代表性に配慮した『太陽コーパス』の分析法再考

森　秀明（東北大学大学院文学研究科）†

Methodological Reconsideration on the 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。近代日本語の確立期をカバーしているため、語や文法の経年変化分析に使用されることが多い。しかし、代表性に配慮して設計されたサンプリングコーパスではないため、用例頻度や PMW で分析しても正確な結果が得られない場合がある。このため森（2014）では PTA という調整頻度で補正する分析を試みた。しかし、PTA の効果は限定的である上、代表性も担保できない。そこで今回はより代表性を有する分析法を検討した。この結果、著者名が判明している記事の記事数や分析対象の語が出現する記事の文字量で割合分析を行う方法がより有効であると考えられた。今後『太陽コーパス』で経年変化分析を行う場合は、用例頻度だけでなく、記事数や文字量でも分析することをお勧めしたい。

１．研究の目的

皆さんは『太陽コーパス』で用例検索を行った際、その調査結果に疑問を持ったことはないだろうか。『太陽コーパス』は本当に正確な値を示しているのか。そんな疑問から、森（2014）では『太陽コーパス』におけるデータの偏りを観察した。その結果、『太陽コーパス』では、記事の長さに 27 字～51,705 字というばらつきがあり、出版年ごとにジャンルの構成比も異なるため、用例頻度や PMW（Per Million Words：百万語当たりの出現頻度）で経年変化を比較しても、正確な分析にならない場合があると考えられた。そこで森 (2014) では PTA（Per Number of the Text Average Letters：一記事平均文字数当たりの頻度） という調整頻度を考案して記事の長さによる影響を均衡化し、ロジスティック回帰分析によってジャンルの偏りを補正する方法を試みた。しかし PTA は文字数に連動して用例頻度が増加しない語の分析ではあまり効果がない。しかもその補正結果が正確かどうかは、結局、外部の指標に頼るしかない。このため今回はより代表性を持った分析法を検討する。

２．『太陽コーパス』の代表性

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を代表性と言う。現在、コーパスの代表性を担保する方法には主に次の 2 つが用いられている。一つは、推定対象の言語をある程度反映している図書館の蔵書などを現実母集団とし、そこからデータを無作為抽出する方法。もう一つは、データを超大規模に収集することで自己均衡化させ、推定対象言語のコンパクトな相似形を作る方法である（マケナリー＆ハーディー，2014；石川，2012 など）。『太陽コーパス』は特定の雑誌の全文コーパスであるから、このような統計学的な意味での代表性は担保されていない。これまで『太陽コーパス』が代表性を持つと主張されてきた根拠は、田中（2012）で述べられている次の言葉に集約されている。

† hideaki@moriharuo.com

> コーパスの重要な要件のひとつである代表性の担保については、対象とした総合雑誌『太陽』が、分量の多さ、ジャンルの広さ、執筆陣の多彩さ、読者層の厚さの四点で、当時の文献資料としては格別の価値を持っていることから、『太陽コーパス』にも「代表性」が備わっていると見ることもできる。（田中，2012）

この主張は、これまでコーパス言語学で議論されてきた統計学的な意味での代表性とは異なる観点から「代表性」を主張したものである。このため、『太陽コーパス』がこれらの「代表性」を持っていても、用例頻度が統計学的に正確な値を出すことは担保されない。例えば 1925 年に日本で出版された書籍の中でアジアという地名が使用された回数に対し、1925 年の雑誌『太陽』に出現するアジアという地名の用例頻度がその何万分の一かの縮尺になっている可能性は担保できない。その可能性を確実に担保するには、1925 年に出版された書籍から無作為サンプリングを行ってコーパスを作る以外、方法はないと考えられる。

その一方で、田中（2012）が指摘する「分量の多さ、ジャンルの広さ、執筆陣の多彩さ、読者層の厚さ」という 4 つの特徴は、図書館書籍の性格とよく似ている。図書館の蔵書はある年に出版された書籍の中で、特に流通量が多かったものを中心に、社会的な需要を考慮して幅広いジャンルの書籍が集積されたものだ。雑誌『太陽』は、博文館が当時刊行していた『日本商業雑誌』『日本大家論集』『日本農業雑誌』『日本之法律』『婦女雑誌』を廃刊して一冊に統合した総合雑誌である。その内容は「百科全書的」で、創刊号は 28 万 5 千部、創刊以後約 10 年間は 10 万部弱の発行数があったと言われている（上野，2007）。雑誌『太陽』は単一の雑誌ではあっても、そのジャンルの広さや当時を代表する執筆陣、流通規模の大きさから、図書館書籍のミニチュア版的な性格を持ち合わせていると見なすことができる[1]。雑誌『太陽』が、統計学的に図書館書籍のミニチュアになっているのなら、『太陽コーパス』は堂々たる代表性を持っていると言えるだろう。これは『現代日本語書き言葉均衡コーパス』（以下 BCCWJ と呼ぶ）の「図書館書籍」が代表性を持っているという議論と同じである。

