15.

文章の難易度と語彙の関連性に関する考察

— 学年の違いを特徴づける語彙的要素とは何か —

本研究は，国語教科書のテキストデータをもとに文章の難易度とそこに出現する語彙にどのような関連性があるか考察する．考察においては，国語教科書のテキストデータを形態素解析したうえで，各テキストに出現する語彙の品詞や語義の出現比率を調査したあと，多変量解析によるデータ解析を行う．その結果に基づいて文章の難易度と語彙の出現傾向に深い関連性があることを指摘する．本研究の分析結果は，テキスト分類のための有効な指標になるとともに，わかりやすい文章に求められる要件を考えるうえでも重要と考えられる．

15.1 研究背景と目的

テキストマイニングは，定型化されていない文章の集まりから自然言語処理の技術や統計的な分析手法を駆使し，有用な知識を発掘することを目指す研究領域であるが，文章の読みやすさを測定するリーダビリティー研究も，テキストマイニングの下位領域の一つとして位置づけられる．リーダビリティー研究では，一文あたりの語数などの表層情報をもとに，文章の難しさをランクづけすることを目指している．とりわけ英語を対象とするリーダビリティー研究は，Flesch(1948) や Smith&Kinkaid(1970) など，古くからの先行研究があり，回帰式による読みやすさの計算式が提案されてきた．日本語においても，坂本 (1967)，建石ほか (1988)，佐藤 (2011)，柴崎・原 (2010) の研究があり，ウェブシステムとして計算式が実装されている例もある[1)]．

こうしたリーダビリティーの研究では，次の二つの問題をめぐり，盛んな議論が行われている．

1) 難しさを決定する要因は何か．
2) 個々の要因をどのように重み付けし，公式化するか．

1) の問題に関しては次の事実を考慮する必要がある．文章の難しさは，いくつもの要因が複雑に絡み合って決まっていく．マクロな要素としては，話題や文章全体としてのまとまり具合などが考えられ，ミクロな要素としては，語彙の難しさ，文法構造の難しさ，語や文の長さなどが考えられる．具体的な研究例として，柴崎・原 (2010) は小学校 1 年から高校 3 年までの国語教科書を用いて，線形回帰分析を使い，日本語のリーダビリティー公式を提案しているが，①文章中の

[1)] 佐藤 (2011) および柴崎・原 (2010) は以下のサイトで，測定システムが実装されている．
- 佐藤 (2011)：http://kotoba.nuee.nagoya-u.ac.jp/sc/readability/index.html
- 柴崎・原 (2010)：http://readability.nagaokaut.ac.jp/readability

平仮名の割合，②一文の平均述語数，③一文の平均文字数，④文の平均文節数の四つの要素をとりあげ，数式化を行っている．2) の問題に関しては，何らかの基準データを用いて統計的な手法で数式化を行っており，主成分分析や回帰分析などがよく用いられる．

本研究では，柴崎・原 (2010) を発展させ，文章の難易度と語彙の質的側面に関する分布の関連性を考察する．とりわけ次の2点を検討する．1) 小学1年から中学3年までの国語教科書を基本データにして，内容語（名詞，動詞，形容詞，形容動詞）の一文単位での平均出現回数を調べた．2) 国語教科書に出現した語彙に対して「分類語彙表–増補改訂版」(以下，「分類語彙表」) と照合し，多義語（複数の語義を有するもの）と単義語（語義が一つのみのもの）の出現率を調べた．そして，これらの調査結果の妥当性を検証すべく，「現代日本語書き言葉均衡コーパス」(Balanced Corpus of Contemporary Written Japanese，以下 BCCWJ) に収録された教科書データに対して，正準判別分析 (canonical discriminant analysis) を行った．結論として，小学1年から中学3年のテキストは，四つの離散的グループ（1; 小学1年〜2年，2; 小学3年〜4年のグループ，3; 小学5年〜中学1年のグループ，4; 中学2年〜中学3年）に分類できることを示したうえで，BCCWJ の教科書データに対して，語彙分布から文章の難易度グループを判別する統計タスクを行った．

15.2　調査方法とデータについて

15.2.1　概要

本研究では，基本データと評価データの2種類の教科書データを使用した．基本データは，柴崎・沢井 (2007) によって構築された国語教科書45冊分のテキストデータである．評価データは，BCCWJ に含まれる教科書データである．この2種類のデータの詳細については15.2.3項で述べる．これらの基本データと評価データに対して，2種類の調査を行った．

