福山平成大学経営学部紀要
第 15 号（2019），63－78 頁

コレスポンデンス分析を用いた英文テキスト分類における語数調整法と単語の選択基準

福井正康[*1]・渡辺清美[*1]

[*1] 福山平成大学経営学部経営学科

要旨：コレスポンデンス分析（対応分析）を用いたテキスト分類では、テキストからの単語選択の方法に決まりがなく、利用者の経験を頼りに分析を行うことが多かった。これに対して、著者らは例文を使った試行的な分析によって、安定的な結果を導く語数の調整法と利用する単語の選択基準を求めた。その結果、二重調整法及び、0 比率に基づく単語選択という新しい２つの方法を提案した。

キーワード：英文テキスト分類、コレスポンデンス分析、対応分析、語数調整

1. はじめに

コレスポンデンス分析（対応分析）を用いたテキストの分類は、新しい手法として、近年用いられるようになってきた[1]-[3]。その方法には、すべての単語を用いる方法、出現語数の高い単語だけを用いる方法、またはテキストごとに単語の総数（以後、総語数とする）を調整してそろえたデータを利用する方法などが考えられる。しかし、どのような方法が標準的か、あまり議論されることなく、分析者の判断で使われているのが現状である。

著者らはこのような状況に対してコレスポンデンス分析の安定性を疑問視してきた。語数調整法や単語の選び方によって、結果が大きく変わることはないのであろうか。この論文は、この問題に対して、どのように単語の語数を調整するのか、どのように単語を選ぶのかという問いについて考えようとするものである。

コレスポンデンス分析は 2 次元分割表を用いた行変数と列変数の分類手法である。例えば図 1 左のような分割表からは、同図右のような結果が得られ、図 2 左のような分割表からは同図右のような結果が得られる。但し、ここでは行成分について図に加えていない。２つの分割表を比較すると、図 2 の分割表の D 社の列が図 1 の同じ列の 10 倍になっている。この絶対数だけの違いにより、結果は大きく異なってくる。

	A社	B社	C社	D社
1	10	19	13	5
2	13	8	15	16
3	18	11	14	8
4	16	8	13	10
5	12	10	6	8
6	3	12	7	11

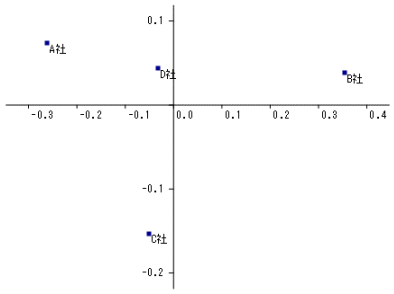


図 1　コレスポンデンス分析結果 1

	A社	B社	C社	D社
1	10	19	13	50
2	13	8	15	160
3	18	11	14	80
4	16	8	13	100
5	12	10	6	80
6	3	12	7	110

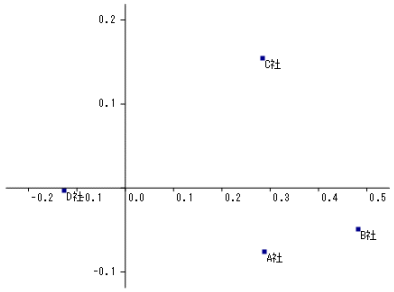


図 2　コレスポンデンス分析結果 2

実際著者らが行っている英語教科書のコレスポンデンス分析による分類では、教科書によって単語数や総語数がかなり異なることがある。著者らはこの問題に対して、すべての教科書の単語の総語数をある一定値に定め、その数をもとに各単語の語数を比例配分して求め、教科書ごとの総語数の違いを取り除いてきた。その後、全教科書の合計で出現語数の多い単語から適当な個数を選び出して分析に利用した[4]-[6]。この処理は分析者の経験によるところが大きく、著者らが常々正当性に不安を感じてきたところである。この論文では、まず単語の選び方による結果の安定性の問題を中学 1 年の歴史的な英語の教科書を題材にして議論し、その後に新しく提案する単語の調整法について説明する。最後に、著者らが提案する方法をいくつかの状況の異なる場合に適用してみる。

適用する例として、まず教科書の数を増やした場合について議論する。これは同じ次元数で比較すると、寄与率を下げる問題である。寄与率を下げた場合、もちろん 2 次元で表示するには問題があることは承知しているが、選択する語数によって、結果にどのような差があるのだろうか。次に、教科書の学年を中学 3 年に変えて、同じ分析を実行する。教科書の難易度が上がることにより、単語の種類が増え、総語数も増加する。次に教科書以外の入試問題を試してみる。この場合は、教科書と異なり、単語の種類は増えるが、全体の総語数は少なくなる問題である。以上述べたような状況に著者らの提案する方法を適用し、どのような点に注意すべきか考察する。

2. コレスポンデンス分析による教科書分類の安定性

前章でコレスポンデンス分析がある列の度数の大きさに影響を受けることを見たが、この章では 2 組の英語教科書を用いて、語数調整や利用する単語数によって、結果がどのように変わるか調べてみる。使用する教科書の 1 組目は単語の総語数が大きく異なる教科書の組である。それに含まれる教科書（単語数，総語数）は以下のとおりである。

Choice-1（466, 5629）、Dening-1（3844,38435）、Kanda-p1（200, 1732）、Seisoku-1（736, 13915）、Sunshine-1（340,1291）、Union-1（935,5476）　以上全単語数 4555