しかし、用例レベルで考えた場合、ある年に出版され図書館に収蔵された書籍の用例に対し、同じ年に雑誌『太陽』に書かれた記事の用例が、統計学的に一定の縮尺になっている保証はない。図書館書籍でアジアという語が使用される回数と雑誌『太陽』でアジアが使用されている回数を結びつける統計学的な根拠が見出し難いからである。

だが、著者を基準に考えた場合はどうであろうか。ある年の図書館書籍の著者の多くは、雑誌『太陽』の記事を書いた著者の多くと重なっているのではないか。雑誌『太陽』には当時を代表する執筆陣が記事を書いている。図書館に収蔵される書籍も当時を代表する書籍である。その著者の多くが一致している可能性はかなり高いと考えられる。当時の平均的な図書館の蔵書目録を入手し、その著者名と雑誌『太陽』の著者名の多くが一致しているなら、『太陽コーパス』は著者レベルでは、統計学的に一定の代表性を持っていると言っても過言ではないだろう。

しかし、残念ながらこの検証は難しい。当時は図書館が未整備で、毎年一定数の書籍を

[1] 『太陽』は 1928 年（昭和 3 年）2 月に廃刊となる。廃刊当時の流通量は不明だが、その量が激減していたことは想像に難くない。この意味で、田中(2012)が指摘する 4 つの特色がどの年代まで保たれていたかは、今後十分に検討していく必要がある。

安定して購入できるような体制にはなかった。内閣統計局（1912）『日本帝国統計年鑑　第31』（p. 553）[2]によれば、1910年の図書館数は全国で374館（官立・私立の合計）、その蔵書合計は2,643,264冊で平均7,000冊程度である。しかも中には1,000冊前後しかない図書館もある。当時の平均的な図書館像を決めるのも難しく、当時の蔵書目録を入手するのはさらに困難である。このためここで著者レベルでの『太陽コーパス』の代表性を実証することは難しい。

ただし、大まかな目安ならつけられる。表1は、当時の書籍の出版数と、『太陽コーパス』で氏名が判明している著者数である。

表1 近代の出版物数[3]と『太陽コーパス』の氏名判別著者数

	1895年	1901年	1909年	1917年	1925年
著述	8,334	18,963	34,066	46,012	18,028
編集	17,712				
翻訳	124	35	57	118	
合計	26,170	18,998	34,123	46,130	
『太陽コーパス』氏名判別著者数	238	212	245	155	245

使用した統計書は年によって集計の仕方が異なるが、基本的に著述は普通出版物、編集は雑誌だと思われる。表1の「著述」の冊数がBCCWJで言えばその年に出版された全ての書籍の数＝「出版書籍」の母集団の数である。表1からごく荒く推定すれば年1、2万冊が出版書籍の母集団の数となる。ここから図書館に収蔵する書籍を選ぶとして、平均7,000冊しか蔵書のない図書館が、毎年何千冊も追加購入することは考えにくい。かといってあまりに少ない冊数では、図書館書籍自体が近代日本語の代表性を失ってしまう。いま仮に推定出版書籍数のおよそ1/10～1/20に当たる1,000冊を一年当たりに購入される図書館書籍の母集団だとしてみよう。この1,000冊を著者1,000人と読み替えるなら、その1,000人の中に『太陽コーパス』の氏名判別著者が含まれている可能性はかなり高いと言えるだろう。今、その割合が何%になるのかは分からない。しかし、重要なことは用例頻度の場合その代表性を担保する統計学的な根拠は見出し難いが、著者数で考えれば確実に何%かの代表性は担保できるということである。著者数で分析する場合、「『太陽コーパス』には代表性がない」という帰無仮説は統計学的な根拠を持って棄却されると考えられる。

３．指標としての記事数

言語の経年変化を分析する場合、用例頻度で分析するということは、例えばアジアと言う地名に対して「亜細亜」という漢字表記が何例出現し、「アジア」というカタカナ表記が何例出現しているかを調べ、その割合の変化を観察することである。一方これを著者数で観察するということは、例えば代表性を持った1,000人の中で何人が漢字で表記し、何人がカタカナで表記するかの割合の変化を見ることである。厳密に言えば用例頻度割合と著者数割合は異なる現象を観察していることになる。しかし言語変化は、つまるところそれを使用する人間の言葉遣いの変化であるから、著者数割合を使用しても言語学的に意義の