(1)　a. 品詞分布の調査（以下，調査1）：学年別の品詞の分布に関する調査
　　b. 語義分布の調査（以下，調査2）：学年別の語義の数に関する調査

調査1は，語の文法的・形態的振舞いと文章の難易度の関係性を検討するためのものであり，調査2は，語の意味的振舞いと文章の難易度の関係性を検討するためのものである．15.2.2項で調査方法の詳細を，15.2.3項で使用データの詳細を，15.2.4項で統計手法の詳細を述べる．

15.2.2　調査方法と予想される結果

A.　調査1について

調査1では，形態素解析済みのデータを用いて品詞単位で語彙の出現頻度を調べる方法で行った．これは，品詞の分布を調査することで，名詞をはじめとする各内容語の占める割合が学年によって差があるかどうかを調べることを目的とする．このことに関連する先行研究の観察として，「名詞率と漢語率は正の相関である」(樺島，1963)，「要約的な文章では名詞率が高い」(樺島，1988)，「内容語が多く含まれた表現は要約的で難易度も高い」(佐野，2008) という指摘がある．さらに，柴崎・沢井 (2007) では，「学年が上がるほど，語種の割合において漢語の割合が大きくなる」という調査結果を報告しており，これらの先行研究の分析結果を整理して考えてみた場合，次のような予測が成り立つ．それは，学年が上がるほど語彙表現の難易度も高くなるという前提を是とするなら，学年が上がるほど名詞の出現頻度は多くなることが予測される．この予測の妥当性を調査1で検討する．

B.　調査2について

調査2では，学年別語彙の語義の分布を調査する．調査は「分類語彙表」に基づいて，個々の語彙がいくつの意味をもっているかを調べる方法で行った．

調査2の基本となる考えについて述べる．基本語に相当する語彙は，さまざまな文脈で自由に用いられるため，語義の数が多く，曖昧性が高い．たとえば，「取る」という語には，手の中におさめる，握る，獲得する，持つ，処理する，没収する，体から一時的に離す，除き去る，導き入れるなど，多数の語義が内包されている．一方，難解語は専門語などからわかるように，使用文脈が限定されていると同時に，語義の曖昧性が低い．たとえば，「希釈する」は溶液に溶媒を加えて濃度を薄めるという意味だけである．語義の数と基本語としての度合いに部分的な相関関係が認められるなら，学年別の多義語と単義語の分布はテキストの難易度にも何らかの関係性をもっていると考えられる．

上記の問題意識のもとで，「分類語彙表」を用いて調査を行った．「分類語彙表」を用いた理由としては，大量の語彙の多義性の有無を人の内省で判断することは難しいが，「分類語彙表」では語義に対応した意味分類を行っているため，項目の登録数を調べることで，多義性の有無について判断ができると考えたからである．「分類語彙表」には各語の意味分類が体系的に網羅されているので，この分類に基づいて，国語教科書に出現した全単語が，どの項目に分類されているかを整理すれば，複数項目に登録されている語を多義語，単一の項目として登録されている語を単義語に分けることができる．調査方法としては，コンピュータプログラム「Python(http://www.python.jp/)」を介して，語彙素の基本形，発音，品詞をベースに「分類語彙表」と対応をとり，該当するすべての意味分類を「類–部門–中項目–分類項目」の形式で出力するような計算プログラムを作成した．これで調査対象の全単語に意味分類を付与した後，集計を行った．

15.2.3　調査データについて

本研究では，分析の基礎資料として国語教科書を使用する．その理由として，国語教科書は児童生徒の発達段階に応じて編集されたものである．そのため，一般的な書物と違って，内容の適切さや構成の難しさなどが統制されているものと考えられる．このことは，言語表現においても反映されていると考えることができ，表記，表現，文法，語彙においても学年間で何らかの差が確認できるのではないだろうか．教科書データのこうした特徴を利用し，学年別に出現する語彙を定量的に分析することで，学年によって変化する文章の難易度に対して，語彙がどのようなバイアスを与えるのかについて，基礎調査ができると考えた．

