『語学教育研究論叢』第 35 号

コーパスから抽出した用例に含まれるノイズへの対応

上村　圭介・髙野　愛子

Evaluation of the influence of retrieval errors in corpus-based studies of Japanese

Keisuke KAMIMURA & Aiko TAKANO

要旨

本稿では現代日本語書き言葉均衡コーパス（BCCWJ）から抽出した順接接続詞の用例に対する全件チェックの結果をもとに、抽出ノイズが分析結果に及ぼす影響を検討した。対象接続詞別に抽出結果の適合率と再現率を明らかにし、接続詞によっては形態素情報を利用した抽出条件では見逃してしまう用例が多数含まれることが分かった。さらに、接続詞別の出現頻度は抽出ノイズの前後で同等性が棄却されること、および対応分析の次元得点を利用したクラスター解析の結果に異同が生じることを確認した。本稿の分析から、コーパスから得られる用例については、その適否についての精査が必要であることが改めて示されたほか、大規模であるとしてもコーパスは有限であり、全数チェックによるデータ精査の結果をコーパスの改善に結び付けることが必要であると結論付けた。

Abstract

This article examines the effect of retrieval errors on statistical analyses, based on the result of data cleaning against the occurrences of conjunctive words and phrases retrieved from a large linguistic corpus of Japanese. The authors first conduct a comparison of the precision and recall rate for each conjunctive, to find out that morphology-based retrieval expressions do not effectively return relevant occurrences for some conjunctives, thus lowering the recall. Chi-squared tests show that the frequencies of conjunctives are not equivalent before and after data cleaning. Although, according to correspondence analysis, the relative positions of the conjunctives do not significantly change regardless of data cleaning, it may affect the cluster formation of conjunctives. The authors conclude that data cleaning against retrieval results may need closer attention than conventionally assumed, and that cleaning results may better be fed back for the improvement of the corpus.

1. はじめに

大規模な日本語コーパスが整備されたことで、コーパスに収録された用例をもとにした研究が広く行われるようになっている。しかし、現在一般的に利用可能なコーパスでは、文脈上の意味や機能を正確に利用した用例の抽出はできない。その結果、コーパスから得られた用例の中には、利用者が意図する意味・機能に合致しない用例が少なからず含まれている。「ノイズ」とも呼ぶべきこれらの用例に対しては、集計・分析の対象外にするなど適切な処理が行われなければ、分析結果や考察を損なうものとなるおそれがある。一方で、大規模なコーパスであればあるほど、抽出した用例を精査し、そのようなノイズを除去することは現実問題として難しく、一般的には小規模サンプルを取り出し、コーパス全体における傾向を推定することになる。

髙野・上村（2017、以下「元研究」と呼ぶ）では、現代日本語書き言葉均衡コーパス（BCCWJ）の形態素情報を利用して順接の接続詞の用例を抽出し、類似の意味・機能をもつ順接接続詞の文体的な使い分けを、外延的な文体的特徴を通じて説明するのではなく、実際にこれらの接続詞が用いられる

文章のレジスターやジャンルと結びつけることで説明している。その際、抽出結果における用例のすべてについて順接接続詞としての適否を人手によって判断している。

本稿では、BCCWJ から抽出した用例について行った全数チェックを基に、ノイズ除去作業が分析結果に及ぼす影響について報告するとともに、日本語研究におけるコーパス整備の今後のあり方について検討する。

2. 研究の方法

2.1. 先行研究

コーパスを利用した研究において、ノイズは大きな問題である。コーパスとは、言語研究での使用を想定した、現実の言語についての大規模で、代表的であって、網羅的なコンピュータ処理可能なデータであるとされる（石川、2012:13）。代表的であり、かつ網羅的であるということは、ある意味でコーパスとは実際の言語運用の縮図であることになる。

だとすれば、その中に話者が現実に生成した誤用や文脈的に不適切な使用が含まれることは許容すべきであろう。しかし、コンピュータ処理可能な形式にするためのデータ化の過程において生じる誤入力や入力漏れなどの「誤り」は、分析結果や考察を損なうおそれがあり、コーパス作成の過程において、あるいは用例抽出後の集計や分析の過程で補正する必要がある。テクスト的な内容に加えて、様ざまな付加的な情報を加えたリッチなコーパスであれば、コーパス中に出現する形態素や文構造などの情報の正確性が問われることになる。このように、現実のコーパスには言語研究の対象とするには適さない様ざまな不適切な情報が含まれる。