[2] http://kindai.ndl.go.jp/info:ndljp/pid/974420（2015.01.31 閲覧）

[3] 1895～1909年は『大日本帝国内務省統計報告』、1910年～1925年は『日本帝国統計年鑑』による。http://kindai.ndl.go.jp/（2015.01.31 閲覧）

ある観察をしていると考えられる。

ただし、同じ著者でも学術的な論文の場合は漢字で表記し、大衆的な読み物の場合はカタカナで表記することも考えられる。このため、一冊の書籍や一つの記事を単位とし、その書籍や記事が漢字表記、カタカナ表記、併用、未使用のどれになるかを観察した方がより実際的だと思われる。このように記事数と言う単位で観察しても、その根本は著者に根ざしているため、この記事数も一定の代表性を持っていると考えられる。

問題は、その代表性がどれぐらいあるかである。母集団 1,000 人のうち『太陽コーパス』と一致している著者が 100 人しかいない場合、代表性は 10%しかないように思える。しかし、『太陽コーパス』の 100 人が母集団のごく平均的な傾向を示しているなら、例えば 1909 年や 1925 年の著者数は 245 人であるから、100 人÷245 人＝40.8%は母集団のごく平均的な傾向を示していることになる。残りの 145 人だけが非常に偏った表記法を使用しているとは想定しにくいので、『太陽コーパス』が相当の割合で母集団の正確な姿を反映している可能性がある。その一方で母集団と一致した 100 人が平均より偏った表記法を使用していた場合、『太陽コーパス』が母集団平均と大きくかけ離れた姿をしていることも考えられる。

この問題は分析対象の言語現象にどのような要因が影響しているかに関わっている。例えば外国地名を漢字表記するかカタカナ表記するかの場合なら、学術書などの硬い文章では漢字が用いられ、大衆向けの柔らかい文章ではカタカナが用いられることなどが考えられる。これをジャンル的に見れば、社会科学などは漢字が使われやすく、文学などではカタカナが使われやすいなどの現象となって現れる可能性がある。雑誌『太陽』の編集方針が学術的な記事に偏っていたり、ジャンル構成が母集団の傾向と大きく異なっている場合、『太陽コーパス』の代表性は低い可能性がある。その逆に当時の母集団平均と同じような文章の硬軟度やジャンル構成で編集されていたとしたら、『太陽コーパス』の代表性は高い可能性がある。これ以上は想像の域を出ないが、雑誌『太陽』が百科全書的な総合雑誌であり、商業的に大きな成功をおさめた雑誌であることを考えれば、『太陽コーパス』の代表性が高い場合の方が多いのではないかと思われる。

ここまでは、『太陽コーパス』の中で著者名が判明している記事を対象に考察してきた。『太陽コーパス』の中で、著者名が判明している記事はおよそ 7 割である。残りの 3 割は無署名でその多くは雑誌記者が執筆していると考えられる。これらの無署名記事はどのように扱えばよいだろうか。これまでの代表性の議論から言えば、雑誌記者が図書館書籍の母集団に含まれている可能性は低いと思われる。また、雑誌記者の場合、編集部の方針によって表記法などの言葉遣いに一定の制約がかかっている可能性もある。このため基本的に無署名記事は除いて分析した方が正確な結果が得られると考えられる。

特に無署名記事では表 2 に見られる〈小話〉〈世界のラヂオ〉〈新刊紹介〉などのように、同じ号に同じ題名で書かれた複数の短文記事が観察される（以後、これを同号同名記事と呼ぶ）。これらは本来ならまとめて一つの記事として掲載されてもおかしくない内容だが、雑誌を読みやすくする意図からか、特に 1925 年の長文記事の間に埋め込まれるように編集されている。これらを別々の一記事と認定すると、同一の著者と思われる無署名記事を何回もカウントしてしまうため、同一著者の言葉遣いを過大に評価してしまうことになる。同号同名記事を統合して一記事と見なした上で署名記事の言葉遣いと比較し、その傾向に大きな違いがあるなら、これらを分離して観察する方法が妥当だと思われる。