基本データは，柴崎・沢井 (2007) によって構築されたもので内訳は以下のとおりである．

(2)　a. 小学校6学年（上下）×3種類（光村図書・東京書籍・教育出版）の合計36冊
　　b. 中学校3学年 ×3種類（光村図書・東京書籍・三省堂）の合計9冊

(2) の全データを「MeCab（形態素解析エンジン）」と「UniDic（形態素解析辞書）」で解析し，延べ頻度460,756語の形態素解析済みデータを作成した．そして，助詞類や助動詞類などの機能語を取り除き，内容語に相当する「名詞，動詞，形容詞，形状詞（「形容動詞」に相当するUniDic上の品詞名）」のみを集計した．

次に，評価用データとしてBCCWJの教科書データを利用した．BCCWJのデータを利用するにあたって，次の2点を考慮し，データとしての適性を判断した．評価データは，基本データの分析で得られた知見の妥当性を検証するためのもので，1) 基本データとは異なるものであること，2) （学年との対応を検討するために）どの学年のテキストであるかが明記されたテキストデータであることが求められるが，この2点の条件を満たすものとしてはBCCWJの教科書データがもっともよいと考えた．

BCCWJには，100万語規模の小学校・中学校・高等学校で採用された各教科の教科書が収録されており，コーパスサイズとしては本研究が用いる基本データよりも大きい．また，国語以外に外国語，技術家庭，芸術，数学，社会，生活，理科のデータも入っており，基本データ以上の網羅性をもったテキストデータベースである[2]．本研究では，小1から中3（9学年）までの教科書に収録された161編の文章を対象に調査1と調査2で得られた知見の妥当性を検証すべく計算機実験を行った．実験では，BCCWJの教科書データ161編の文章に対して，調査1，調査2で導きだされた四つのグループを指定し，品詞情報や語義情報から難易度を判別するタスクを行った．

15.2.4　統計分析について

統計分析では，判別分析を行った．一般に判別分析では，データの判別ルールを作成することと，それに基づき新規のデータを自動的に分類することが主目的になるが，本研究では，異なる変数群で判別タスクを実行し，判別精度の変化を確認した．具体的には，三つのタスクを実行した．1) 調査1の品詞分布から難易度を判別させるタスク，2) 調査2の語義分布から難易度を判別させるタスク，3) 調査1の品詞分布と調査2の語義分布から難易度を判別させるタスクを行った．これら三つのタスクにおける判別精度を比較することで，本研究が仮定する語彙の分布が難易度の判別にどの程度，貢献するのかを明らかにするとともに，品詞分布と語義分布の分類指標としての妥当性を明らかにする．

15.3　結果

15.3.1　調査1の結果

一文における名詞，動詞，形容詞，形状詞の平均出現回数を計算した．その結果，二つの傾向が確認された．第一に，形容詞，形状詞に関しては，どの学年においても，一文あたりの平均出現回数が0.2回から0.3回に留まっており，学年の差を特徴づける顕著な違いが見られない．第二に，名詞，動詞に関しては，若干のぶれはあるものの大まかな傾向として，学年が上がるにつれ，出現回数が上昇傾向にあることが観察できる．

[2] BCCWJの教科書データのサンプル数は次のとおりである．小学校教科書が94編，中学校教科書が67編，高校教科書が251編である．本研究の基本データが小学校から中学校までのデータであるため，BCCWJに関しても小学校から中学校までのデータのみ使用した．

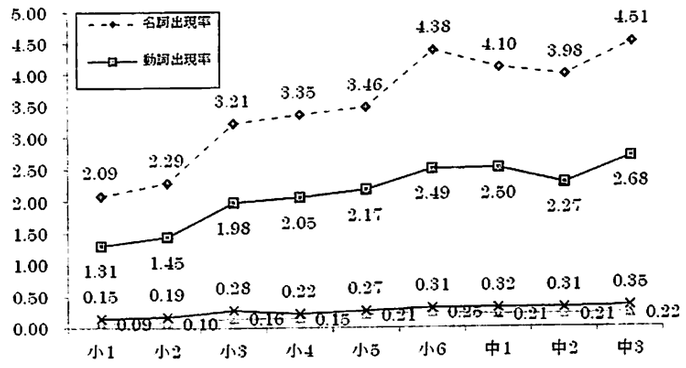

図15.1　名詞，動詞の学年別の一文あたりの平均出現回数の変化

図15.1が示すように，1学年のテキストでは，名詞は一文に平均2.09回が出現しているのに対して，9学年のテキストでは，平均4.51回出現しており，ほぼ2.5倍になっていることがわかる．また，動詞に関しても名詞ほどではないが，1学年から9学年の間で約2倍の違いがあることが明らかになった．

