <.05 で表示。切片 –2.006** (.188)　1895 年 4.828** (.381)　1901 年 3.049** (.262)　1909 年 .503** (.160)　1917 年 .073 (.157)　（以上の参照カテゴリは 1925 年），総記 1.653** (.189)　哲学 2.026** (.437)　歴史 2.870** (.339)　社会科学 2.181 **(.203)　自然科学 1.879 * (.857)　技術・工学 5.480** (1.003)　産業 2.154** (.445)　芸術・美術 2.430** (.164)　言語 3.025 (1.780)（以上の参照カテゴリは文学）

ジャンルの効果は個々の偏回帰係数に個々の平均ジャンル割合をかけて求めた（この分析法を「PTA・ジャンル統制法」と命名する）。ロジスティック回帰分析を使用した言語分析の仕組みについては、横山・真田(2007a)を参照のこと。また、ロジスティック回帰分析を使用してデータの乱れを修正する方法は、横山・真田(2007b)に詳しい。

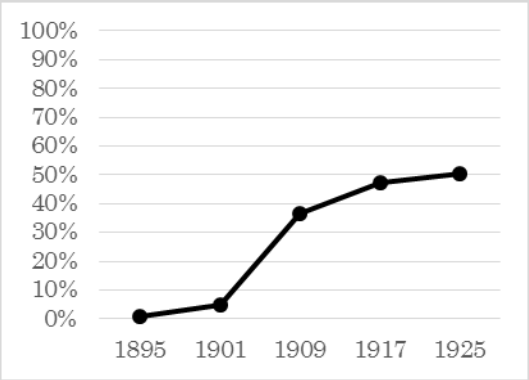

図 17　PTA・ジャンル統制法による芸術割合

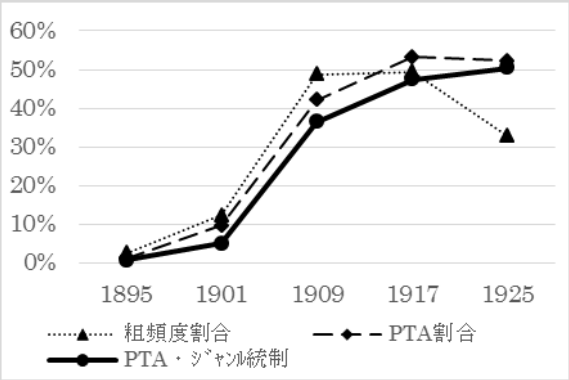

図 18 分析法による芸術割合の比較

PTA・ジャンル統制法の分析結果は、PTA 割合の分析結果より、全体的に「芸術割合」が低くなる。PTA・ジャンル統制法における 1925 年の値は 50.4%で、これはほぼ予想どおりの結果である。PTA 割合分析では 1917 年で一度高くなり、1925 年で若干落ち込んでいたが、ジャンルの統制を行うと一度も落ち込むことなく、全体がゆるやかな S 字カーブを描く。言語はその変化過程で S 字カーブを描くとされており、方言研究では井上(2000)などを初めとして多くの事例が報告されている。このような知見に合致する結果が得られたことからすると、PTA・ジャンル統制法による分析は、有効である可能性が高いだろう。

図 18 はこれまでに行った 3 種類の割合分析の結果を比較したものである。この３つのグラフの有効性を比較するため、条件を次の①～③にして「芸術・美術」を従属変数、出版年を説明変数としたロジスティック回帰分析を行った。①粗頻度を使用した場合の疑似決定係数は McFadden で.183、②PTA で重み付けした場合が.233、③PTA で重み付けし、さらにジャンルを説明変数に加えた場合が.349 となった。このことからしても、PTA・ジャンル統制法による分析がこの中では最も信頼できると考えられる。

