# テキストマイニング技術を応用したレポート課題の教育効果測定

植　田　　麦

## はじめに

稿者は『古事記』『日本書紀』を中心とした、古代日本文学および日本語学の研究者である。同時に、明治大学に所属する教員でもある。所属先では自分の研究対象を授業内容として教示する科目も担当しているが、いわゆる文章表現法の科目、担当授業名では「国語表現」も担当している。受講生数は年度によって差もあるが、おおむね 50 人から 150 人程度である。2020 年度は 90 人であった。

これは文章表現法に相当する授業科目の担当者に限ったことではないが、多くの教員がレポートの採点に頭を悩ませている（椿本弥生・柳沢昌義・赤堀侃司(2008)）。また、文章表現法の科目を担当している場合は学期中の添削もあって、授業期間の負担は相当である。

2000 年代に入ってからは積極的に、機械的なレポート採点についての研究が進められている（椿本弥生・赤堀侃司（2004）、渡邊博之（2008）等）。しかしながら現時点で、稿者を含め多くの教員は、機械的なレポート採点を導入していない。それは機械的レポート採点の精度や教育環境の問題もあるのだろうが、導入のための障壁が高すぎることに原因があるのではないか。

また、文章表現法の科目の担当者としては、最終的に提出されるレポートもさることながら、授業期間内に提出された複数のレポートを確認し、任意の時点・授業内容において、受講生の文章表現能力がどの程度向上したのか（あるいは、しなかったのか）を計測する必要がある。たとえば、レポートの体裁から用語の使い方、一文ごとの質、論の構成などの能力である。そのため、仮に提出されたレポートを機械的に採点しても、教育効果自体の計測に寄与すると

ころは少ない。とすると、機械的なレポート採点を導入する動機は小さくなってしまう。そして、文章表現法の科目担当者は、みずからの経験に基づいた教育を継続し、大量のレポートの添削・採点にいそしむこととなる。

如上の状況に鑑み、稿者はコンピュータを用いたテキストマイニング技術の応用によるレポート課題の教育効果測定を提案したい。と、このように述べると矛盾しているように感じられるかもしれない。文章表現法の科目において、機械的な採点が必ずしも実務に適しているわけではない状況を、稿者みずからが確認したためである。しかし、上に述べたのは、現在研究されている機械採点の方法と文章表現法の科目を担当している教員の状況との乖離である。一般的なスペックのパソコンがあって、無償で入手可能、かつ導入難度の低いアプリケーションソフトがあれば、機械的な技術と教員の業務との懸隔を狭めることができるのではないか、と考えるのである。そして、テキストマイニング技術の導入によって教育効果測定を行うことができれば、より適切な指導が可能となる。

以上のことから、本稿では稿者が 2020 年に担当した「国語表現」の授業をもとに、テキストマイニング技術を用いた文章表現法の教育効果測定の方法とその実践を示す。

## 1．稿者の「国語表現」

本稿執筆時、すなわち 2020 年は、大学教育にとっては大きな変化の一年であった。2019 年末に発生した新型コロナウィルス（COVID-19）は、本稿執筆時点でもその収束を見せることなく蔓延している。国内の大学は春学期（前期）授業期間に対面型授業を行うことができず、ほぼすべての授業がオンラインで行われた。稿者の担当する「国語表現」も同様である。

明治大学では 1 コマの授業あたり 100 分・13.5 回を設定している。総授業時間は 1350 時間であり、90 分・15 回授業と同様の授業時間である。しかしながら、2020 年度は当初の混乱もあり、授業回数としては 12 回が設定され、残りの 1.5 回分は課題等補講によってまかなうことが要請された。稿者が担当する「国語表現」は 4 単位分、つまり 2 コマの授業がセットになっているため、27（13.5 × 2）回の授業が予定されていた。それが学年暦の変更により 24 回となってしまったが、極力、授業内容の変更はしないようにした。以下は、その 24 回の授業の概略である。

第 1 回～第 6 回　：同音異義語やアカデミックワードなど、主として「語」の理解に関わる内容
第 7 回～第 10 回　：文のねじれや多義文など、主として「文」の修正に関わる内容
第 11 回～第 20 回　：資料の読解やレポートの論理構成など、主として「構成」に関わる内容
第 21 回～第 24 回　：学習内容に基づいた総括的内容

このうち、第 4 回・第 12 回・第 14 回・第 16 回・第 18 回に中間レポートを課し、最終回である第 24 回に成績評価対象となる最終レポートを課した。以下に、課したレポートの概略（テーマおよび指定文字数）を示す。

【第 4 回】
レポートテーマ：日本において、一定年齢以上の独身者に対する税（いわゆる「独身税」）を課すべきか、字数：800 字以上 1600 字以内
【第 12 回】
レポートテーマ：2013 年「家計調査報告」に基づいた分析レポート、字数：600 字以上 1200 字以内
【第 14 回】
レポートテーマ：選択的夫婦別姓を認めるべきか、字数：800 字以上 1600 字以内
【第 16 回】
第 14 回で作成したレポートについて、パラグラフライティングに注意した上でリライトせよ、字数：800 字以上 1600 字以内
【第 18 回】
レポートテーマ：地方分権の現状をふまえつつ、その問題点とそれに対するあなたの考えについて述べよ、字数：800 字以上 1600 字以内
【第 24 回（最終レポート）】
レポートテーマ：A）行政と民間の役割分担について述べよ　B）NPO 活動を質・量ともに促進するためにはどうしたらよいか（A もしくは B のいずれかを選択してレポート作成）、字数：2000 字以内