例えば、田野村（2012）は、BCCWJ に含まれるサブコーパスの一部である、「Yahoo! ブログ」および「Yahoo! 知恵袋」のデータの中に、現代日本語の書き言葉の基準とするには不適当なデータが含まれていることを指摘している。

また、引用により同一のテクストが同じ本文に繰り返し出現することの影

響を評価した研究（Sigley, 2016）によれば、コーパス全体の 1％程度はそのような繰り返し出現する引用等のテクストであるという。このような繰り返しテクストの影響は、分析対象によって異なり、語彙項目や文法項目では分析結果への影響は少ないと考えられるが、項目間の共起が関わる場合には分析結果が歪められるおそれがあるという。

馬場（2015）は、BCCWJ の長単位で接続詞に分類される語彙素のうち、使用頻度の高いものについてサンプル調査を行い、接続詞別およびレジスター別の解析精度について調べた上で、サンプルにおける誤解析の比率から、BCCWJ 全体の接続詞の出現数の補正を試みている。こうすることで、抽出された用例にノイズが含まれていても、分析においてその影響をできるだけ抑制することができる。

抽出結果におけるノイズへの対応としては、サンプルに基づいた補正をすることが一般的だが、考えられる選択肢としてはほかに抽出結果に対する全件チェックがある。しかし、全件チェックは、作業量が膨大になることや、一貫性をもった基準によるチェックが難しいこと、人手の作業による誤りが排除できないことなどの問題があり、積極的に行われることは多くないと考えられる。

しかし、その一方で、抽出結果についての検証が十分ではないと思われる事例も見られる。例えば、宮内（2013）は BCCWJ に出現する接続詞について「なので」、「だから」、「ですから」の三つの接続詞が「法律」レジスターにおいてもっとも多く出現すると述べているが、筆者らの研究では、「法律」レジスターの中に文と文の論理的な関係を示す順接の接続詞は確認できなかった[1)]。

このような事例があることを考えると、コーパスを使った研究においては、抽出結果の中身について一般に考えられているよりも丁寧に検証する余

[1)]「法律」レジスターに収録されているのは法律の条文から取ったサンプルであるため、ここで対象となる文と文の間の論理的な関係を示すような接続詞は使われていないはずである。

地があると思われる。現実的にはすべての研究において全件チェックを行うことは難しいとしても、抽出結果に含まれるノイズや用例の分布など、コーパスが内在する様ざまな特徴を十分理解し、それが分析にどのような影響を及ぼすものであるかを把握し、場合によっては結果として導かれる解釈や結論を修正する必要があるだろう。

2.2. 検証データ

本稿では順接の接続詞の用例について全数チェックを行った元研究のデータを用いる。このデータは、接続詞の文脈的使い分けの傾向を明らかにするために BCCWJ から抽出された順接の接続詞（接続詞相当句を含む）の用例データである [2)]。

用例の抽出にあたっては、BCCWJ の形態素（長単位）解析上の「接続詞」と、日本語教育上接続詞として導入されることの多い接続詞相当句を対象とした。形態素解析上接続詞であるものだけでなく、他の品詞の組み合わせとして解析された可能性のあるものも抽出するように抽出条件を指定した。

BCCWJ から抽出した用例は 66,202 件に上ったが、その結果、順接の接続詞と認めるべき用例は 51,121 件となった。最初に抽出された用例のうち 22.8％にあたる 15,081 件については、分析の対象とすべきではない「ノイズ」であったことになる。その上で、元研究では、抽出されたすべての用例について順接の接続詞としての適否を判断した。

2.3. 分析手法

本稿では、このようにして得られた接続詞のレジスター別頻度情報について、まず接続詞ごとのノイズ混入の状況について分析する。ノイズ混入の状

[2)] 先行研究や教科書で言及されることの多い順接の接続詞 26 件のうち、BCCWJ での出現頻度の低いものを除いた 15 件を分析対象とした（使用頻度の低い 11 件の接続詞については「その他」として集計した）。また、検索条件の指定にあたっては、文頭に出現するものだけを対象とした。