ここに、単語数は単語の種類数、総語数は各単語の出現回数の総和で、固有名詞は除いている。

この組の場合、教科書の単語の総語数は 1291 から 38435 と約 30 倍もの開きがある。最初にこれらの教科書に現れるすべての単語をまとめて、教科書全体で出現頻度の高い単語の順に並べる。全 4555 種類の単語の中で、頻度の高い方から、2000 語、1000 語、500 語、200 語、100 語、50 語を取り出して、コレスポンデンス分析を実行する。ここで注意することは、例えば 2000 語取り出すとすると、その中には 1 つまたはいくつかの教科書だけで使われて、他の教科書では使われない、すなわち他の教科書では頻度が 0 の単語が含まれている。このように、教科書数×単語数の度数の中で、頻度が 0 となる割合を「0 比率」と呼んで、使用頻度の低い単語が混じる 1 つの指標とする。この指標があまり高い値だと、単語の用法の比較というよりはむしろ教科書独自の話題の比較になりかねない。

コレスポンデンス分析の実行結果を散布図にして図 3.1 から図 3.6 に示す。以後この論文では、図のキャプションにつけた括弧の中の数値は「/」より左が 0 比率、右が 2 次元までの累積寄与率とする。例えば、最初の（0.761 / 0.574）は、0 比率が 0.761、2 次元までの累積寄与率が 0.574 である。

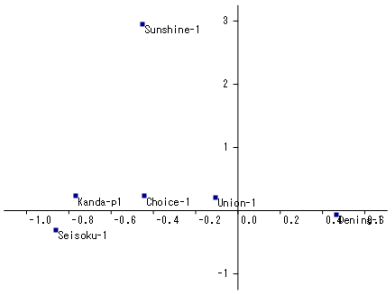


図 3.1　2000 語（0.676 / 0.567）

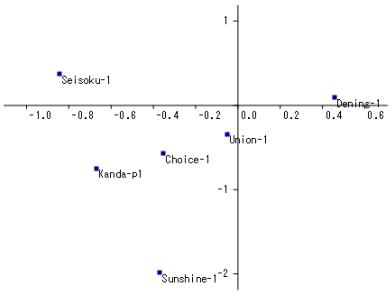


図 3.2　1000 語（0.574 / 0.602）

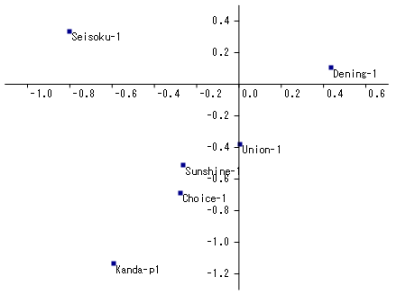


図 3.3　500 語（0.423 / 0.665）

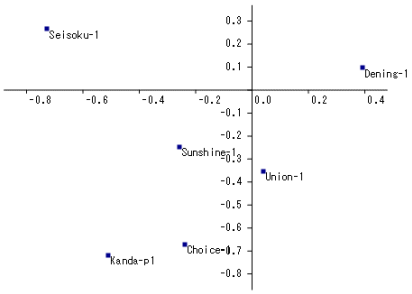


図 3.4　200 語（0.233 / 0.800）

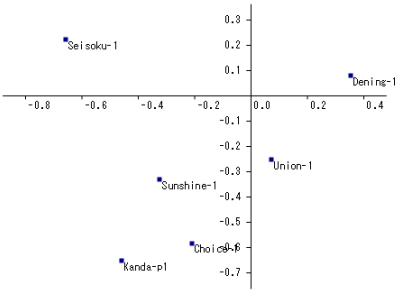


図 3.5　100 語（0.120 / 0.814）

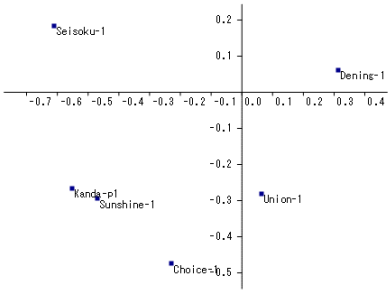
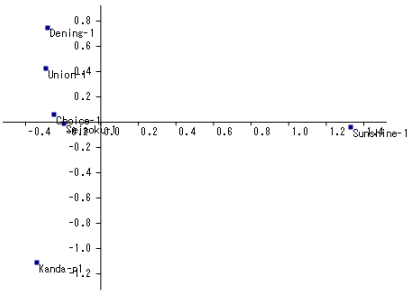


図 3.6　50 語（0.047 / 0.838）

この結果を見ると、まず、Dening-1 の特殊性が目に付く。どの場合でも 1 軸の右端にいる。また、2000 語と 1000 語の間で 2 軸の向きが異なっている。しかし、これは分類という点からは、同じとみなすことができる。また、1000 語と 500 語の間で Sunshine-1 と Kanda-p1 の位置に大きな違いが見られる。500 語以下では多少並びに違いがあるが、ほぼ似たような配置になっている。この実数での比較は、教科書ごとの語数の差が大きく、また単語の選択順についても語数の多い教科書からの影響が大きいことから、あまり良い方法とは言えない。

次に教科書間の単語数の差をなくすために、すべての教科書でまず初めに総語数を 10000 にそろえてから単語の並べ替えを行った結果を図 4.1 から図 4.6 に示す。