15.3.2　調査2の結果

調査2では，語彙の意味的側面に対する調査を行った．15.2.2項のBで述べた方法で，「分類語彙表」との対応づけを行った[3]．単義語と多義語の集計結果を表15.1に示す．

表15.1　多義語，単義語の出現頻度

	異なり頻度		延べ頻度	
学年	多義語	単義語	多義語	単義語
1学年	418	347	1837	972
2学年	793	656	4172	2084
3学年	1156	1077	6006	3234
4学年	1339	1193	7657	4253
5学年	1717	1617	8895	5631
6学年	2175	2294	12496	7966
7学年	2774	3000	13928	9332
8学年	2726	3073	15274	11467
9学年	2744	3101	15285	11641

表15.1の異なり頻度に注目した場合，1学年から5学年では多義語（例：手，取る，万能，悪口，甘い，山）のほうが多いのに対して，6学年を境界に多義語と単義語の頻度が逆転し，単義語（例：学童，割り込む，患者，喫茶店，強風，金属）の数が一貫して多くなるという結果が示

[3] 「分類語彙表」との対応づけは，名詞は81%，動詞は80%，形容詞は95%，形状詞は85%の割合で成功している．対応づけに失敗した語として，名詞においては固有名など「分類語彙表」に収録されていない語であり，動詞，形状詞，形容詞においてはUniDicの表記と「分類語彙表」の表記が一致しない語であった．