況を見る上では、検索式によって抽出された結果がどの程度目標とする用例を含んでいると言えるのかを最初に検討する。

その上で、ノイズの除去の有無によって、元研究で行われた分析結果がどのように異なるかを比較する。具体的には、接続詞の出現頻度の同等性、対応分析により得られる次元得点から見た接続詞およびレジスター間の相対的位置関係、そして次元得点を利用した接続詞とレジスターのクラスター形成について、ノイズを除去したデータと、ノイズを除去する前のデータとでどのような変化が見られるかを検討する。

3. 抽出結果の検討

3.1. 長単位における品詞情報の精査

BCCWJ から抽出された用例がどの程度適切であったかを検討する上では、抽出された用例が、第 1 に形態素解析上接続詞であるか、第 2 に文法上または機能上接続詞とみなすべきかという 2 点から検討する必要がある。それぞれの組み合わせから抽出結果は表 1 に示す①から④までのグループに分けられる。

表 1　形態素解析の結果と実際の機能の関係

		機能上	
		接続詞	接続詞以外
解析上	接続詞	①	②
	接続詞以外	③	④

①は、BCCWJ の形態素情報において接続詞とされており、使用実態としても接続詞として扱うべき用例である。

②は、BCCWJ の形態素情報において接続詞とされるが、文法的または機能的な観点から接続詞としてみなすべきではないものである。以下の例に見られる「そこで」は、形態素情報では接続詞となっているが、「その建設会社で」という意味であり、代名詞＋助詞と解析されるべきものである（墨付

き括弧内は BCCWJ におけるサンプル ID）。

⑴私は建設会社に二年間勤めました。そこでエンジニアの夫と知り合ったのです。【PM25_00067】

③は、外形上接続詞の可能性がありながら解析誤りのために接続詞以外の品詞として解析されているもので、機能上も接続詞として用いられているものである。以下の例の「よって」が相当する。

⑵このように米国のモバイルデータサービスは広がっているが、市場開拓に多額の投資を行うリスクがあるため、需要を見極めながら慎重に展開している。よって当分は、パソコンに付随して利用されるニッチなサービスとしての存在が続くだろう。【PM25_00067】

この例における「よって」は、BCCWJ の解析結果によると動詞連用形「ヨッ」＋助動詞「テ」として扱われているが、機能的には接続詞とみなすべきである。このグループの用例は、抽出条件で捕捉できなかった「見逃し」に相当する。

④は、解析上または機能上のいずれにおいても接続詞でないものである。しかし、無関係の文字や形態素の連続ということではない。外形上接続詞に見える（ただし、形態素解析上は接続詞とされない）もののうち、文法上または機能上接続詞として扱うべきではないものが相当する。

BCCWJ の形態素情報を用いて用例を抽出すると①と②が抽出結果に含まれる。このうち、①の用例は確実に接続詞と言えるものだが、②に相当する用例が多数

含まれていれば、抽出結果の信頼性は低くなる。また、抽出結果に②が少なかったとしても、見逃しにあたる③の用例が多ければ、やはり抽出結果の信頼性が高いとはいえない。

一般に、抽出された用例（①＋②）に含まれる①の割合$\left(\frac{①}{①+②}\right)$を適合率（precision）と呼び、本来抽出すべき用例（①＋③）のうち実際に抽出された①の割合$\left(\frac{①}{①+③}\right)$を再現率（recall）という。抽出結果の信頼性を高めるには、両方の値がともに高いことが望まれるが、両者は必ずしも両立しない。そのため、この二つの値の調和平均がF尺度として用いられる。調和平均は、適合率および再現率のそれぞれの逆数の平均（算術平均）の値を求め、さらにその値の逆数をとったものである$\left(\frac{2}{\frac{1}{適合率}+\frac{1}{再現率}}\right)$。この値は、適合率か再現率のいずれかが低い値であると、それに引きずられて低下する特徴をもつことから、母集団からのデータ抽出に関する総合的な指標として広く用いられる。