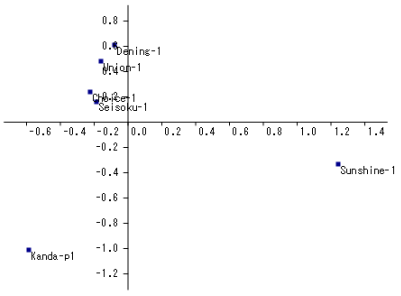


図 4.1　2000 語（0.670 / 0.553）

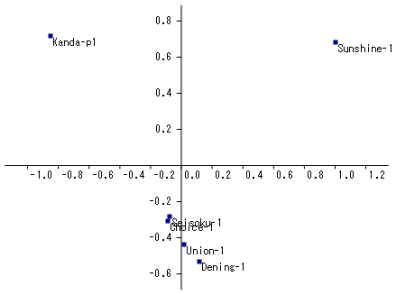


図 4.2　1000 語（0.549 / 0.585）

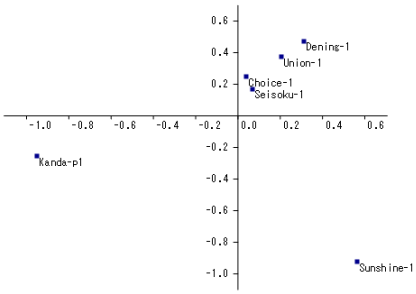


図 4.3　500 語（0.411 / 0.626）

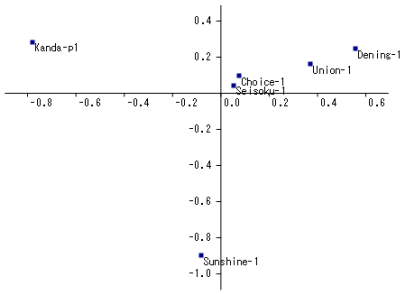


図 4.4　200 語（0.237 / 0.670）

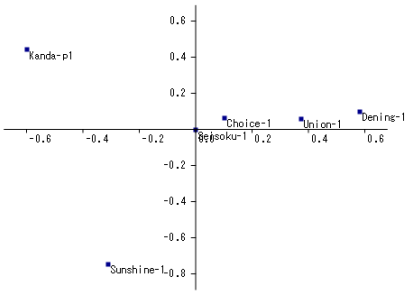


図 4.5　100 語（0.118 / 0.691）

図 4.6　50 語（0.048 / 0.709）

この場合、2000 語以下で、Sunshine-1, Kanda-p1, Dening-1 の位置関係を見ると、500 語で一時的に 2 軸の正負が変わったことを除き、右回転して連続的に変化しているように見える。この程度の変化はどのレベルまでの単語を使うかということで、自然な変化のように思われる。

100 語まで取った場合の最小の調整済み語数と最大の調整済み語数は、それぞれ 4500（Dening）と 5769（Kanda-p）であった。元々総語数が多い教科書は少なく、総語数が少ない教科書は多くなっているが、教科書間に大きな差はない。これらのことから、分析結果の安定には総語数が似ていることが重要であるように考える。

この判断を踏まえ、総語数が似た教科書で、実データを用いて分析を行ってみた。利用した教科書（単語数，総語数）は以下のとおりである。

Choice-1（466, 5629）、Drill-1（505,5603）、Jack&Betty-1（613, 4309）、National-1（426, 4827）、Taisho-1（633,4268）、Tsuda-p1（469,3804）　以上総単語数 1390