表 2　1925 年 04 号の記事配列（開始から 20 記事目まで／全 78 記事）

No.	題名	文字数	No.	題名	文字数
1	昨年の今月	654	11	日米海軍勢力の比較	5,337
2	普選実施後の政党	9,408	12	〈世界のラヂオ〉	267
3	〈和田豊治氏母堂米寿に寄せられた詩歌〉	434	13	明治初年外交物語（その七）苦心の犯人捜索	7,176
4	時事漫吟	905	14	〈世界のラヂオ〉	583
5	〈小話〉	126	15	新人有馬頼寧	5,650
6	赤露印象記	6,276	16	〈冬の日に〉　丹下生	82
7	〈世界のラヂオ〉	634	17	〈小話〉	65
8	普選実施の影響と女子参政権問題	6,458	18	戦場の悪戯者―空想の兵器―運命の弾丸―	7,364
9	〈世界のラヂオ〉	329	19	〈小話〉	65
10	〈新刊紹介〉	570	20	今は我れ　丹下生	42

４．指標としての文字量

記事数という指標は、一定の統計学的な代表性を有していると考えられる。しかし、『太陽コーパス』の記事には 27 字～51,705 字というばらつきがある。記事数で分析する場合、27 字の記事も 51,705 字の記事も同じ 1 記事となるが、その扱いで良いものだろうか。

図書館書籍を日本語の代表と見なす考え方の中は、その当時、大量に流通していた書籍の方が日本語の代表としてふさわしいという前提があると思われる。短い記事しか依頼されない著者と長い記事を依頼される著者では、日本語を代表する代表度に差があると考えられる。例えば 1,000 字の記事 10 本に外国地名がカタカナ表記されていたとする。一方、10,000 字の記事では漢字表記されていたとする。その場合、カタカナ：漢字の比率は 10：1 でいいのだろうか。これが口語・文語の割合ならどうだろう。1,000 字の口語記事 10 本と 10,000 字の文語記事 1 本の場合、雑誌の口語：文語比率は本当に 10：1 でいいのだろうか。

雑誌の編集者の立場で考えた場合、記事の硬さ・柔らかさの比率や、口語・文語の比率は、当然コントロールの対象になったと思われる。これらの分量を最も読者層に受け入れられやすい比率とすることで、雑誌の販売量の最大化を図ったと考えられる。このように編集者が市場のニーズに配慮することによって反映された代表性を「市場代表性」と名付けるなら、記事数より文字量の方が市場代表性が高いと考えられる。つまり先の例でいえば、10：1 ではなく 1：1 と数える方が、より市場代表性を反映していると考えられる。

記事の硬さ・柔らかさや口語・文語の比率などは、言葉遣いの比率に大きな影響を与える。特に言語の交替現象を観察する場合、新しく使用されるようになった言葉遣いは、まず、話し言葉や柔らかい記事から使用される傾向がある。この割合がコントロールされた文字量は記事数以上に母集団の正確な姿を反映している可能性がある。また、雑誌の編集者は無署名記事も含めて様々なコントロールを行っていたと考えられるため、無署名記事を削除しない方がより市場代表性を有している可能性がある。ただし、このような市場代表性は、統計学的に立証できる類のものではないと思われる。このため、統計学的に一定の代表性を有すると考えられる記事数と併用しながら、比較検討する方法が妥当であろう。

５．ケーススタディ

ここでは 2 つの先行研究を取り上げ、記事数、文字量を指標とした割合分析の有効性と問題点を検討する。記事数、文字量を指標とするだけでなく割合分析も行うのは、『太陽

コーパス』における出版年ごとの不均衡性を平準化するためである。これまで割合分析は主に言語現象を観察する目的で使用されてきたが、出版年の影響を除く効果も高いと考えられる。例えば外国地名表記の経年変化を調べる場合、出版年ごとの文字数や記事数が異なるため、単純な頻度では比較できない。これを割合分析すればこれらの要因は相殺されて比較可能な値になると考えられる。

$$\text{カタカナ割合} = \frac{\text{カタカナの頻度} \times \cancel{\text{出版年の影響}}}{(\text{カタカナの頻度} + \text{漢字の頻度}) \times \cancel{\text{出版年の影響}}}$$

５.１井出（2005）「外国地名表記について―漢字表記からカタカナ表記へ―」の再分析

井出（2005）は、外国地名が漢字表記からカタカナ表記へ移り変わっていく経年変化を分析した研究である。この研究では、先駆的な試みとして分析の指標に記事数が使用されている。初めに井出（2005）が記事数を指標に採用した考え方を見てみよう。

> 頻度ではなく記事数を指標にしたのは，地名の場合，記事の種類によって，同一記事内に同一語が繰り返して出現している場合があり，頻度よりも記事数の方が指標としてまさっていると考えられるからである。年代別の使用の推移を見ようとするなら，一つの記事に何語出現するかということは無視し，出現した記事を１として数えた方がより正確にその推移の変化を見ることができると思われる。（井出，2005，p. 159）