このように、稿者の担当した「国語表現」では、最終試験に相当するものも含めて 6 回のレポートを課した。このうち本稿での分析対象として、調査型レ

ポートである第 12 回に課したものと、修正提出のもととなった第 14 回に課したものを除いた 4 回分のレポートを用いることとする。以下、論考の都合上、第 4 回レポートを「第 A 回」、第 16 回レポートを「第 B 回」、第 18 回レポートを「第 C 回」、最終レポート（第 24 回）を「第 D 回」と称する。

## 2．テキストマイニングツール

上述のとおり、テキストマイニング技術を導入するにあたり、導入にかかる障壁の低いツールの利用を企図した。そのため、MTMineR（https://mjin.doshisha.ac.jp/MTMineR/html/menu.html）および KH Coder（https://khcoder.net/）を利用する。いずれも無料でダウンロード・利用が可能なテキストマイニングツールである。

テキストマイニングとは、その名称のとおり文章（テキスト）を掘り起こして（マイニング）、新たな知見を引き出すものである。現在では特に、大量のテキスト群を機械的に分析する技術を指すことがある。たとえば KH Coder を用いた研究では、議事録やアンケートの分析、Twitter や雑誌記事の分析例がある。いずれも、ひとの目では恣意的な分析になりかねないものを、統計的な処理によって研究の蓋然性を担保するものである。

テキストマイニングでは、分析対象となることばをどのように分けるかが重要となる。たとえば「すももももももものうち（李も桃も桃のうち）」と書かれたとして、これを「すもも／も／もも／も／もも／の／うち」と形態素解析し、かつ「すもも（名詞）」や「も（助詞）」のように適切な品詞として分類することができなければ、実用に耐えない。そのために必要となるのが形態素解析器と辞書である。形態素解析器としては、ChaSen（https://chasen-legacy.osdn.jp/）や MeCab（https://taku910.github.io/mecab/）などが知られている。これらには解析用の辞書がオプションで添付されている。MTmineR では形態素解析器を利用者自らがセットアップする必要がある。一方、KH Coder は ChaSen が内蔵されているため、利用者自身の負担は少ない。

以下、MTmineR と KH Coder の特徴について示す。

### 2.1　MTmineR

MTmineR は、金明哲によって開発された、プログラミング言語である R をベースにしたテキストマイニング用のソフトウェアである（金明哲（2016））。ただし、公開されているのはコンソールとなる MTnineR 本体のみであり、テ

キストマイニングのためのツールなどは利用者が自分でセットアップする必要がある。具体的には、R・JAVA・MeCab・CaboCha などである。ただし、KH Coder などのソフトウェアを併用する場合は、必ずしも MeCab・CaboCha 等の導入は必須ではない。

MTmineR では「Summary（データの要約：分析対象となるテキストの文字数、文の数などの集計）」「Length（文や段落の文字数の集計、延べ語数（Token Number）と異なり語（Token Type Number）での集計が可能）」「KWIC（指定したキーワードについて、その前後の表現を抽出する）」などの機能がある。

これらのうち、Summary と Length は非常に有用である。というのは、文章表現法の科目を担当していれば経験的に「数百字の文章に段落が 1 つしかないレポートは、内容が整理されていないことが多い」「140 字を越える文は、高確率で悪文」などの知見をもっている。そのため、「数百字を越えるが段落が 1 つしかないデータ」「140 字を越える文をもつデータ」のリストがあれば、それらのデータの書き手に対して速やかに指導を行うことができる。

### 2.2　KH Coder

KH Coder には形態素解析器として ChaSen が組み込まれており、公式ウェブサイトで公開されているチュートリアルにのっとってサンプルファイル（夏目漱石『こころ』）の分析を行えば、一通りの使用法が理解できる。また、樋口耕一（2020）では KH Coder を利用した実践研究が、アプリケーションソフト開発者自身によって解説されている。本稿執筆時点ではバージョン 3.x の使用が推奨されているため、以下の解説はバージョン 3 に拠る。

KH Coder は、原則として単独のテキストを分析対象とする。そのため、複数のテキストを対象とするときは、ひとまとめのテキストに加工する必要がある。テキストファイルであれば、KH Coder を利用してまとめることができる。しかしながら、「学科」や「学年」などの属性を外部変数として利用したい場合は、テキストファイルとしてまとめるのではなく、csv ファイルでまとめるとより効果的である。

また、たとえば「独身税」を「独身」「税」のように 2 語として検出するのではなく、1 語の「独身税」としたい場合は、強制抽出語として設定することが可能である。逆に任意の語を分析対象としたくない場合は、除外語として設定することができる。このように、特段の知識がなくても処理が容易であることに KH Coder の特性がある。

分析方法としては、対象テキストの使用語数をみる「抽出語リスト」の作成

や、任意の語の前後にある表現を確認する「KWIC コンコーダンス」に加え、対象テキストの構成をみる「共起ネットワーク」「階層的クラスター分析」「自己組織化マップ」等が利用可能である。また、外部変数が設定されている場合は「対応分析」によって変数ごとの傾向をみることもできる。