結果を図 5.1 から図 5.6 に示す。

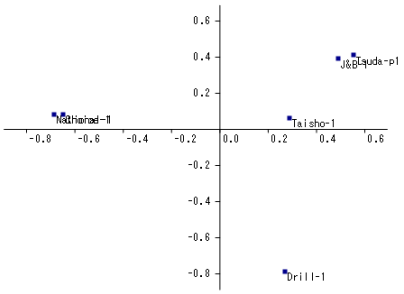


図 5.1　すべて（1390 語）（0.627 / 0.543）

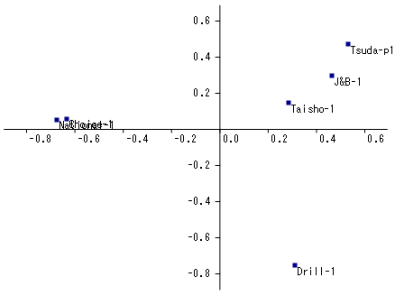


図 5.2　1000 語（0.546 / 0.571）

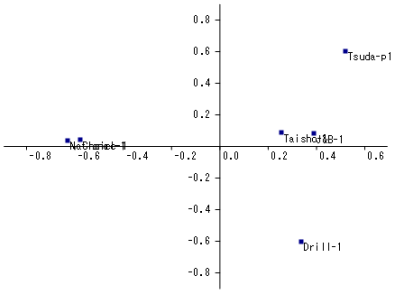


図 5.3　500 語（0.353 / 0.634）

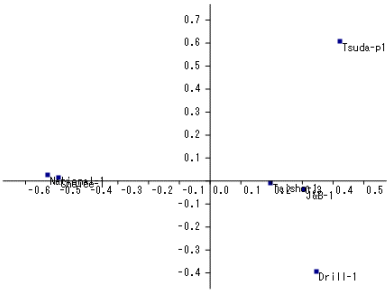


図 5.4　200 語（0.133 / 0.697）

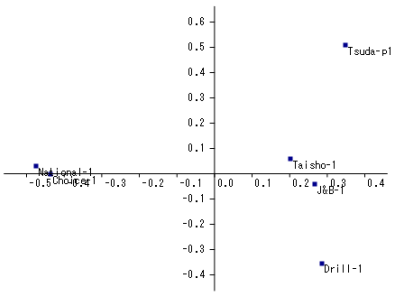


図 5.5　100 語（0.053 / 0.723）

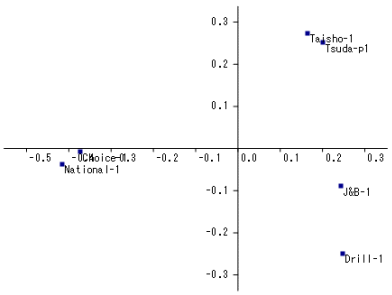


図 5.6　50 語（0.013 / 0.743）

これらの結果を見ると、全体的に安定しており、特に 500 語から 100 語程度まではほとんど変化がみられない。このようなことから、各教科書の総語数がそろっていることが安定的な結果を与える条件のように思われる。

3.　語数調整の提案

前章までで、コレスポンデンス分析による教科書分類の基本的な考え方は、語数をそろえることであることが分かった。これを言い換えると、分析は単純に単語の構成比率だけを用いるべきで、総語数などに差をつけて複雑にしてはならないということである。総語数は教科書の難易度にもつながることであるので、取り入れて分析を行いたいと考えるで

あろうが、コレスポンデンス分析はあくまで単語の構成比率によって、単語がどのような使われ方をしているのかに着目すべきである。

以上のことを考えると、標準化の方法が分かってくる。まず、単語選択に総語数による差が出ないように、全体で一度標準化を行う。次に、単語ごとに全教科書の語数の合計を求め、多い順に並べ替える。その後、語数の合計順に個数を定めて単語を取り出す。その際、どのような分析を行うか、例えば基本的な単語の使われ方を調べるか、より特徴を強く出すようにするかなど、単語の内容と 0 比率を見ながら取り出す単語数を決定する。取り出した単語数によって、総語数は少しではあるが、教科書によって異なる。前章で述べたように、元々語数の多い教科書は、調整され選ばれた単語での総語数は少なくなる傾向がある。これを再度、教科書ごとに総語数が等しくなるように調整し、コレスポンデンス分析にかけるデータとする。この方法を今後二重調整法と呼ぶことにする。二重調整法によるデータは単語の使用比率という意味でデータの特色が明確になる。

この処理を簡易的に行うには、最初の段階の標準化だけを行ってもよい。これまで著者らは直感的にこの方法を使ってきたが、これはそれほど誤った方法ではない。最終的な語数の取り出しは、教科書の場合、基本の単語と少し応用的な単語について特徴を見たいのであれば、現在の中学 1 年生の必要単語 300 語程度から考えて、0 比率は 0.2 以下が良いように思う。これについてはさらに検討する。

最初に与えた、総語数に差がある教科書の組に二重調整法を用いてみよう。結果を図 6.1 から図 6.6 に示すが今回は詳細を見るために、3 次元まで含めて議論する。図の左側が 1 次元と 2 次元の散布図、右側が 1 次元と 3 次元の散布図である。