3.2. 精査の具体例

前節①にあたる分析対象とする用例をBCCWJから抽出する過程において、慎重に検討を要する用例があった。それは、「それで」、「で」である。両者は同義の接続詞であり、その意味・用法には「順接」とは異なる「添加」があるため、「順接」を対象とする今回の分析では「添加」に当たる用例を排除する必要が生じた。

この「順接」か「添加」かを区別する作業は、非常に困難を要するものであった。「それでいいです」の指示詞「それ」＋助詞「で」のような明らかなノイズとは異なり一見して判断できないうえ、用法「順接」「添加」が用例上では類似しており、どちらか迷う例が多かったためである。

「添加」とは、日本語記述文法研究会編（2009：85）の分類・定義によると「先行部で示された情報に、同種の別の情報をさらに付け加えて示す」もので、「そして」「それで」「それと」「あと」などの接続詞が類義となるもの

[3] 『三省堂国語辞典　第七版』（2014）においてのみ、「添加」にあたる用法を「話を続ける」ものとして一番目に挙げている。

である。また、多くの国語辞典では、一番目に「順接」としての用法、二番目にこの「添加」にあたる用法の記述がある[3)]。『明鏡国語辞典　第二版』（2010）の記述では「相手に話を促したり話をさらに続けたりするときに使う。そして。それから。で。」となっており、談話の標識としての役割があることがわかる。

⑶お気持ちは十分にわかりました。それでですね、これはオレの勝手な判断なんですが、今日を最後にオレは須藤さんへの取材はヤメにします。　【LBp3_00062】

上記⑶は「気持ちが十分わかったから取材をやめた」と考えると「順接」になるが、話を続ける談話標識としての機能の役割のほうが強いと考えると「添加」になる。この用例については、話を続けるための「添加」として判断しノイズとして排除したが、そう判断するまでにかなりの時間を要した。順接としての「それで」「で」も、会話内で出現する傾向があり、前後の文脈が辞書の用例のように短くわかりやすいものではない。そのため、先行部と後続部の関係を判断するには各用例の前後の文脈を一つ一つ丁寧に分析しながら確認する必要があった。さらに、用例数が「それで」「で」だけでも合計約 1 万件（9,818 件）と大量にあったため、精査するのにかなりの時間を要した。

以上のように精査した結果、「それで」「で」は「順接」よりも話題の「添加」を示す談話標識としての用例が多くを占めており大量に排除されることとなった。もし、この用法の区別をせずに分析した場合、結果は大きく異なっていたはずである。このようなことも生じうるため、用例の適否については、人の目を通す丁寧な精査が必要であるといえる。

3.3. 全体的な抽出精度

今回抽出した用例の全件を精査した結果を表 2 に示す。抽出結果のうち、

形態素解析上の接続詞という条件で抽出された用例（表 1 の①＋②に相当）は 38,565 件あった。そのうち、機能上も接続詞と認められるものは、30,141 件（同①）であることから適合率は 0.781 となる。一方で、機能上接続詞として認めるべき用例（同①＋③）は 51,123 件であるので、再現率は 0.589 となる。

品詞情報に基づき接続詞として抽出された語句については 8 割近くが機能的にも接続詞であったことになるが、本来接続詞としてみなすべき語句のうち、接続詞として解析され、かつ抽出できたものは 6 割に満たなかったということである。