文章表現法の教育効果測定にあたっては、「抽出語リスト」が有効である。また、本稿では使用しないが、「共起ネットワーク」「階層的クラスター分析」「対応分析」等も利用価値が高い。「抽出語リスト」では、各品詞ごとに頻用される語の一覧を表形式で出力することができる。また、「共起ネットワーク」等では対象のテキスト（群）の記述内容について、その傾向を分析することが可能である[注1]。

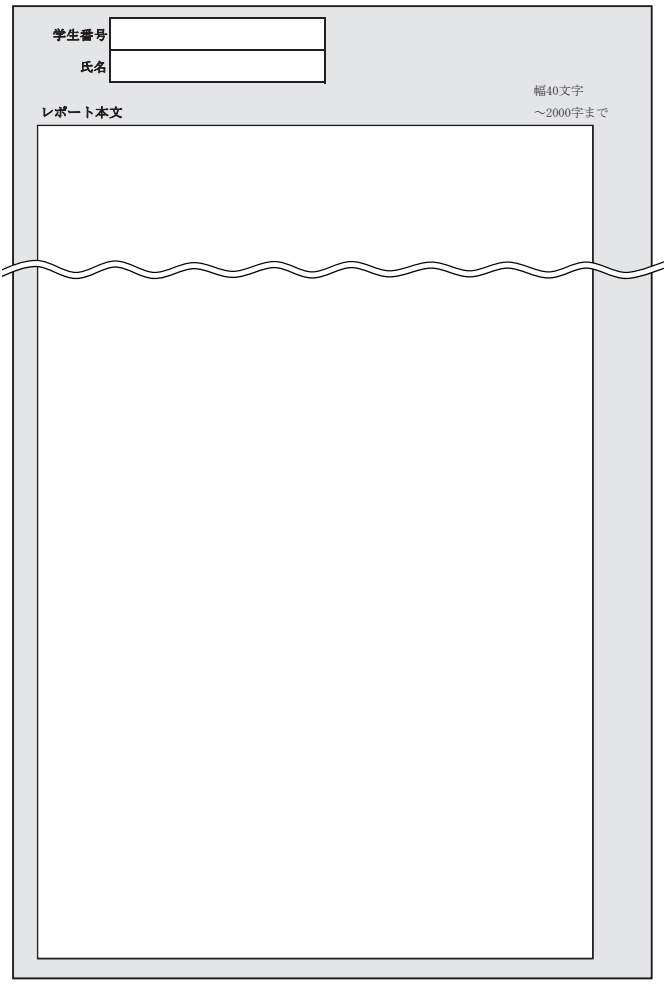
## 3．テキストマイニング技術の導入

「国語表現」にテキストマイニング技術を導入するために必要なのは、分析のための環境と適切なデータであった。環境については、上述のとおり、導入障壁は低い。より重要であるのは、分析のためのデータを適切なかたちで準備することである。

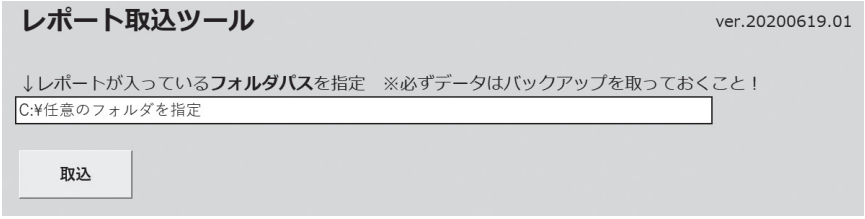
稿者は過去にもテキストマイニング技術を授業分析に用いるための試みを重ねてきた。たとえば、.doc や .txt など、word やテキスト形式でのデータ提出を受講生に課した。しかし、1 つのファイルに「氏名」「学生番号」「レポート本文」などが混在しているため、これを Excel に 1 つずつ編集していくのはきわめて煩雑である。

そのため、次に Excel のファイルをひな形として配布し、A 列に氏名・B 列に学生番号・C 列にレポート本文を記入してもらうことにした。しかし、Excel の使用に戸惑う受講生が多く、データの大半は稿者自身が編集し直す必要があった。このように、受講生から提出されたレポートを分析するといっても、データを適切に収集することは容易でない。

そのため、目的とする分析技術に適したデータを収集するためのフォーマットを作成する必要があった。分析にあたっては、提出者の識別と同定が可能で、かつノイズの少ないテキストデータが必要である。しかも、MTmineR と KH Coder では分析方法が異なるため、それぞれに適したファイルを作成しなければならない。さらに、受講者自身が適切なかたちで課題を提出できる形式にすることが必須である。以上の状況に鑑み、稿者は以下のようなフォーマット（画像 1）を準備した。




画像１　レポートファイル




画像２　レポート取り込みツール

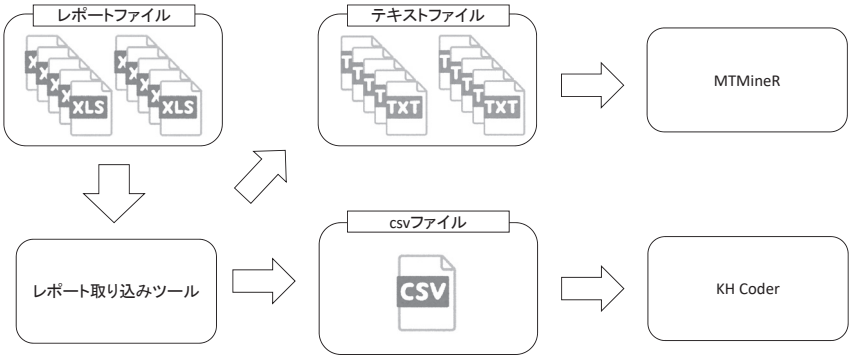
これは、Excel 用のファイル形式である .xlsx で作成したもので、「学生番号」「氏名」「レポート本文」の記入欄を備えている。このファイルを以下「レポートファイル」とよぶことにする。