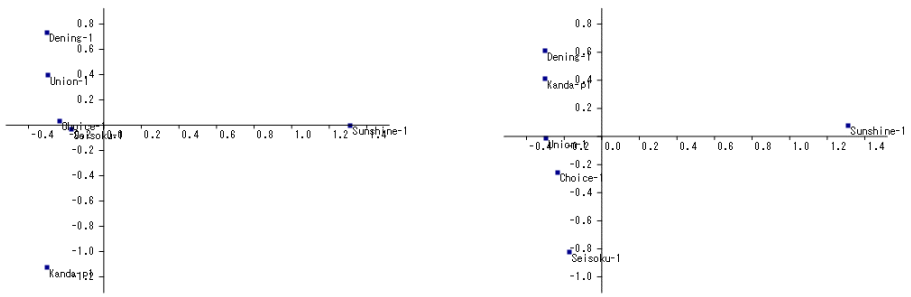


図 6.1　2000 語（0.670 / 0.548）

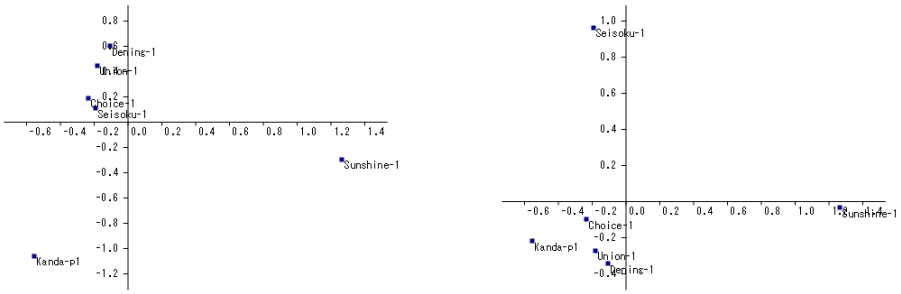


図 6.2　1000 語（0.549 / 0.583）

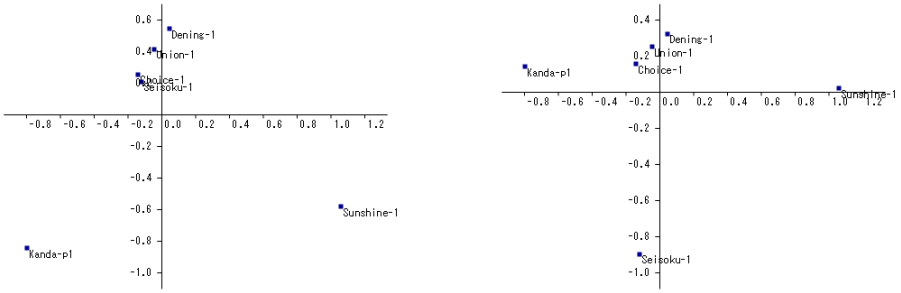


図 6.3　500 語（0.411 / 0.621）

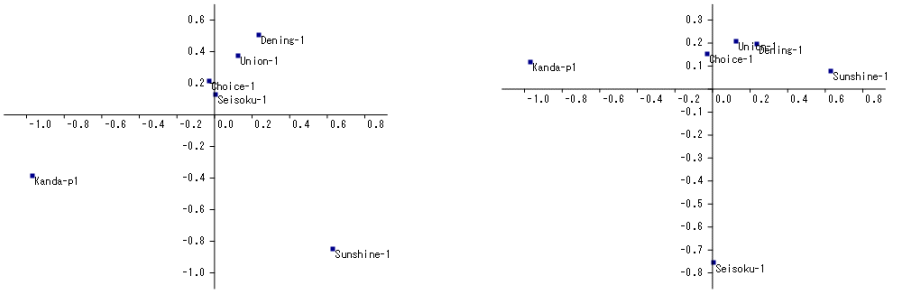


図 6.4　200 語（0.237 / 0.666）

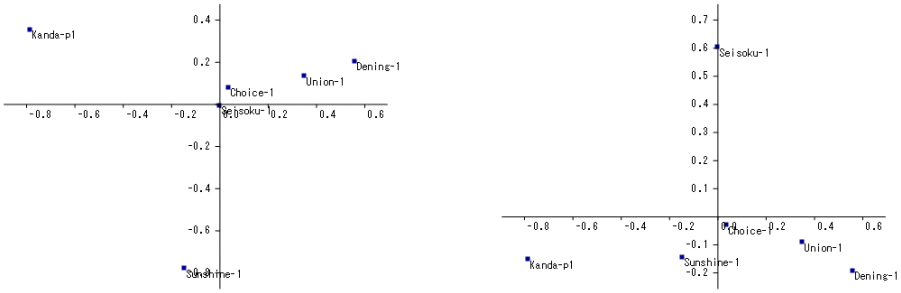


図 6.5　100 語（0.118 / 0.677）

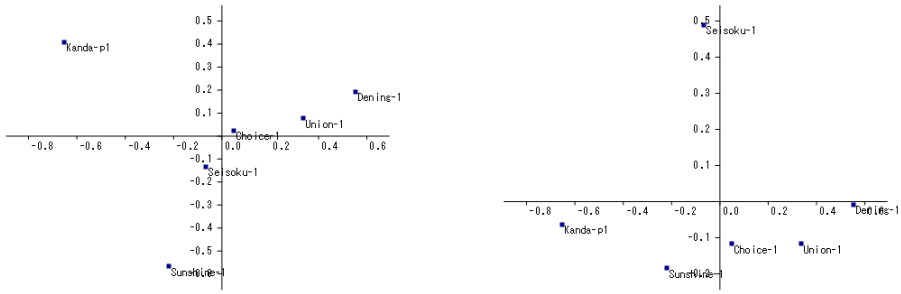


図 6.6　50 語（0.048 / 0.696）

各図の左側を見ると、選択する単語数が少なくなるにつれて、点はあまり位置関係を変えずに連続的に時計回りに回っているように見える。また 3 次元目は、符号が変わることはあるが、Seisoku-1 と他の教科書に分かれる様子は同じである。以上のように結果はかなり安定している。

4.　新しい語数調整法の様々な状況への適用

著者らは教科書の総語数の違いを二重調整法と名付けた方法で調整し、語数の選択については、0 比率を単語の数を決める目安とすることを提案した。しかし、これはデータが中学 1 年の教科書という限定された議論であり、もう少し状況の異なる場合に適用可能かどうか調べてみる必要がある。そのため、この章では懸念される典型的な 3 つの場合に、著者らの手法を適用し、結果を調べてみよう。 1 つは比較する教科書の冊数が多い場合である。もう 1 つは学年が上の教科書の場合である。最後は、単語は難しいが総語数の少ない、入学試験長文読解の問題の場合である。

4.1　比較する冊数が多い場合の検討

これまでは図を見やすくするために、6 冊の教科書を使って分析してきたが、ここではこれを増やし、18 冊を使って議論する。これにより、同じ次元までだと一般的に累積寄与率は低下する。利用する教科書は以下のとおりである。

Choice-1（466, 5629）, Dening-1（3844, 38435）、Kanda-p1（200， 782 ）、 Seisoku-1 （ 736, 13915）、Sunshine-1（340, 1291）、Union-1（935, 5476）、Drill-1（505, 5603）、Globe-1（387, 1495）、Inoue-p1（163, 2779）、Jack&Betty-1（613, 4309）、Kanda-1（405, 3858）、Mombusho-p1（253, 1609）、National-1（426, 4827）、Pacific-1（484, 3321）、Standard(p)-1（816, 8250）、Standard(t)-1（826, 4945）、Taisho-1（633, 4268）、Tsuda-p1（469, 3804）