井出（2005）では、地名のような特徴語[4]的性格を示す語の場合、用例頻度より記事数の方が正確だと主張されている。しかし、なぜ記事数の方が指標として優っているのかについて、理論的な考察がなされていない。このため、井出（2005）では、同号同名記事を統合する必要性や署名記事と無署名記事を分離して観察する必要性について、検討されていない。井出（2005）では、最終的に 1925 年にカタカナ表記が急激に増加したと結論づけられているが（p. 170）、その結論には疑問が残る。以下、これを再分析してみる。

井出（2005）では、21 の地名について個別に観察が行われている。しかし、21 の地名ごとに分析した結果、分析に適さないほどデータ数が少なくなっている地名が散見される。計量分析では少しでもデータ数が多い方がより正確な分析となることから、ここでは 21 の地名を合計した分析を行う。初めに用例頻度、記事数、文字量を指標とし、割合分析を行わずに経年変化を観察する。ここで使用するのは記事を統合したり無署名記事を除いたりしない、全数での観察である。

図 1 の用例頻度を観察すると、1917 年の漢字地名がそれまでの 2 倍弱使用されていることが目につく。図 2 で 1917 年の記事数を観察すると、記事数はむしろ減少していることから、この現象は一記事当たりに使用されている漢字地名が増えていることを意味している。1917 年は 1914 年に始まった第一次世界大戦や 1917 年に起きたロシア革命に関する記事などが多く、増加の原因にはそれらの記事で漢字地名が多用されたことが考えられる。問題

[4] 特徴語とは、あるテキストに頻出し、そのテキストの性格を特徴づけるような語を意味する。例えば海外の事情を紹介したテキストなどでは外国地名が頻出し、それが特徴語となる場合がある。美術・芸術、戦争・平和などのように、テキストのテーマに深くかかわる語は、特徴語となる可能性がある。

はこのような増加が雑誌『太陽』独自の現象なのか、日本語全体の現象なのかである。第 3 節で想定した例で考えれば、図書館書籍 1,000 冊から用例を抽出しても図 1 のような現象が観察されるなら、日本語全体の現象と言える。しかし、様々なジャンルの書籍 1,000 冊の合計で、なお漢字を使用した外国地名がそれまでの 2 倍弱にもなることは考えにくい。よって、この用例頻度はあくまでも雑誌『太陽』の姿を現したものと思われる。

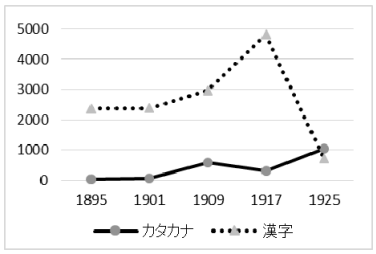


図 1 表記別外国地名用例頻度

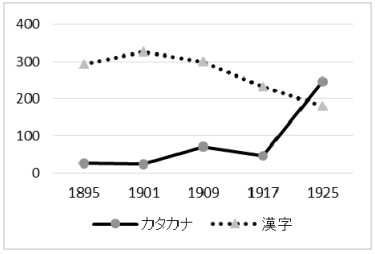


図 2 表記別外国地名記事数

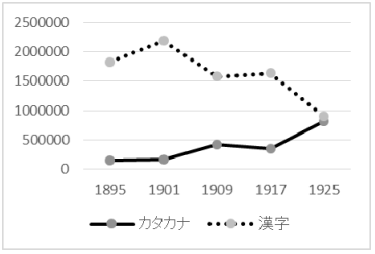


図 3 表記別外国地名記事の文字量

図 2 では、1925 年で外国地名をカタカナで表記する記事の本数が急増する現象が目につく。これと図 3 の文字量を比較すると、外国地名をカタカナで表記する記事の文字量はさほど増加していない。図 2 の現象は 1925 年のカタカナ表記をしている記事が、ごく短い文字数で書かれ、さらにその記事数が多いことを示している。これには表 2 で観察した同号同名記事の問題が反映されていると考えられる。同号同名記事は同一著者（または同一の属性を持った複数の雑誌記者）によって書かれていると思われ、これを重複してカウントすると著者を単位にした正確な分析はできない。図 3 は文字量である。文字量には、統計学的な代表性は考えにくく、読者のニーズを反映した市場代表性が推定されるだけである。しかし、図 3 を見る限り、図 1、2 に見られるような明らかな偏りは観察されない。

次に同号同名記事を統合した場合の記事数を観察する（以後これを統合記事数、統合前の記事数を単純記事数と呼び分ける）。図 4 は、統合記事数のグラフである。同号同名記事を統合した結果、1925 年の偏りは解消され、図 3 の文字量のグラフに近くなった。