このレポートファイルを取り込むツール（画像 2）を .xlsm 形式で準備した。これも Excel 用のファイルで、任意のフォルダにあるレポートファイルを一括で取り込み、レポート本文・学生番号・氏名を表にすることができる。さらに、レポート本文 1 文字目が空白である（字下げされている）と 1、空白でない（字下げされていない）と 0 で判定する機能を加えている。また、取り込み時に、学生番号をファイルネームとしたレポート本文のみのファイルを出力し、一つのフォルダに収納するように設定されている。

以下、この取り込み用ファイルを「レポート取り込みツール」とよぶ。

出力されたファイルのうち、レポート取り込みツールで集約したデータはそれのみを抽出して csv 形式にして KH coder の分析用ファイルとして利用し、個別のテキストファイルにまとめたものは MTmineR での分析用ファイルとして利用する。

以上の手順を図（画像 3）として示す。




画像３　手順

## ４.「国語表現」のテキストマイニング

2020 年の「国語表現」は、受講生にはレポートファイルで課題を提出させ、提出された課題を稿者がレポート取り込みツールで集約した。ただし、このツールでもいくつかのエラーが発生した。たとえば、受講生のパソコン環境によっ

ては、作成したレポートのファイルに想定外の変更が加えられ、取り込み時にブランクデータとして取り込まれてしまう事例があった。しかし、それらは限定的で、修正も容易であった。

受講生がレポート執筆の経験を重ねるごとに変化することとして、

1. 語彙が豊富になる
2. 文章の体裁が整う
3. 文章自体の作成能力が向上する

の 3 点を仮説として設定したい。この 3 点を確認するために、

1. 語彙の豊富さの計測と観測
2. 常体の使用・段落冒頭の字下げ・段落数・文の数等の計測と観測
3. 1 文あたりの文字数・高使用頻度語の計測と観測

を行う。

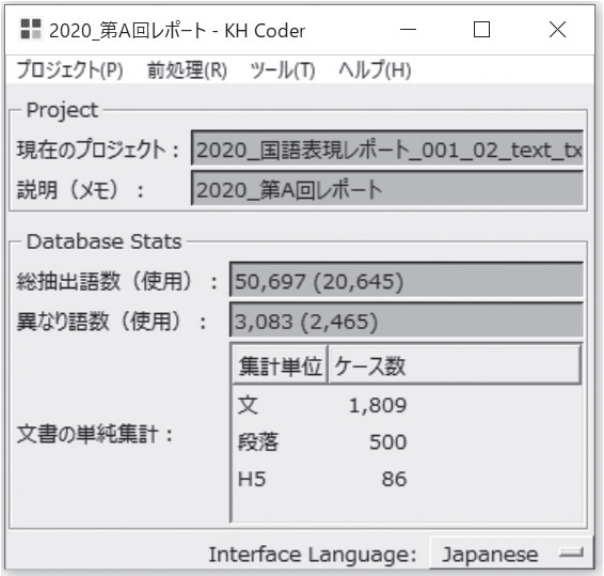
### 4.1 語彙の豊富さ

複数のテキストにおいて、語彙が豊富であるかを示す指標としては、異なり語数（V（N））を延べ語数（N）で割った TTR（Type Token Ratio）が広く知られる。しかし、分析対象となるテキストの分量が増えるほど延べ語数が増えるのに対し、異なり語数は延べ語数と同じ比率では増えない。従って、TTR は分量の異なるテキスト群での比較には向かない。

そのため、語彙の豊富さを示す指標として、TTR に変わるものが多く提起されている。今田水穂（2018）では、TTR を補正する指標を提案している。また鄭弯弯・金明哲（2018）では、11 種類の指標を比較した結果、s について「文章の長さと言語の依存率が最も低い指標」として評価している。本稿では鄭・金（2018）にもとづき、語彙の豊富さを示す指標として s を使用する。なお、参考として TTR の数値も示す。

以下、第 A 回から第 D 回までの異なり語数と延べ語数の一覧を示す。レポートの分析については、KH Coder を使用した。先述のとおり、KH Coder の使用については公式サイトのチュートリアルに詳しいが、ここでは KH Coder における異なり語数・延べ語数の表示について確認しておく。

画像 4 では、総抽出語数（述べ語数）が「50,697（20,645）」、異なり語数が「3,083




画像 4　KH Coder 3 の異なり語数と延べ語数の表示

(2,465)」と表示されている。これは、述べ語数自体は 50,697 語であるが、KH Coder で分析対象とする語が 20,645 語であることを示している。残りの 30,052 語は助詞および助動詞等である。