表 2　調査対象となった接続詞の抽出数

	接続詞として解析			接続詞以外で解析			全体		
	対象数	対象外	適合率	対象数	対象外	適合率	対象数	対象外	適合率
だから	14,598	431	0.971	14	2	0.875	14,612	433	0.971
そこで※	−	−	−	10,747	1,563	0.873	10,747	1,563	0.873
したがって	7,534	0	1.000	71	0	1.000	7,605	0	1.000
ですから	3,684	2	0.999	6	0	1.000	3,690	2	0.999
そのため※	−	−	−	3,296	262	0.926	3,296	262	0.926
それで	1,506	1,671	0.474	1,038	2,474	0.296	2,544	4,145	0.380
その結果※	−	−	−	1,998	267	0.882	1,998	267	0.882
なので	1,440	1	0.999	2	3	0.400	1,442	4	0.997
そのために※	−	−	−	1,264	1,044	0.548	1,264	1,044	0.548
よって	912	2	0.998	121	1	0.992	1,033	3	0.997
それゆえ※	−	−	−	809	0	1.000	809	0	1.000
結果※	−	−	−	413	579	0.109	413	579	0.416
ゆえに	143	0	1.000	256	0	1.000	399	0	1.000
で	324	6,317	0.049	−	−	−	324	6,317	0.049
それゆえに※	−	−	−	259	0	1.000	259	0	1.000
その他※	−	−	−	686	462	0.598	686	462	0.598
合計	30,141	8,424	0.782	20,980	6,657	0.759	51,121	15,081	0.772
接続詞	30,141	8,424	0.782	1,508	2,480	0.378	31,649	10,904	0.743
接続詞相当句	−	−	−	19,472	4,177	0.823	19,472	4,177	0.823

無印は形態素解析上の接続詞、※は接続詞相当句（形態素解析上接続詞以外のもの）

解析上接続詞とされた用例と、解析上接続詞とされなかった用例の適合率は、それぞれ 0.782 と 0.759 であり、特に大きな差があるわけではない。ただし、接続詞以外の品詞に解析された語句の適合率を詳しく見てみると、接続詞としても解析されうる語句の適合率は 0.378 であるのに対して、接続詞以外の品詞にしか解析されない語句の適合率は 0.823 であった。これは、接続詞として解析されている用例の中に、接続詞としての用例がより多く集中しているということであり、形態素解析情報の有効性を示すものと言えよう。

適合率と再現率の総合的指標である F 尺度は 0.672 であり、必ずしも高い数値とは言えない。このことから分かるように、形態素解析上「接続詞」とされた用例だけを検討することは、接続詞の全体的な傾向を把握する上では適当ではないと言える。用例を抽出する上では、抽出ノイズを減らし、適合率を高めることと同時に、単純な抽出条件では見逃してしまう用例を捕捉し、再現率を高めることが求められると言えよう。

3.4. 接続詞別の抽出精度

接続詞別の適合率にはかなりのばらつきが見られることが分かる。「ですから」のように抽出された用例の数が多く適合率が高いものがある一方で、「で」のように用例数が多く適合率が著しく低いものもある。

「で」の適合率が低いのは、話題の転換を示す談話標識としての用例が多く含まれることが要因と考えられる。同じように、適合率の低い「それで」については、接続詞としての用例の中に、順接とは異なる機能をもつものが多く含まれるからである。また、接続詞相当句の中にも、「そのために」、「結果」など、適合率が低いものがある。「そのために」の場合には「そのためには」の一部である「そのために」が抽出されてしまったことが、「結果」の場合には「結果は」のように名詞として機能しているものが多いことが主な要因と考えられる。

前節の通り、機能上接続詞である用例のうち、形態素解析上接続詞とされ

ていない語句によるものは 4 割に達する。形態素解析情報だけに頼ると、用例の 4 割を見失うことになる。

次に個別の接続詞についての適合率と再現率を見る。この二つの値を比較することができるのは、形態素解析上、接続詞と、接続詞以外の品詞の 2 通りに解析されている語句に限られる。これに該当するのは、「だから」、「ですから」、「なので」、「よって」、「ゆえに」、「したがって」、「それで」の 7 組である。表 3 は、これら 7 組について、適合率、再現率、F 尺度を示したものである。

適合率と再現率に注目すると、これらの 7 組は三つに分けられる。一つめは、適合率と再現率が両方とも高いもの、言い換えれば、形態素「接続詞」で用例の大半がカバーされており、その他の形態素に解析された用例もほとんどないというものである。「だから」、「ですから」、「なので」、「したがって」の 4 組が該当する。もう一つは、適合率は高いが再現率が低い（またはやや低い）ものである。形態素「接続詞」によって抽出される用例に不適切なものは少ないが、形態素「接続詞」による抽出でカバーできなかったものが多いことを意味する。「ゆえに」と「よって」が該当する。三つめが、適合率と再現率のいずれも低いものである。このグループには「それで」が該当する。