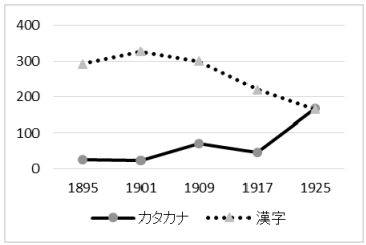


図 4 表記別外国地名統合記事数

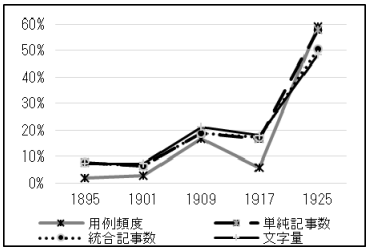
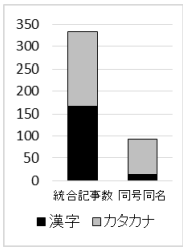
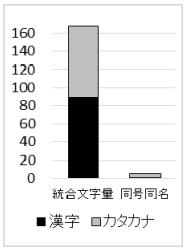


図 5 外国地名の指標別カタカナ割合



縦軸：記事数
図 6 記事数



縦軸：万字
図 7 文字量

図 5 は、用例頻度、単独記事数、統合記事数、文字量を指標として算出したカタカナ割合である。統合記事数と文字量のグラフの形状はほぼ一致し、1925 年の値が約 50%になる。一方、単純記事数は 1917 年まではこれらと同じだが、1925 年は 60%弱で、用例頻度の値と同じになる。図 6 は統合記事数と同号同名記事の本数を比較したグラフである。これを見るとカタカナを使用した同号同名記事だけで約 100 本になることが分かる。図 7 は同じものを文字量で描いたグラフである。文字量に直すと、カタカナを使用した同号同名記事は

約 1.4 万字しかなく、ほとんど影響力を持っていない。井出（2005）は、単純記事数に基づいて分析したため、1925 年のカタカナ割合を過大評価していると考えられる。

ただし、図 5 の統合記事数や文字量割合のグラフが直ちに代表性を持っているとは見なし難い。図 8 は、一記事当たりに 1、2 回しか外国地名が出現しない低頻度出現記事と、一記事当たりに 3 回～366 回出現する高頻度出現記事に分け、さらに著者名が判明しているかいないかを加味して全体を 4 つのグループに分けたグラフである。指標には文字量を使用している。今、議論を単純化するために低頻度記事を一般記事、高頻度記事を専門記事と見なすと、著者名が判明している一般記事では、カタカナ割合は一定の割合で増加していたことが分かる。著者不明の記事は、雑誌『太陽』の記者による記事と思われるため、これらのカタカナ割合は編集方針によって統制されていた可能性がある。著者名が判明している専門記事も類似の傾向を示しているが、総じてカタカナ割合が高い。

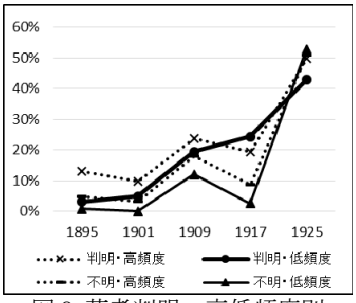


図 8 著者判明・高低頻度別カタカナ割合

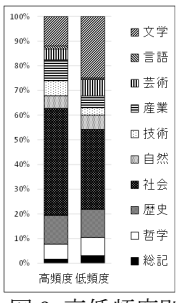


図 9 高低頻度別ジャンル

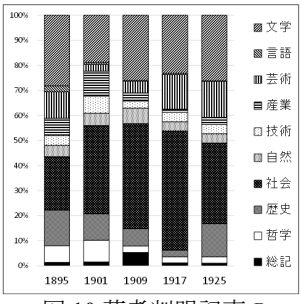


図 10 著者判明記事の出版年別ジャンル

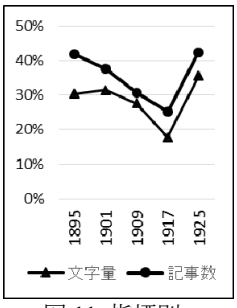


図 11 指標別低頻度記事割合

図 9 は、図 8 の著者判明記事のジャンルを高低頻度別に描いたグラフである。高頻度記事では社会のジャンルが多く、低頻度記事では社会が減って文学が増えている。図 10 は著者判明記事のジャンルを出版年ごとに描いたものである。ジャンル構成は出版年によって変化しており、特に 1909 年と 1917 年で社会のジャンルが多い。図 11 は文字量と記事数の指標別に著者判明記事の中で低頻度記事がどれぐらいの割合になるかを示したものである。特に 1909 年と 1917 年で低頻度記事が低下している。図 10 のグラフと図 11 のグラフには連動性が見られる。