語彙の豊富さの変化をみるとき、助詞・助動詞等の量がどの程度意味をもちうるのかをみる目的からも、次の表 1 では、助詞・助動詞等を含む数値と含まない数値を示す。

表 1 にみるとおり、第 B 回で TTR・s ともに減少したのち、第 C 回で上昇している。ただし、第 D 回では TTR が「助詞・助動詞等を含む」「助詞・助動詞等を含まない」のいずれにおいても減少しているのに対し、s は横ばいもしくは上昇している[(注 2)]。第 D 回において TTR が減少したのは、延べ語数が増加したのに対して異なり語数が同じ比率では増加していないためである[(注 3)]。

翻って第 A 回から第 D 回までの s を確認すると、確認したように第 B 回で数値が下がったのち、第 C 回では上昇に転じている。

この状況についてまず数値の差が有意であるか否か、そして有意であるとすればその変化が偶発的に発生したのか、あるいは要因があって発生したのかが問題となる。何かしらの要因を想定しうる場合はさらに、受講者側にそれがあ

表１　第Ａ回から第Ｄ回までの語数と指標

| | 異なり語数（助詞・助動詞等を含む） | 異なり語数（助詞・助動詞等を含まない） | 述べ語数（助詞・助動詞等を含む） | 述べ語数（助詞・助動詞等を含まない） | TTR（助詞・助動詞等を含む） | TTR（助詞・助動詞等を含まない） | s（助詞・助動詞等を含む） | s（助詞・助動詞等を含まない） |
|---|---|---|---|---|---|---|---|---|
| 第A回 | 3083 | 2465 | 50697 | 20645 | 0.07 | 0.12 | 0.72 | 0.73 |
| 第B回 | 2868 | 2209 | 55121 | 21999 | 0.06 | 0.11 | 0.71 | 0.72 |
| 第C回 | 3456 | 2827 | 51318 | 21827 | 0.07 | 0.13 | 0.73 | 0.74 |
| 第D回 | 4555 | 3726 | 84845 | 36962 | 0.06 | 0.11 | 0.73 | 0.74 |

るのか、あるいは教授内容･教授者にそれがあるのかを考えなければならない。

しかしながら、このｓの変化については、たとえば別の年度を対照群として設定すればより妥当性のある推測が可能であるが、2020 年度のデータのみでは状況の確認にとどまらざるをえない。この点については、別稿での課題とする。

### 4.2　文章の体裁

文章表現法の科目担当者のみならず、授業でレポートを課す教員に共通する悩みは、提出されたレポートの体裁が整っていないことであろう。たとえば、最初の一文字が下げられずに始まっている段落、改段せずに書かれた数百字の文章の羅列などをみると、それだけでレポートを採点する意欲が減退しかねない。

表２　データ取り込みの例

| 学生番号 | 文字数 | 文章数 | 段落数 | 段下げ | 敬体 |
|---|---|---|---|---|---|
| 1300000001 | 1758 | 41 | 5 | 1 | 1 |
| 1300000002 | 1195 | 24 | 12 | 1 | 1 |
| 1300000003 | 1668 | 36 | 9 | 1 | 1 |
| 1300000004 | 1201 | 18 | 5 | 1 | 0 |
| 1300000005 | 1748 | 28 | 6 | 1 | 1 |
| 1300000006 | 2228 | 46 | 20 | 1 | 1 |
| 1300000007 | 1237 | 25 | 5 | 1 | 1 |
| 1300000008 | 1243 | 25 | 6 | 1 | 1 |
| 1300000009 | 796 | 14 | 6 | 1 | 1 |
| 1300000010 | 954 | 18 | 3 | 1 | 1 |
| 1300000011 | 1886 | 48 | 8 | 0 | 0 |
| 1300000012 | 1412 | 25 | 5 | 1 | 1 |
| 1300000013 | 1551 | 40 | 10 | 1 | 1 |
| 1300000014 | 2198 | 47 | 10 | 1 | 1 |
| 1300000015 | 1706 | 24 | 7 | 0 | 1 |
| 1300000016 | 1999 | 43 | 8 | 1 | 1 |
| 1300000017 | 1900 | 39 | 9 | 1 | 1 |
| 1300000018 | 1493 | 23 | 4 | 1 | 1 |
| 1300000019 | 1671 | 26 | 8 | 1 | 1 |
| 1300000020 | 1284 | 30 | 6 | 1 | 1 |

注意：学生番号および数値はダミーである。

しかしながら、こういった「難点」は文章表現法の科目担当者にとっては指導しやすいことがらともいえる。であれば、機械的に指導のポイントが抽出できれば、ガイドとして機能する。また、このようなポイントを文章表現法の教授内容に加えている場合、複数回のレポートを課すときにそのポイントの推移をみれば、教育効果がどの程度あったのかを計測することが可能となる。

稿者の担当した「国語表現」では、専用のレポートファイルとレポート取り込みツールを導入した。先述のとおりこのツールでは、取り込んだレポート本文の一文字目がスペースであるかどうかをチェックし、スペースであれば「1」、そうでなければ「0」の値を返すように設定されている。また、学生番号をファイルネームとしたレポート本文のテキストファイルを作成する。このテキストファイルを MTmineR で解析すれば、段落数・段落ごとの文字数・一文あたりの長さを求めることができる。

上の表 2 は、MTmineR の Summary 機能を利用して抽出したデータに、Length 機能を利用して抽出した段落のデータを加えたものである。また取り込みツールによって値を返した字下げの数値を加えた。さらに、取り込みツールで抽出したレポートの本文データから「です。」「ます。」を含む文を検索し、常体で書かれたレポートには「1」、敬体で書かれたレポートには「0」の値を与えている。