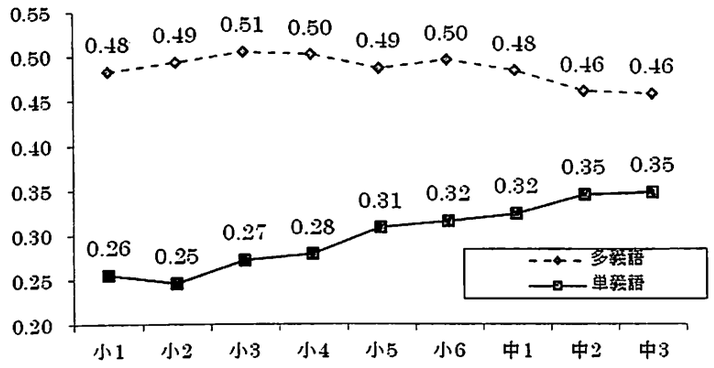

図 15.2　多義語と単義語の学年別の出現比率

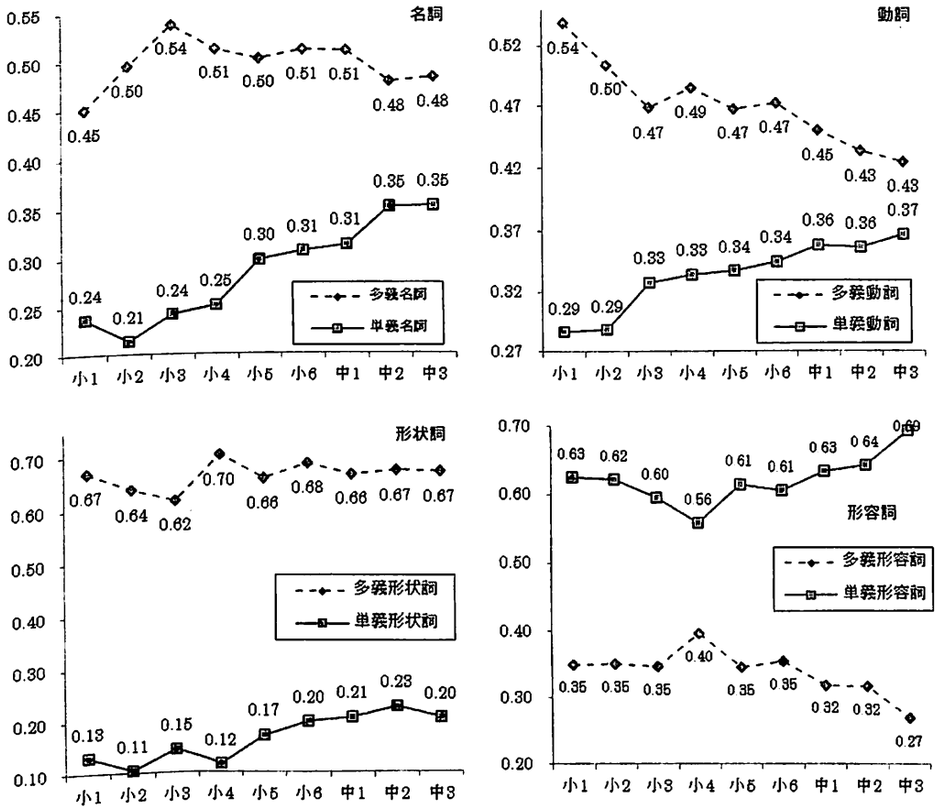

図 15.3　品詞別の多義語・単義語の出現比率

された．延べ頻度に注目した場合，当然の結果であるが，多義語のほうが（単義語に比べ）一語あたりの出現率が高く，多義語は平均 4.3 回～5.6 回出現しているのに対して，単義語は平均 2.8 回～3.7 回出現している．このことを踏まえ，表 15.1 の多義語と単義語の延べ頻度を内容語の延べ頻度で割り出現比率を求めた．その結果，図 15.2 の分布が確認された．

図 15.2 で確認できることとして，多義語に関しては，0.5 前後でほとんどの学年で変化がないのに対して，単義語に関しては学年が上がるに連れて，出現率の上昇傾向が確認できる．このことを踏まえ，さらなる検証として品詞単位で同様の調査を行った．とりわけ，名詞，動詞，形容詞，形状詞に対して単義語と多義語の頻度を集計し，当該品詞の延べ頻度で割った値をグラフ描画した（図 15.3）．

図 15.3 で注目すべきは次の 3 点である．1 点目は，全体的な変化パターンとして名詞，動詞，形状詞に関しては，学年が上がるにつれ，単義語の出現率が上昇していること．2 点目は，動詞に関しては学年が上がるにつれ，多義語が減少傾向にあること．3 点目は，名詞，形状詞，形容詞に関しては，多義語の出現率の増減は ±0.1 に収まっており，大きくは変化しないことである．

15.3.3　調査 1, 2 の統合と検証

調査 1 で用いた品詞分布，調査 2 で用いた語義分布を変数に，9 学年のセグメント化を行うため，階層的クラスター分析を行った．クラスター法はウォード法，サンプル間の距離は平方ユークリッド距離で測定した．

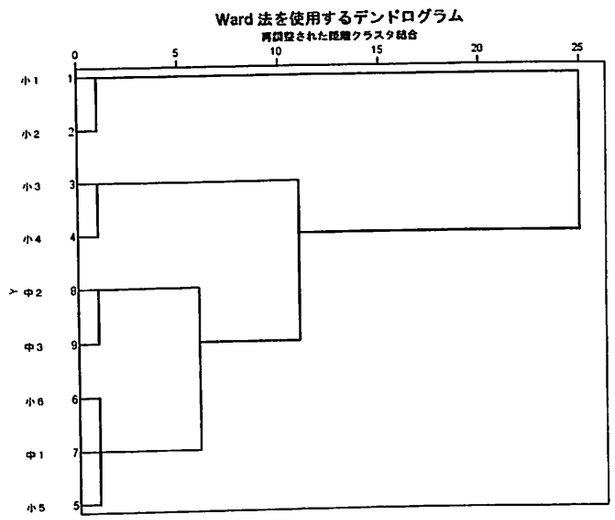

図 15.4　クラスター分析の結果

図 15.4 から次の 3 点が明らかになった．1) 小 1・小 2 のグループとそれ以外で大きな分岐が確認される．2) 小 3・小 4 のグループとそれ以外で分岐し，さらに，3) 小 5 から中 1 のグループと中 2・中 3 のグループで分岐している．クラスター数の検討のため，Kruskal の方法の多次元尺度構成法でデータ分析を行った．尺度モデルはユークリッド距離を使用した．図 15.5 が示すとおり，四つのグループでまとまりをなしていることが確認された (Stress=.01853，RSQ=.99836)．