表 3　接続詞と接続詞以外の形態素の用例数

	用例数	適合率	再現率	F 尺度
だから	14,612	0.971	0.999	0.985
ですから	3,690	0.999	0.998	0.998
なので	1,442	0.999	0.998	0.999
よって	1,033	0.997	0.882	0.936
ゆえに	399	1.000	0.358	0.527
したがって	7,605	1.000	0.990	0.995
それで	2,544	0.474	0.592	0.526

このように、接続詞によって抽出精度には大きなばらつきがある。

BCCWJ の形態素情報に基づいて特定の形式を取り出す場合には、見逃しのないよう配慮するとともに、抽出された用例が意図したものであるか慎重なチェックが必要となろう。

なお、BCCWJ ではデータ全体の 1％については「コア」データとして形態素解析の結果を人手により修正し、「ノンコア」データよりも高い解析精度をもつ。しかし、元研究のデータでは、ノンコアデータとコアデータの間ではノイズの比率に有意な差は確認できなかった（$\chi^2(1)=0.30871$、$p=.5785$）。これは、本稿では形態素上接続詞であるかということに加えて、機能上の順接接続詞かどうかを判断したことによるものと思われる。言い換えれば、品詞上の接続詞を抽出するだけでよいなら、形態素解析情報は、用例を効率的に抽出するために有効であることを示すということであろう。

4. 分析結果への影響

元研究では、頻度情報をもとに、順接の接続詞とそれが出現するレジスターの結びつきの独立性をカイ二乗検定によって検討し、その上で対応分析を実施し、接続詞間、レジスター間の遠近関係を示した。さらに、対応分析の結果として得られる次元得点によってクラスター解析を行い、レジスターおよび接続詞それぞれのクラスター形成の特徴を分析した。

本稿では、抽出ノイズがこれらの分析結果にどのような影響を及ぼすか検討する。ノイズを除去した後の分析結果が、ノイズを除去する前の分析結果と相当程度異なっていれば、ノイズを除去することの妥当性を示すことができる。一方、ノイズの除去が、分析結果に大きな違いをもたらさないなら、コーパスから得た用例からノイズを除去することには、少なくとも元研究で行った全件に対するデータのクリーリングには、積極的な意味はないということになる。

4.1. 単純集計への影響

ノイズ除去の有無は接続詞別集計表の同等性を棄却するものとなった

（$\chi^2(15)=6556.2$、$p<.001$）。当然予想されることであるが、ノイズを取り除いたことで接続詞別の集計表は質的に異なるものとなった。同様に、レジスター別の接続詞頻度についても有意な差が確認された（$\chi^2(10)=701.23$、$p<.001$）。

接続詞別、レジスター別の両方とも、用例からノイズを取り除くことは接続詞の頻度情報に有意な違いをもたらすものであることが分かる。抽出されたままの用例とノイズ除去後の用例を同じデータとみなすことは適当ではない。

4.2. 対応分析への影響

次に、ノイズ除去が元研究で行った接続詞とレジスターの関連に関する対応分析にどのような影響を及ぼすものかを検討する。ここでは、次元得点の変化から接続詞とレジスターの相対的な位置関係がどの程度変化したかを検討する。

図 1（p.256）は、ノイズ抽出をした場合と、しない場合の対応分析の結果（第 1 軸～第 3 軸の次元得点）である。接続詞の次元得点については、第 1 軸で「結果」と「で」が入れ替わったほか、「それで」が「ですから」、「ゆえに」が「そこで」が変動幅は小さいものの入れ替わっている。第 2 軸では、「なので」が原点をまたいで変動したのが特徴的である。第 3 軸では「なので」、「結果」、「それで」において他の接続詞との入れ替わりが見られる。第 1 軸については、ノイズ除去前のほうが次元得点の散らばりが小さく、接続詞の特徴を弁別する力が弱くなっていることがうかがわれる。

次元得点の変化からは接続詞、レジスターのいずれについても相対的位置関係に若干の異同が見られるものの、これらの異同は、ノイズ除去の有無による対応得点の相関関係を否定する水準ではない[4)]。つまり、ノイズ除去の有無により対応分析の結果は大きくは変わらなかったことになる。