このような概要を各回の課題ごとに作成し、総合的に数値をまとめたものが、以下の表 3 である。

表 3　レポートの概要

| | レポート総数 | 敬体 | 字下げなし | 字数 平均値 | 字数 中央値 | 段落数 平均値 | 段落数 中央値 | 単段落 | 文数 平均値 | 文数 中央値 |
|---|---|---|---|---|---|---|---|---|---|---|
| 第A回 | 87 | 20 | 41 | 1003.7 | 920 | 6 | 5 | 11 | 20.2 | 19 |
| 第B回 | 84 | 3 | 13 | 920 | 1045.5 | 7.6 | 6 | 1 | 21.14 | 20 |
| 第C回 | 81 | 2 | 5 | 1083 | 1031 | 7 | 6 | 0 | 21.07 | 21 |
| 第D回 | 86 | 1 | 5 | 1704 | 1743.5 | 10.3 | 9 | 0 | 33.1 | 33 |

第 A 回から第 C 回までは指定字数が同じであったこともあり、字数・段落数ともに大きな変化はない。第 D 回も字数・段落数の比率は第 C 回までとさほどの変化はない。つまり、授業期間を通じて文字数に対する段落数は大きく変化しない。一方、敬体で書かれたレポートは第 B 回以降、ほとんどみられなくなる。また、字下げのないレポートも第 B 回から減少していくが、第 C 回と第 D 回でもそれぞれ 5 件のレポートは字下げをしていなかった。また、全文を単独の段落で書くレポート（表中では「単段落」）も第 B 回では 1 件のみ、第 C 回からは 0 件となった。

このように、常体の使用・段落の字下げ・段落数については、数値化するこ

とで学習効果を確認することができる。また、体裁が適切でない提出者についての指導も効果的に行うことが可能である。

### 4.3　文章の構成能力

以下の表 4 は MTmineR の Length 機能を利用して、一文の文字数を一覧に抽出したものである。MTmineR では文字数の間隔を任意に変更できるため、本稿では 20 字ごとに区切り、200 字以上は一括して示す（s>=200）こととした。

稿者の経験では、提出されたレポートで一文が 140 字を越えるものは文意をくみとりにくいものが多い。また一文が長いほど、文の主語と述語が対応していない、また並列関係が整理されていないなど、悪文となっている可能性が高い。この一覧で学生番号と文字数を対応させれば、悪文を含む可能性のあるレポートを容易に見つけ出すことができる。

表 4　一文の文字数一覧

| 学生番号 | s1-20 | s21-40 | s41-60 | s61-80 | s81-100 | s101-120 | s121-140 | s141-160 | s161-180 | s181-199 | s>=200 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1300000001 | 1 | 22 | 13 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1300000002 | 1 | 7 | 7 | 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1300000003 | 0 | 18 | 10 | 7 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1300000004 | 1 | 2 | 4 | 4 | 2 | 3 | 0 | 1 | 0 | 0 | 0 |
| 1300000005 | 1 | 5 | 10 | 6 | 3 | 2 | 1 | 0 | 0 | 0 | 0 |
| 1300000006 | 2 | 17 | 20 | 2 | 3 | 1 | 1 | 0 | 0 | 0 | 0 |
| 1300000007 | 1 | 10 | 10 | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 1300000008 | 2 | 8 | 9 | 4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1300000009 | 0 | 3 | 6 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1300000010 | 0 | 5 | 8 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1300000011 | 4 | 25 | 14 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1300000012 | 1 | 7 | 8 | 7 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1300000013 | 1 | 25 | 11 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1300000014 | 7 | 14 | 18 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1300000015 | 0 | 3 | 9 | 5 | 2 | 2 | 3 | 0 | 0 | 0 | 0 |
| 1300000016 | 1 | 11 | 24 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1300000017 | 0 | 17 | 14 | 4 | 3 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1300000018 | 0 | 6 | 6 | 4 | 4 | 2 | 1 | 0 | 0 | 0 | 0 |
| 1300000019 | 2 | 7 | 8 | 3 | 4 | 1 | 0 | 0 | 0 | 0 | 1 |
| 1300000020 | 2 | 17 | 6 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |

注意：学生番号および数値はダミーである。

さらに、第 A 回から第 D 回までの全レポートを対象として一文の長さを計測し、その数値を集約したのが、次に示す表 5 である。

表 5　一文の文字数概要

| | s1-20 | s21-40 | s41-60 | s61-80 | s81-100 | s101-120 | s121-140 | s141-160 | s161-180 | s181-199 | s>=200 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第A回 | 157 | 644 | 507 | 265 | 102 | 44 | 19 | 8 | 4 | 2 | 6 |
| 第B回 | 119 | 629 | 535 | 282 | 114 | 41 | 25 | 10 | 5 | 1 | 15 |
| 第C回 | 130 | 606 | 538 | 255 | 112 | 31 | 13 | 8 | 4 | 3 | 7 |
| 第D回 | 187 | 1023 | 936 | 412 | 158 | 64 | 28 | 12 | 4 | 4 | 18 |