４．代表性に配慮した分析法

これまで見てきたように、『太陽コーパス』は 1 記事文字数、出版年ごと記事数・文字数・ジャンルにばらつきがあり、さらに著作権問題によって一部のデータが非公開になっているなど、非常に不均衡なコーパスである。本稿ではこのばらつきを均衡化するため、5 種類の分析法を検討した。その中では、テキスト平均文字数当たりの調整頻度（PTA）を使用し、ロジスティック回帰分析によってジャンルを統制する分析法（PTA・ジャンル統制法）が最も有効だと考えられた。『太陽コーパス』に何らかの代表性があるとしても、その代表性の姿はデータの不均衡性によって乱されている。これを均衡にしていく分析法、すなわち均衡性に配慮した分析法がそのまま代表性に配慮した分析法にもなると考えられる。

ただし、ジャンルの統制を行うためには、近代日本語がどのようなジャンル割合になっているのかを広く調査し、それを使用するのでなければ真の代表性はないという考え方も成り立つ。田中(2012)によれば、国立国語研究所では「通時コーパス作成」のため近代語の

資料選定が行われているというが、その研究はまだ途上にあるようだ。外部資料としては『近代女性雑誌コーパス』の利用も考えられるが、データ量が少なく使用が難しい。このような状況下でも言語変化のS字カーブのように、すでに知られている知見との整合性に留意することで、一定の代表性は確保できると思われる。言語変化の分析を行った際、その結果があまりに歪んでいるなら、分析法を再検討する必要があるだろう。

もう一つ配慮が必要なのが、口語体と文語体の問題である。『太陽コーパス』では口語体と文語体が混在しているが、これらは語の選択や文法体系が異なり、基本的に異なったレジスターだと考えられる。Biber ほか(2003)では代表性の問題に関連して、「全体的一般化というものは、どのレジスターにとっても正確でないことが多く、むしろ現実には全く存在しないような言語の記述をしてしまうことになる。」(p.41.)と注意を喚起している。

本稿では詳しく触れられなかったが、「美術・芸術」の使用傾向において口語体・文語体による違いは見られなかった。しかし分析対象によっては口語体と文語体で異なる振る舞いを見せることも予想される。例えば文語体では使用率がどんどん増加するのに、口語体では徐々に使用率が減少する形式の場合、口語体と文語体を混合させて分析すると、あたかもその使用率は一定であるかのように見える。これは年代とともに文語体の記事自体が減少するためである。しかし近代日本語でそのような言語は、現実には全く存在しない。近代日本語では基本的に口語体と文語体という書き言葉しか、現実には存在していないのである。このため常に口語体と文語体による使い分けに留意し、使い分けがあった場合はこれらを分離して記述することが、もう一つの代表性に配慮した分析法だと思われる。

参考文献

石川慎一郎(2012)『ベーシック　コーパス言語学』ひつじ書房

井上史雄(2000)『東北方言の変遷』秋山書店

樺島忠雄(2009)「語彙量の実態」計量国語学(編)『計量国語学辞典』朝倉書店，pp.93-97.

国立国語研究所(編)(2005)『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』博文館新社

後藤 斉(1995)「言語研究のデータとしてのコーパスの概念について　―日本語のコーパス言語学のために―」『東北大学言語科学論集』4 . pp. 71-87.

後藤 斉(1996) 「コーパスとしての新聞記事データ―終助詞「かしら」をめぐって―」『東北大学言語学論集』5. pp. 37-46.

Stubbs, M. (2002) *Words and Phrases; Corpus Studies of Lexical Semantics*. Blackwell Publishing. 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

田中牧郎(2012)「近代語コーパスにおける資料選定の考え方」近代語コーパス設計のための文献言語研究 成果報告書 (国立国語研究所共同研究報告 12-03)

Biber, D.,Conrad,S., Reppen,R. (1998) *Corpus Linguistic; Investigating Language Structure and Use*. Cambridge University Press.　齊藤俊雄・朝尾幸次郎・山崎俊次・新井洋一・梅咲敦子・塚本聡(訳)(2003)『コーパス言語学　―言語構造と用法の研究―』南雲堂

横山詔一・真田治子(2007a)「フィールド言語学にロジスティック回帰分析は寄与しうるか」情報処理学会研究報告. 人文科学とコンピュータ研究会報告 (9) ， pp. 9-16.

横山詔一・真田治子(2007b)「多変量S字カーブによる言語変化の解析　―仮想方言データのシミュレーション―」『計量国語学』26-3 ， pp.79-93.