二重調整法を用いた分析結果を図 7.1 から図 7.6 に示す。

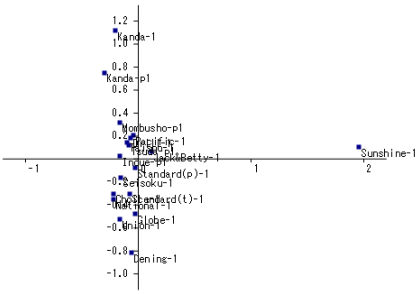


図 7.1　2000 語（0.747 / 0.239）

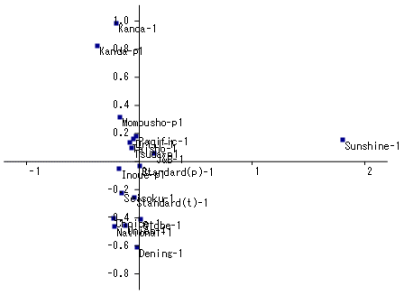


図 7.2　1000 語（0.602 / 0.264）

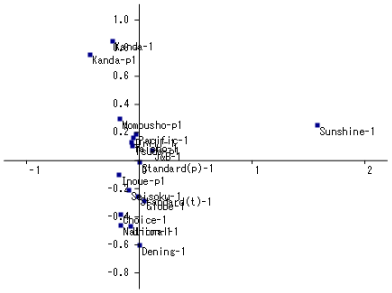


図 7.3　500 語（0.426 / 0.296）

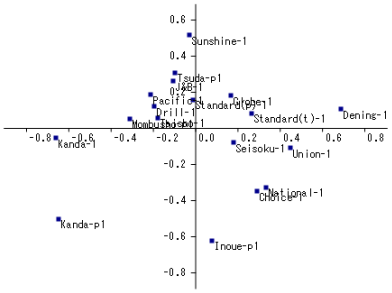


図 7.4　200 語（0.209 / 0.321）

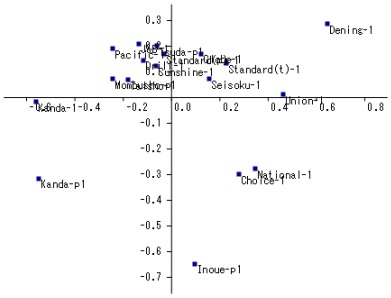


図 7.5　100 語（0.078 / 0.421）

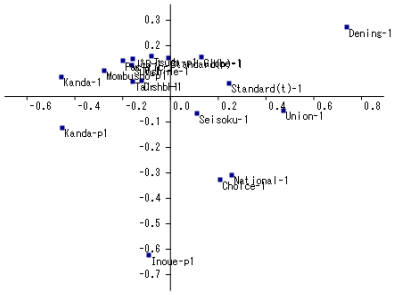


図 7.6　50 語（0.026 / 0.459）

これを見ると、500 語と 200 語の間で結果が大きく異なっている。200 語は 0 比率が 0.207 であり、著者らの提案した 0.2 に近く、教科書数が多い場合でも著者らの 0 比率が 0.2 以下は比較的安定しているという主張はほぼ成り立っている。

4.2　教科書の難易度が全体的に高い場合の検討

これまでは中学校 1 年の教科書を用いて分析を行ってきたが、ここでは少しレベルを上げて、中学校 3 年の教科書を用いてみる。結果を図 8.1 から図 8.4 に示す。

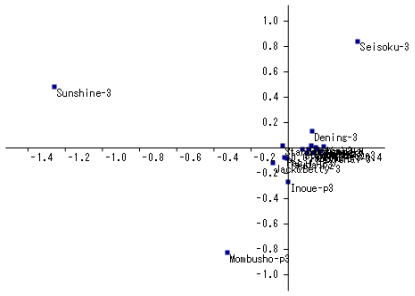


図 8.1　2000 語（0.411 / 0.272）

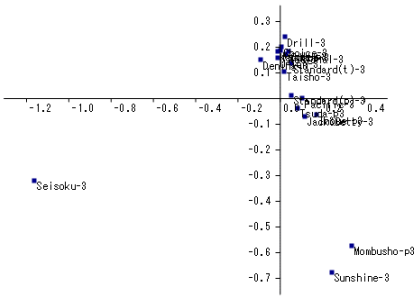


図 8.2　1000 語（0.223 / 0.320）

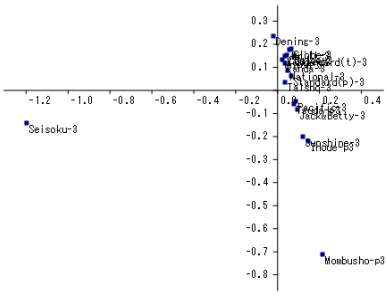


図 8.3　500 語（0.089 / 0.405）

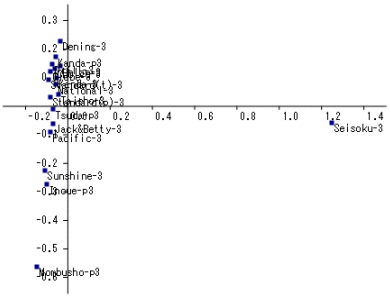


図 8.4　200 語（0.023 / 0.529）

これを見ると、難易度の低い教科書に比べて、同じ語数を選択した場合、低い 0 比率を与えている。また、散布図は 2000 語と 1000 語の間で大きく変化しており、後は軸の正負を除いて連続的な変化が示されている。このことから、難易度が高い教科書の場合でも 0 比率は 0.2 以下で分析を行う方が良さそうである。また、中学 3 年生の習得単語数はほぼ 1000 語と言われており、それも 0 比率 0.2 以下の条件にほぼ合っている。