表をみると、一文の長さの分布は変化していない。140 字以上の文もほぼ同様の推移をみせている。2020 年度の「国語表現」では、読みやすい分量の文を書くように指導したが、その効果は薄かったといわざるをえない。提出されたレポートのなかで 140 字をこえる文をみると、修飾表現が整理されていないた

めに文意のとりづらいもの、並列関係が適切で無いものなどがあった。ただし、特定の受講者に固定しているのではないため、指導は全体的なものとならざるをえない。

このように、一文の長さを数値化することで、授業の課題が浮き彫りとなった。

つづいて、使用頻度の高い語の変化をみる。まず、課題内容によって受ける影響の少ないと想定される副詞および副詞 B（ひらがなのみの副詞）を確認したい。分析にあたっては KH coder を使用し、上位 20 語を示す（表 6）。

表 6　副詞

| 第A回 | | 第B回 | | 第C回 | | 第D回 | |
|---|---|---|---|---|---|---|---|
| 実際 | 33 | 実際 | 37 | 実際 | 32 | 実際 | 48 |
| 必ずしも | 19 | 必ず | 22 | 特に | 16 | 特に | 32 |
| 仮に | 15 | 特に | 18 | 一度 | 9 | 最も | 22 |
| 少し | 15 | 既に | 7 | 最も | 9 | 同時に | 9 |
| 本当に | 14 | 別に | 7 | ある程度 | 7 | ある程度 | 7 |
| 最も | 13 | 全く | 6 | 当然 | 7 | 仮に | 7 |
| 当然 | 11 | 必ずしも | 6 | 同時に | 7 | 既に | 7 |
| 特に | 11 | 仮に | 4 | 比較的 | 7 | 一層 | 6 |
| 全く | 8 | 極めて | 4 | 更に | 6 | 初めて | 6 |
| 比較的 | 8 | 最も | 4 | 全く | 6 | 当然 | 6 |
| 共に | 6 | 多々 | 4 | 未だに | 6 | 極めて | 5 |
| ある程度 | 5 | 当然 | 4 | 少し | 5 | 更に | 5 |
| 更に | 5 | 未だ | 4 | 未だ | 5 | 全く | 5 |
| 少なくとも | 5 | ある程度 | 3 | 一層 | 4 | 共に | 4 |
| 多少 | 5 | 一概に | 3 | 極めて | 4 | 互いに | 4 |
| 到底 | 5 | 果たして | 3 | 元々 | 4 | 次に | 4 |
| 一概に | 4 | 少し | 3 | 依然として | 3 | 次第に | 4 |
| 一見 | 4 | 同時に | 3 | 仮に | 3 | 徐々に | 4 |
| 単に | 4 | もう一度 | 2 | 既に | 3 | 常に | 4 |
| 果たして | 3 | 何故 | 2 | 必ずしも | 3 | 比較的 | 4 |
| 決して | 3 | 概ね | 2 | もう一度 | 2 | 未だ | 4 |
| 現に | 3 | 現に | 2 | 何故 | 2 | 一概に | 3 |
| 初めて | 3 | 互いに | 2 | 果たして | 2 | 決して | 3 |
| 徐々に | 3 | 徐々に | 2 | 初めて | 2 | 今や | 3 |
| 大いに | 3 | 度々 | 2 | 別に | 2 | 必ずしも | 3 |
| 同時に | 3 | 同じく | 2 | | | | |
| | | 比較的 | 2 | | | | |
| | | 本当に | 2 | | | | |
| | | 未だに | 2 | | | | |

第 A 回から第 D 回への推移をみると、「実際」「仮に」「最も」「当然」「特に」「全く」など、20 語でも更に上位の語は入れ替わりが少ない。下位の語では「少なくとも」「多少」「到底」など、使用のなくなる語もあるが、多くは第 B 回以降でも使用される。

副詞 B（表 7）でも傾向は同様である。「さらに」「まず」の使用は全回にわたって頻度が高い。その他の語も、頻度に差はあるものの、第 B 回以降も使用さ

表７　副詞Ｂ

| 第A回 | | 第B回 | | 第C回 | | 第D回 | |
|---|---|---|---|---|---|---|---|
| さらに | 51 | さらに | 35 | さらに | 54 | さらに | 73 |
| まず | 48 | どう | 29 | より | 36 | まず | 55 |
| もし | 25 | まず | 27 | まず | 26 | より | 37 |
| どう | 24 | ほとんど | 16 | そう | 13 | どう | 21 |
| そう | 21 | そう | 10 | もちろん | 10 | これから | 20 |
| あまり | 17 | なぜ | 10 | なぜ | 8 | そう | 11 |
| なぜ | 13 | あくまで | 8 | まだ | 8 | すでに | 10 |
| やはり | 12 | より | 7 | こう | 7 | なぜ | 10 |
| これから | 11 | これから | 6 | これから | 7 | ほとんど | 10 |
| より | 11 | もちろん | 6 | どう | 7 | もちろん | 10 |
| また | 10 | むしろ | 5 | ほとんど | 7 | もう | 9 |
| かなり | 9 | もし | 5 | もし | 6 | わずか | 9 |
| ますます | 9 | わずか | 5 | もっと | 6 | やはり | 7 |
| むしろ | 8 | あまり | 4 | あまり | 5 | とても | 6 |
| もちろん | 8 | あまりに | 4 | もはや | 5 | また | 6 |
| すでに | 7 | いまだ | 4 | いまだ | 4 | かなり | 5 |
| ほとんど | 7 | しっかり | 4 | なかなか | 4 | ますます | 5 |
| もう | 7 | そのまま | 4 | もともと | 4 | あくまでも | 4 |
| かつて | 6 | たしかに | 4 | やはり | 4 | こう | 4 |
| もっと | 6 | もう | 4 | あくまで | 3 | ともに | 4 |
| あまりに | 5 | あくまでも | 3 | あくまでも | 3 | あまり | 3 |
| こう | 5 | あらかじめ | 3 | いまだに | 3 | いかに | 3 |
| このように | 5 | こう | 3 | しっかり | 3 | そのうち | 3 |
| そもそも | 5 | さほど | 3 | すぐ | 3 | そのまま | 3 |
| たしかに | 5 | そもそも | 3 | たしかに | 3 | そもそも | 3 |
| たとえ | 5 | とても | 3 | | | たしかに | 3 |
| とても | 5 | ほぼ | 3 | | | どうしても | 3 |
| | | まだ | 3 | | | なかなか | 3 |
| | | やはり | 3 | | | ひいては | 3 |
| | | | | | | ほぼ | 3 |
| | | | | | | まだ | 3 |
| | | | | | | よく | 3 |