[4)] Wilcoxon の順位相関係数に基づく検定による。

4.3. クラスター解析への影響

最後にクラスター解析への影響について検討する。表 4 および表 5 は、ノイズ除去の有無でクラスター解析の結果がどのように変化したかを示す。元研究におけるクラスター解析で認められた四つのクラスターのうち二つについてはクラスター内部の結びつきを含めて変化が見られなかった。残りのクラスターについては、単独でクラスターを形成していた国会会議録が Yahoo！知恵袋とクラスターを形成し、その代わりに Yahoo！ブログが単独のクラスターを形成することになった。

接続詞についてもノイズの除去の有無によるクラスター解析の変化が確認できる。元研究で四つ認められたクラスターのうち、「なので」、「で」、「結果」から形成されるクラスター以外の三つに異同が見られる。さらに、第 4 クラスターについてはクラスター内部の構造への変化も確認できる。

クラスター解析においては、次元得点によって定義される接続詞やレジスターの距離によってクラスターが形成されるため、ノイズ除去の前後で次元

表 4　ノイズ除去の有無によるクラスターの変化（レジスター別）

ノイズ除去後	ノイズ除去前
教科書、白書、広報誌	（変化なし）
知恵袋、ブログ	国会会議録、知恵袋
国会会議録	ブログ
書籍、新聞、ベストセラー、雑誌、図書館	（変化なし）

表 5　ノイズ除去の有無によるクラスターの変化（接続詞別）

ノイズ除去後	ノイズ除去前
なので、で、結果	（変化なし）
ですから、よって	ですから、その他、よって
その他、ゆえに、そのため、その結果	したがって、ゆえに、そのため、その結果
したがって、そこで、それゆえ、それゆえに、それで、そのために、だから	そのために、それゆえ、それゆえに、だから、そこで、それで

得点に有意な違いがなかったとしても、形成されるクラスターが異なることは考えられる。

5. 考察

4. の分析から、①頻度情報全体はノイズの除去によって有意に異なるものとなったこと、②接続詞とレジスターの相対的な位置関係には有意な変化が見られなかったこと、および③クラスター解析の結果には移動が生じたことが確認された。このことは、コーパスを利用した日本語研究にどのような意味をもつのだろうか。

そもそも、コーパスの使い方は二つに大きく分けられる。一つは、コーパスから得られた用例の頻度情報を利用して、計量的に日本語の語彙的、文法的な特徴を明らかにしようとするものである。もう一つは、日本語の語彙的、文法的な特徴に関して立てられる主張を裏付ける実例をコーパスの中から示すものである。

同じくコーパスを使った日本語研究でありながら、前者はコーパスを量的な実証の材料とし、後者はコーパスを質的な実証の材料とするものということができる。後者のようなコーパスの使い方では、ノイズ除去作業の有無が分析の大勢に影響するようなことはないだろう。前者のタイプの研究ではノイズの除去の有無が頻度情報に大きな影響を及ぼすことから、抽出された用例の適否の検討が重要である。

もちろん、抽出した用例のすべてについて分析対象としての適否を判断する作業が常に可能であるとは限らない。それでも、BCCWJ は 1 億語レベルの大規模コーパスであるといっても有限である。前川（2013）は、BCCWJ における異なり語数は 18 万語にすぎず、そのうち約半数の用例は 10 未満であり、100 以上の用例をもつ語は 3 万語にとどまると述べる。実際に BCCWJ に出現する語句（語彙素レベル）[5] を見ても、1,000 件以上の用例が

[5]『現代日本語書き言葉均衡コーパス』語彙表
<http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html>

ある語彙素の数は、短単位で 6,653、長単位で 4,879 に過ぎない。

コーパスはその大規模性がことさらに強調されるが、実際に研究に必要な語句や表現の用例数がそれほど多くないのであれば、文脈情報を考慮した意味・機能の悉皆的な精査をより積極的に行うべきではないだろうか。そういう情報が集約されることで、BCCWJ やその他の大規模コーパスは、言語データの供給源としてより高い信頼性をもつことになるだろう。