4.3　総語数が小さい場合の検討

著者らが提案した方法を入試問題に適用してみよう。取り扱う入試問題は、以下の大学の主要読解問題である。

センター入試（221,495）、九州大学（328,723）、広島大学（316,559）、大阪大学（352,707）、京都大学（262,523）、東京大学（178,347）、東北大学（446,982）、北海道大学（305,643）

分析結果を図 9.1 から図 9.6 に示すが、今回は 3 次元目に違いが見られたので、散布図は 3 次元まで示す。

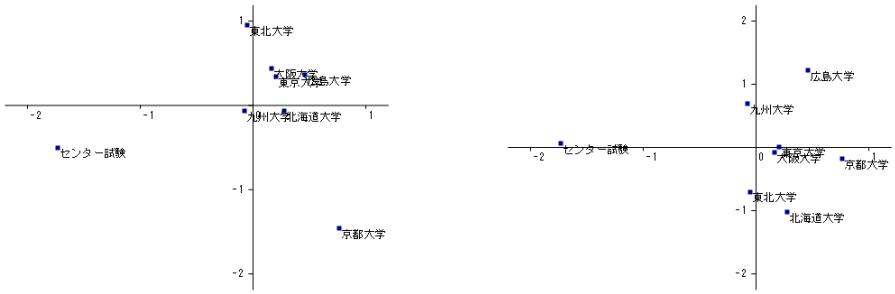


図 9.1　すべて 1584 語（0.810 / 0.312）

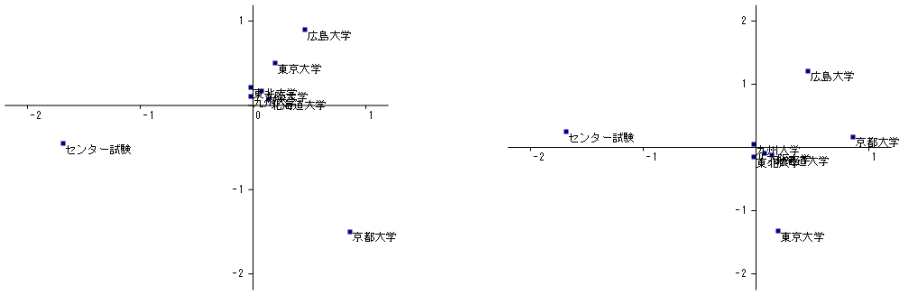


図 9.2　1000 語（0.772 / 0.356）

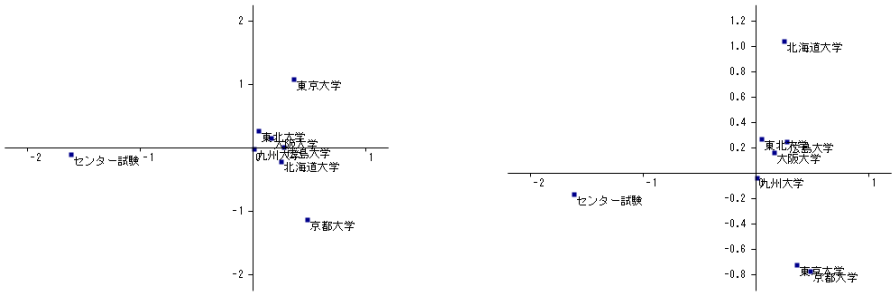


図 9.3　500 語（0.677 / 0.363）

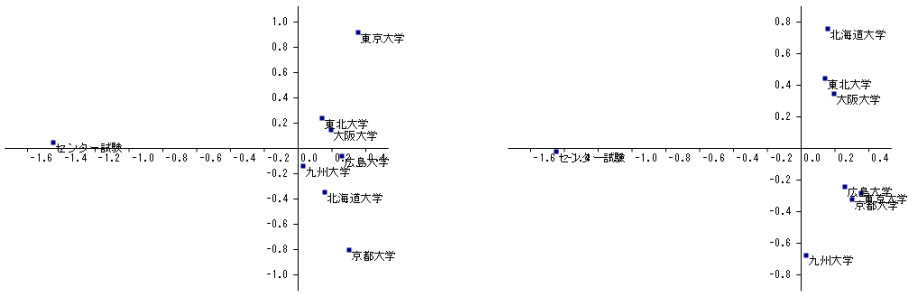


図 9.4　200 語（0.513 / 0.382）

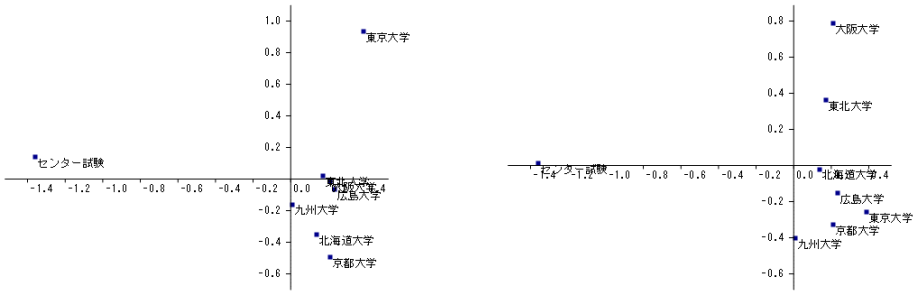


図 9.5　100 語（0.375 / 0.445）

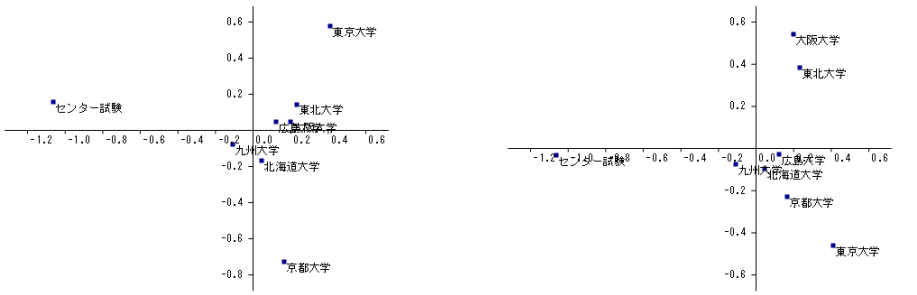


図 9.6　50 語（0.223 / 0.535）

この結果によると、2 次元までだと 500 語以下は同じような傾向を示しているが、3 次元は 200 語と 100 語の間を境に傾向が少し異なっている。安定的な結果が得られるのは、100 語以下のように思われる。文章の量が少ない場合、単語はその年の出題傾向や偶然性に左右される。そのため 0 比率は高めに出ることになり、100 語でも 0.3 以上になる。このことから、網羅的な教科書とは異なり、文章の量の少ない入試問題においては多少 0 比率を高く選んでも良いのかもしれない。但し、それでも語数が少ないことから、比較は基本的な単語の使い方を比べることになりそうである。教科書と同じレベルの比較を行うためには、もう少し年度を重ねて分量を増やす必要があるように思う。

5.　おわりに

著者らはこの論文で、今まで専門家の経験を頼りに行ってきたコレスポンデンス分析によるテキスト分類の方法を見直し、経験によらず、ある程度の知識で一般的に行える方法を提案した。それは以下の手順に基づく。

1. 教科書ごとに集めた単語の総語数を一定にし、各単語の語数はその値からの比例配分で求める。

2. 教科書を集めて単語ごとの語数を合計し、語数合計の多い単語から順番に並べ直す。
3. 単語の語数順に利用する単語数を決めるが、それには 0 比率という指標を用いて、これが 0.2 以下の範囲で、単語の難易度を考えながら決める。
4. 単語数を決めたら、もう一度単語の総語数を一定にするように標準化し、このデータをコレスポンデンス分析に利用する。

この一連の手続きを二重調整法と名付ける。

この方法では、単語の語数を用いず、語数の比を用いることになり、文章の分類が、単語の使用法という意味で明確になる。また、利用する単語の個数も、0 比率という指標を用いることにより、あまり迷うことなく決定できる。

0 比率を定めて分析を実行すると、総語数が多い文章の集まりで選択語数が多くなり、総語数が少ない文章の集まりで選択語数が少なくなる傾向がある。そのため、入試問題などの総語数の少ない文章だと、同じ 0 比率で単語数が少ない分析しかできなくなる。単語数が少ない分析は、簡単な単語の使い方による分析と解釈できるので、もう少し本質的な違いをみるためには、より多くの年度の試験問題を集める必要がある。確かにこれは、正確な分析には多くの文章が必要になるという、一般的な直感とも一致している。

本文中でも述べたが、コレスポンデンス分析に文章の難易度を求めるのは困難である。文章の難易度は別途求め、コレスポンデンス分析の結果と合わせて教科書の分類を考えるべきであろう。