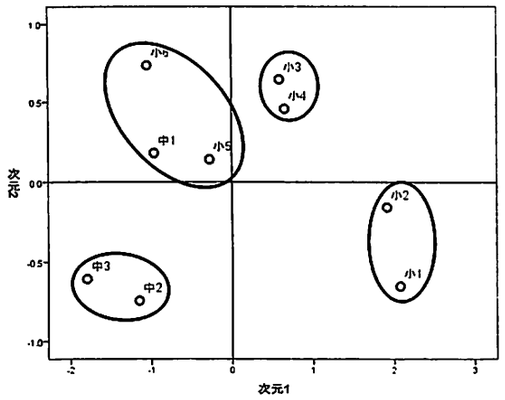

図 15.5　多次元尺度法による分析結果

以上の分析結果を踏まえ，四つのクラスターとして全体をとらえることができると判断した．この四つのクラスターは学年の相違と対応しているため，テキストの難易度を反映するものと分析できる．

15.3.4　BCCWJ の教科書データの分析結果

前節までの分析結果によって，国語教科書のテキストデータは四つの難易度スケールでとらえられることが明らかになった．この分析の妥当性，汎用性を検討すべく，BCCWJ のデータを用いた検証を行った．検証作業においては，BCCWJ の 2012 年度の DVD 版データに収録された教科書データに対して調査 1 および調査 2 と同様の方法で品詞情報と語義情報を付与し，表 15.2 のデータテーブルを作成した．

表 15.2　BCCWJ 分析データ

		一文内の品詞の出現率				語義の出現比率							
学年	難易度グループ	名詞	動詞	形容詞	形状詞	単義名詞	単義動詞	単義形容詞	単義形状詞	多義名詞	多義動詞	多義形容詞	多義形状詞
小 1	1	1.33	0.89	0.11	0.22	0.32	0.33	0.00	0.00	0.52	0.50	0.86	0.81
小 3	2	2.68	1.21	0.08	0.10	0.41	0.27	0.65	0.25	0.49	0.63	0.25	0.65
小 5	3	3.38	1.23	0.08	0.03	0.45	0.31	0.57	0.84	0.35	0.59	0.33	0.12
中 1	3	3.97	2.03	0.26	0.08	0.43	0.34	0.61	0.00	0.47	0.56	0.29	0.91
中 3	4	5.69	1.72	0.28	0.20	0.49	0.44	0.31	0.59	0.41	0.46	0.61	0.31
：	：	：	：	：	：	：	：	：	：	：	：	：	：

表 15.2 の形式で，BCCWJ の教科書 161 編の全データを作成した．統計分析では，難易度グループを従属変数に，一文内の出現率と語義の出現比率を独立変数にして判別分析を行った．判別分析は，三つの条件で行い，それぞれの条件で判別率がどのように変化するかを確認した．判別率を確認したところで，表 15.3 の結果が明らかになった（交差確認は，Leave-one-out cross validation 法で行った）．

表 15.3 の結果から，品詞情報と語義情報をともに従属変数として投入した場合，もっとも判別

表 15.3　判別分析の結果

判別分析の条件	交差確認済み判別率
タスク 1：一文内の品詞の出現率のみで分析	52.2%
タスク 2：語義の出現比率のみで分析	63.5%
タスク 3：一文内の品詞の出現率と語義の出現比率で分析	82.6%