6. おわりに

本稿で分析したように、BCCWJ における形態素情報は、用例抽出の上では有効だが、形態素情報だけでは見逃してしまう用例が相当数存在する。また、形態素情報によって抽出した用例には不適切な用例がかなりの割合で含まれている。そのような用例を取り除かないままで分析を行うと誤った結論に結び付いてしまう。コーパスからの用例の頻度に基づいて研究を行う場合には、抽出ノイズを慎重にコントロールすることが必要である。

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

髙野愛子・上村圭介（2017）「レジスター別出現頻度に基づく順接接続詞の文体差の評価　―現代日本語書き言葉均衡コーパス（BCCWJ）の用例分析から―」『語学教育研究論叢』34 号、273 ～ 293 ページ

田野村忠温（2012）「BCCWJ に含まれるウェブデータの特性について ――データ重複の諸相と BCCWJ 使用上の注意点――」『待兼山論叢：文化動態論篇』46 号、59 ～ 83 ページ

日本語記述文法研究会編（2009）『現代日本語文法 7 第 12 部談話』くろしお出版

馬場俊臣（2015）「BCCWJ の品詞情報の解析精度について―特に接続詞に注目して―」『北海道教育大学紀要（人文科学・社会科学編）』第 66 巻第 1 号、13 ～ 29 ページ

前川喜久雄（2013）「第 1 章　コーパスの存在意義」前川喜久雄編『講座日本語コーパス　1. コーパス入門』朝倉書店、1 ～ 31 ページ

宮内佐夜香（2013）「接続詞「なので」の書き言葉における使用について―『現代日本語書き言葉均衡コーパス』を資料として―」『中京国文学』32 号、106 ～ 93 ページ

『明鏡国語辞典　第二版』（2010）大修館書店

『三省堂国語辞典　第七版』（2014）三省堂

Sigley, Robert. (2016). The problem of repeated text in corpus construction. 大東文化大学語学教育研究所研究会発表資料（2016 年 11 月）

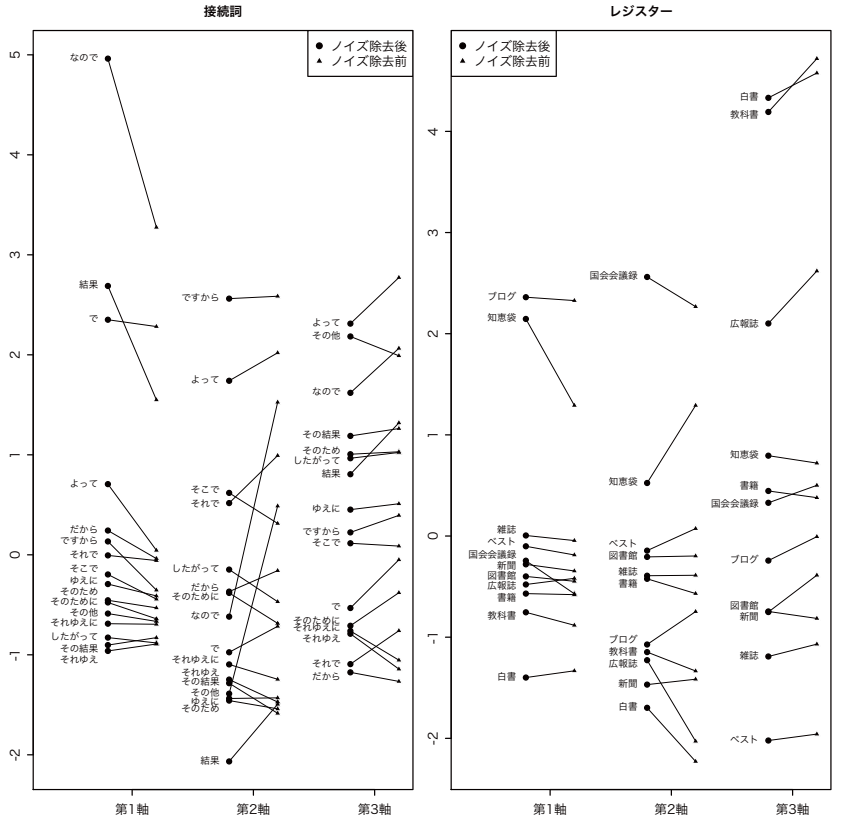

図１　対応分析結果の比較（第１～３軸）