れる。

このように、副詞に限定すれば、授業期間を通じた語彙の変化は大きくはない。頻用される語のうち、「本当に」「果たして」「もし」「そもそも」「なぜ」など、アカデミックライティングでの用語として適切ではない語もみられたため、来年度以降の「国語表現」での指導材料として考慮したい。また、頻用される語を、たとえば社会科学分野で公刊された学術論文の使用語彙と比較すれば、大学生のレポートの特徴語をみいだすことができるだろう。

## おわりに

以上、稿者が担当した2020年度「国語表現」でのレポート課題をもとに、テキストマイニング技術の分析手法を用いて、教育効果を測定した。縷々述べたとおり、本稿では極力簡便な技術の導入を心がけた。

教育効果の測定については、稿者が「国語表現」の授業内容として重視していることがらを中心に行った。その結果、語彙の豊富さについてはわずかながら一時的に低下することが確認された。上述のとおり、これは別の年度と対照することで、より詳細な研究を進めたい。また、レポート執筆の基本的な所作の指導については、即効性が高いことを確認することができた。しかし、一文の長さやアカデミックワードなどについては、まだ授業を改善するための課題があることが明らかになった。

このように、テキストマイニング技術を導入することで、機械的な採点は難しいにせよ、教育効果は確認することができる。また、レポートを採点する際にも補助的手段として運用することが可能である。

なお、レポートの論理構成については論考の都合上ふれられなかったので、別稿での課題としたい。

**参考論文**

今田水穂（2018）:「語彙多様性指標の可視化と単回帰分析による TTR の補正」『言語資源活用ワークショップ発表論文集』3

金明哲（2016）:「教育と研究のためのテキストマイニングツール MTMineR（5.4）」『日本計算機統計学会大会論文集 』30（0）

椿本弥生・柳沢昌義・赤堀侃司（2008）:「人文・社会科学分野を中心とした大学教員によるレポート実施と採点の現状に関する調査」『メディア教育研究』5-2

椿本弥生・赤堀侃司（2004）:「レポート採点支援システムの開発と評価」『日本教育工学会大会講演論文集』20

鄭弯弯・金明哲（2018）:「変動係数を用いた語彙の豊富さ指標の比較評価」『同志社大学ハリス理化学研究報告』58（4）

樋口耕一（2020）:『社会調査のための計量テキスト分析（第 2 版）』ナカニシヤ出版

渡邊博之（2008）:「レポート自動採点支援システムを用いたプログラミング演習の評価」『工学・工業教育研究講演会講演論文集 』2008（0）

**注**

注 1 おおむね「共起ネットワーク」で対象テキスト（群）の概要を知ることができるのだが、初期設定では抽出される語が極端に少ない場合がある。そのため、二次的手段として「階層的クラスター分析」が効果的である。これは、「共起ネッ

トワーク」ほどに直感的な示唆を与えるものではないが、堅実な効果を得ることができる。また、補佐的手段として「対応分析」が有効である。とくに、「学科」などの変数があると、より効果的である。変数を設定しない場合は累積寄与率が低くなる。しかし、学術的厳密性を考えるのではなく、実務的問題として対象テキスト（群）の傾向をみるだけであれば、手段としては有効である。

注2 sはその性質上、分母（述べ語数）が大きくなっても、分子（異なり語数）が極端に少なくない場合、数値としては大きく減少しない。また、異なり語数と述べ語数の合計数が多くなると、同じ割合であってもsの数値は大きくなる。たとえば、異なり語数と述べ語数の比が同じ1:10であっても、2000:20000であるときのsは0.717であるが、4000:40000のときは0.735になる。

注3 第D回では、提出されたレポートの本文中に、参考資料としてウェブサイトのURLが多く含まれており、これが非使用語としてカウントされてしまっている。

補注　本稿で使用したレポートファイルおよびレポート取り込みツールを期間限定で公開する（2021年3月1日～6月1日）。ただし、当該ファイルの使用にかかる不具合等について、稿者は責任を負わない。また、サポートも行わない。

https://www.dropbox.com/s/eo1v9719h1a66hg/RIT_02_02.zip?dl=0

（うえだ　ばく・明治大学准教授）