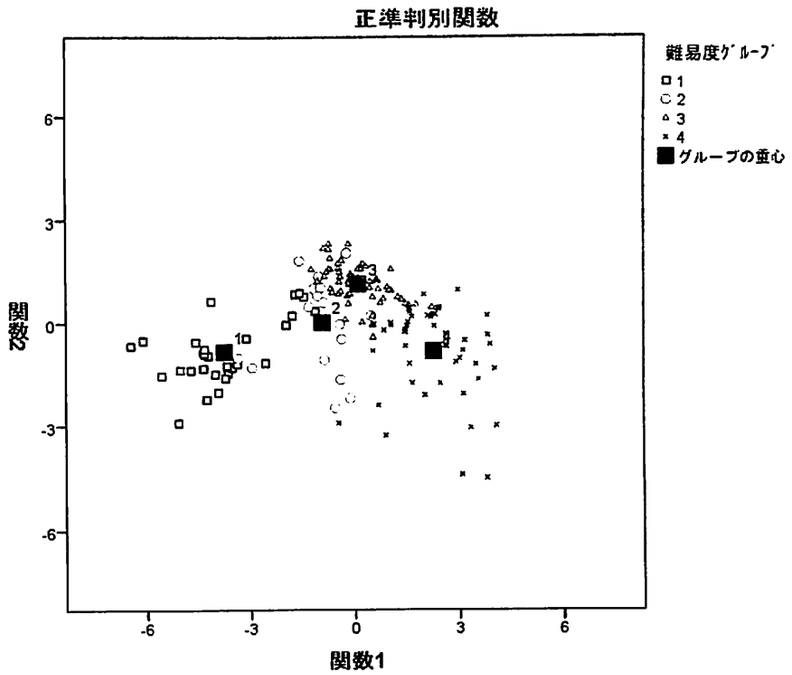

図 15.6　正準判別関数による全グループの散布図

精度が高いことがわかる．そして，品詞情報より語義情報のほうが，判別率が高いことが明らかになった．タスク 3 の三つの判別関数のうち，第一関数と第二関数で全グループの散布図を作成した．

図 15.6 からわかるように，関数 1 によって，1 から，2，3，4 の順で分布していることがわかる．なお，第一関数と第二関数の累積固有値は 97.3% であった．また，Wilks のラムダで関数の検定を確認したところ，第一関数から第三関数までで有意であった (Wilk's Lambda =.153, p = .000)．次に判別関数の係数を確認したところ，第一関数に関しては名詞に関連する変数が高い値を示していることが確認できる（表 15.4）．

以上の結果から，動詞や名詞の一文内の出現回数や多義語・単義語の出現率がテキストの難易度に深く関係している事実が明らかになった．次節では，分析結果を踏まえ，各グループ別，教科別に判別率の詳細を検討する．

表 15.4　正準判別関数係数

	関数		
	1	2	3
動詞の一文平均出現回数	0.410	0.184	−0.481
名詞の一文平均出現回数	0.761	−0.262	0.219
形状詞の一文平均出現回数	−0.377	−0.908	0.487
形容詞の一文平均出現回数	−0.218	0.016	−0.257
多義動詞の出現比率	−0.377	−0.908	0.487
多義名詞の出現比率	−0.535	0.043	0.231
多義形状詞の出現比率	0.030	−0.101	−0.119
多義形容詞の出現比率	−0.459	0.019	0.631
単義動詞の出現比率	0.186	0.075	0.858
単義名詞の出現比率	0.818	0.585	−0.034
単義形状詞の出現比率	0.698	0.212	0.521
単義形容詞の出現比率	0.389	0.963	0.102

15.4　考察

本研究では，学年が上がるにつれ，文章の難易度が上がるということを前提に，各学年の国語テキストに表れる語彙の定量的分析を行った．分析の結果，名詞や動詞などの内容語の一文内の平均出現回数がテキストの難易度の変化に関係していること，単義語の出現率が難易度の変化に関係していることが明らかになった．そして，このことを検証すべく，BCCWJ の 161 編の文章に対して，品詞情報，語義情報の指標を与え，難易度を予測するコーパス実験を行い，87.6% の精度で判別できることが明らかになった．より詳しい考察を行うべく，表 15.3 のタスク 3 の結果に対して難易度グループ別の判別精度を確認した後，科目別の判別精度を確認する．

表 15.5　難易度グループ別の正判別と誤判別の集計

難易度グループ	正判別	誤判別	正判別率	総計
グループ 1（小 1，小 2）	23	4	85.2%	27
グループ 2（小 3，小 4）	25	6	80.6%	31
グループ 3（小 5，小 6，中 1）	45	6	88.2%	51
グループ 4（中 2，中 3）	48	4	92.3%	52
合計	**141**	**20**	87.6%	161

表 15.5 では，グループ 1 からグループ 4 の正判別と誤判別のテキスト数が示されている．グループ 1・2 は，グループ 3・4 に比べ，相対的に誤判別の可能性が高い．次に，教科別の正判別率を確認した．