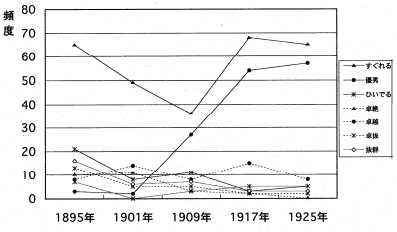
図 8 において、代表性が担保できるのは著者判明のグラフである。これらの高頻度：低頻度記事の割合は、図 11 のようにおよそ 6：4（記事数）または 7：3（文字量）となるため、そのまま合計すると高頻度記事の影響が強く出る。この結果、『太陽コーパス』の著者判明記事割合は図 5 の統合記事数のグラフに近くなる。しかし 1909 年や 1917 年にはジャンルや高低頻度割合の偏りがある。これを補正した場合、特に 1917 年の落ち込みは図 5 より少なくなると考えられる。このため、正確なカタカナ割合は図 5 の統合記事数から図 8 の判明・低頻度の形状にもう少し近づくと思われる。つまり、外国地名のカタカナ割合は 1925 年に急増するのではなく、一定の割合で徐々に増加していた可能性が考えられる。

以上の観察から、用例頻度、単純記事数、無署名記事を使用すると、分析が不正確になる例が確認された。また、著者判明記事の記事数は一定の代表性を持つと考えられるものの、ジャンル等で言葉遣いの使い分けがなされている言語現象では、『太陽コーパス』におけるジャンルの偏りを補正しないと、高い代表性は見込めないことが考えられる。

５.２　田中（2005）「漢語「優秀」の定着と語彙形成―主体を表す語の分析を通して―」の再分析

田中（2005）は明治期に新しく作られた「優秀」という漢語が、「卓越、卓絶、卓抜、抜群」といった古くからある漢語（以後「卓越類」と呼ぶ）や、「すぐれる」といった和語とのかかわりの中で、どのように定着していったのかを分析した研究である。その結果、「漢語「優秀」は，和語「すぐれる」との間に意味的な使い分けをもったことで，語彙の基本的な部分に深く浸透したものと考えられる。」（p. 139）と考察されている。これは、用例の統語的な分析を詳細に行った結果から導かれた結論だが、ここではごく単純に全体の数量的な観点から再分析してみる。

図 12 は田中（2005）に掲載されている用例頻度のグラフである。先にも述べたが、『太陽コーパス』では出版年ごとの文字数や記事数が一定でないため、用例頻度そのものでは偏りが出る。このため、用例頻度を使用して割合分析を行ったグラフが図 13 である[5]。この際、「卓越類」は合計して集計した。図 13 を見ると「優秀」と数量的に競合しているのは「卓越類」であり、「すぐれる」は数量的にはほぼ無関係であることが観察される。