参考文献

［1］ 村上征勝，今西祐一郎，源氏物語の助動詞の計量分析，情報処理学会論文誌，Vol.40, No.3, 774-782, (1999)

［2］ 田畑智司，コーパス言語学のための多変量解析入門，英語コーパス学会第 24 回大会ワークショップ，於日本大学文理学部, (2 Oct. 2004)

［3］ 水本篤，コーパス言語学研究における多変量解析手法の比較 －主成分分析 vs. コレスポンデンス分析－，統計数理研究所共同研究リポート 232,『コーパス言語研究における量的データ処理のための統計手法の外観』, 53-64, (2009)

［4］ 渡辺清美，浅井智雄，赤瀬正樹，「中国小学校英語教科書の語彙の量的分析-日本の現行教科書との比較を中心にして-」，日本言語教育 ICT 学会研究紀要，vol. 4, 47-58, (2017)

［5］ 渡辺清美，浅井智雄，小篠敏明，A Correspondence Analysis of Five Japanese Historical English-as-a-Foreign-Language Textbooks, International Conference on Education, Psychology and Learning, Conference Proceedings 61-73, (2017.8)

［6］ 渡辺清美，福井正康, Quantitative Analysis of Initial Stage English Textbooks in Asia in

Comparison with Textbooks in Japan, 2018 International Symposium on Teaching, Education, and Learning - Winter Session, Conference Proceedings 117-130, (2018.1)

Method of Word Count Adjustment and Word Selection Criteria for English Text Classification Using Correspondence Analysis

Masayasu FUKUI [*1] and Kiyomi Watanabe [*1]

**1 Department of Business Administration, Faculty of Business Administration, Fukuyama Heisei University*

Abstract: In the analysis of text classification by correspondence analysis, there have been very few effective methods for its word selection procedure, and thus the analyses were often carried out based on researchers' mere experiences. In an attempt to find such methods, we conducted an experimental analysis using example sentences to formulate effective methods for word counting adjustment and word selection criteria, which has produced stable results. As a result, we proposed two new methods: double adjustment method and word selection based on 0 ratio.

Key Words: English text classification, correspondence analysis, word number adjustment