田中（2005）より引用（p. 134 ）

図 12〈優秀〉語彙の年次別用例頻度

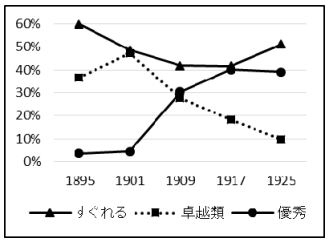


図 13〈優秀〉語彙の年次別用例頻度割合

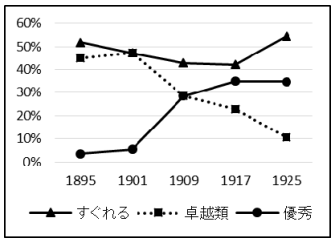


図 14〈優秀〉語彙の年次別統合記事数割合

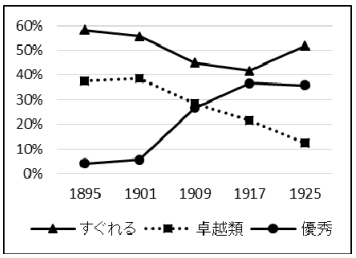


図 15〈優秀〉語彙の年次別著者判明記事数割合

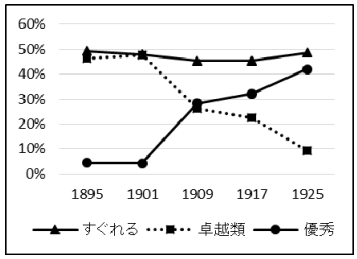


図 16〈優秀〉語彙の年次別文字量割合

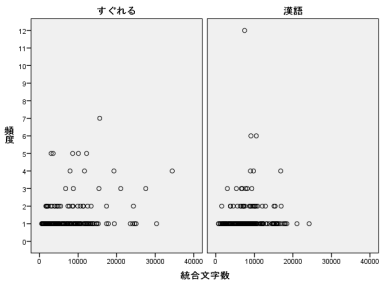


図 17 「すぐれる」と〈優秀〉漢語語彙の文字数別散布図

図 14～16 は、少しずつ形は変化するものの、基本的に図 13 と同じ形状をしている。第 2 節で行った代表性の議論からすれば、この中で統計学的な代表性を持つと考えられるのは図 15 であり、図 13 の用例頻度では代表性が担保できないはずであった。それなのになぜこれほど形状が似ているのであろうか。その理由は、図 17 の散布図にある。図 17 は、記事の文字量を横軸に、一記事当たりの使用回数を縦軸にして描いた散布図である。これを見ると、一記事に用例が 1 回しか出現しない記事が最も多く、大半は 2 回までの出現にとどまっている。この傾向はどんなに文字数が多い記事でも基本的に変わらない。用例頻度

[5] データは発表者が現行の『太陽コーパス』から抽出したものを使用している。また、1925 年 01 号阪谷芳郎「近代文明と発明」は外れ値とみなして除いてある。またこれ以後のグラフでは論点を絞り込むため「ひいでる」は描いていない。

が一記事当たり 1 回であれば、用例頻度と記事数は完全に同一になる。これが平均 2 回になったとしても、互いの出現傾向が同じであれば、割り算をすれば記事数割合と同じになる。代表性が担保できないはずの図 13 が一定の代表性を有すると考えられる図 15 とよく似たグラフになるのは、用例頻度を使用しても、その割合分析の結果が記事数割合とほぼ同様の結果となるからである。つまり、用例頻度を使用しても、割合分析の結果が記事数割合と似た値になる語の場合、概ね正確な分析結果を示すと考えられる。

これらに比べ、図 16 の文字量のグラフは「すぐれる」がほぼ直線的に推移して形状がやや異なる。この理由は「すぐれる」が和語であり、小説や雑学的な記事に現れやすいためだと思われる。小説の文字数は長いものが多く、雑学的な記事は短いものが多い。これらの割合は記事数的には出版年ごとのばらつきがあるが、文字量から見れば常に 5 割前後になっている。これは「すぐれる」と言う語が使用されるタイプの記事が、全ての出版年を通じてほぼ一定であることを示唆しているのかも知れない。第 3 節で検討した市場代表性を重く見れば、図 16 の方が正確な近代日本語の姿を示しているとも考えられる。

以上の観察から、用例頻度割合でも概ね正確な分析となる例が確認された。ただし、それは検索語がどの記事にも同程度の回数で使用され、結果的に用例頻度割合が記事数割合と同じになるからだと考えられる。

６．まとめ

これまで『太陽コーパス』の分析では、用例頻度を使用した研究が多かった。しかし、用例頻度は代表性を統計学的に担保することが難しい。その一方で著者名が判明している記事数は、統計学的に一定の代表性を担保できると考えられる。また、統計学的な証明は難しいが、用例が出現する記事の文字量は、読者のニーズを反映した市場代表性を有していると考えられる。ただし、この 3 種類の指標は、厳密には別々の現象を表していると考えられる。このため、『太陽コーパス』の分析に当たっては、これら 3 種類の指標を併用し、その振る舞いの違いを観察していく分析法が有効だと思われる。

文　献

井出順子（2005）「外国地名表記について―漢字表記からカタカナ表記へ―」国立国語研究所（編）『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，pp. 157-172.

石川慎一郎（2012）『ベーシック　コーパス言語学』ひつじ書房.

上野隆生（2007）「研究プロジェクト 日本近代化の問題点--明治国家形成期の明と暗 雑誌『太陽』の一側面について」『東西南北』2007，和光大学総合文化研究所，pp. 252-285.

田中牧郎（2005）「漢語「優秀」の定着と語彙形成―主体を表す語の分析を通して―」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，pp. 115-141.

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）.

マケナリー&ハーディー（2014）石川慎一郎（訳）『概説コーパス言語学－手法・理論・実践』ひつじ書房．[McEnery, T.& Hardie, A.（2012）*Corpus Linguistics; Method, Theory and Practice.* Cambridge University Press.]

森秀明（2014）「均衡性と代表性に配慮した『太陽コーパス』の分析法試論」『第 5 回コーパス日本語学ワークショップ予稿集』国立国語研究所，pp. 73-82.