表 15.6 から，判別精度が著しく低い科目として，「芸術」と「生活」が挙げられる．その理由を模索すべく，テキストとしての長さを調べてみた．全テキストの平均語数を調べてみたところ，表 15.7 の結果が明らかになった．

表 15.7 でわかることとして，「芸術」と「生活」はいずれも一つのテキストが有する平均語数

表 15.6　教科別の正判別と誤判別の集計

教科	正判別	誤判別	正判別率	総計
外国語	7	0	100.0%	7
技術家庭	8	1	88.9%	9
芸術	28	9	75.7%	37
国語	26	2	92.9%	28
社会	24	1	96.0%	25
数学	27	2	93.1%	29
生活	1	2	33.3%	3
理科	20	3	87.0%	23
総計	**141**	**20**	**87.6%**	**161**

表 15.7　テキストの平均語数

教科	総語数	平均語数
外国語	3250	464. 3
技術家庭	20431	2270.1
芸術	13007	351.5
国語	26917	961.3
社会	40356	1614.2
数学	27970	964.5
生活	433	144.3
理科	25434	1105.8
総計	**157798**	**980.1**

が短いという特徴をもっており，このことが判別精度に影響している可能性が考えられる．比較的良好な判別精度を示している「外国語」，「国語」，「数学」，「社会」を基準に考えると 500 語以上のテキストにおいて，安定した難易度判別ができるといえよう．

15.5　最後に

本研究では，テキストデータがもつ潜在的性質として，文章の難易度，すなわちリーダビリティーに関する問題を取り上げ，語彙の分布がどのように影響するかを検討した．分析の結果として，品詞や語義の出現比率がリーダビリティーに深く関係していることが明らかになった．こうした分析の結果は，教育分野に限らず，製品マニュアルのようにわかりやすい文章が求められる実社会においても貢献するものが多いと考えられる．

テキストの難易度という概念は物質がもつ固有の質量のように実在する概念ではない．このことを前提に考えると，難易度の設定方法および設定スケールは，記述の粒度に帰結するものであり，唯一無二の答えは存在しない．その意味において，難易度というのは，あくまで仮想的なスケールであるといえる．では，そもそも正解がないものに対して，調査し，分析することの意義はどこにあるのだろうか．それは，一言でいえば，テキストを分類することに役に立つからである．リーダビリティー研究がテキストマイニングの延長で位置づけられる経験的根拠はここにある．リーダビリティー研究の知見を利用することで，難しいテキストとやさしいテキストを区別し，さらには難しいテキストに対して，どの程度難しいかを仮想的なスケール上でポジショニングでき，またはやさしいテキストに対して，どの程度やさしいかをポジショニングできる．

今後，精緻化された語彙難易度の尺度を設定し，それをリーダビリティー式の変数に加えていけば，さらに精度の高い式が構築されることが期待できる．

＊ 本研究は，李在鎬・柴崎秀子 (2008)「日本語リーダビリティー公式構築のための国語教科書語彙の分析」（計量国語学会 2008 年度年次大会）を加筆修正したものである．また，本研究は，科学研究費補助金特定領域研究「日本語コーパス」（課題番号:19011003）の援助を受けて行ったものである．

参考文献

[1] 樺島忠夫. 漢語をめぐって, 計量国語学, 27, pp. 14–19, 計量国語学会. 1963.

[2] 樺島忠夫. 日本語はどう変わるか —— 語彙と文字 ——. 岩波書店. 1988.

[3] 佐野大樹. 大規模バランスコーパスにおけるテキスト分類 —— システミック理論の観点から ——, 特定領域研究日本語コーパス平成 20 年度全体会議予稿集, 83–90, 特定領域研究「日本語コーパス」総括班. 2008.

[4] 佐藤理史. 均衡コーパスを規範とするテキスト難易度測定. 情報処理学会論文誌, 52–4. pp. 1777–1789, 情報処理学会. 2011.

[5] 柴崎秀子・沢井康孝 . 国語教科書コーパスを応用した日本語リーダビリティー構築のための基礎研究, 信学技報 NLC2007–32(2007–10). pp. 19–24, 電子情報通信学会. 2007.

[6] 柴崎秀子・原信一郎. 12 学年を難易尺度とする日本語リーダビリティー判定式, 計量国語学, 27–6, pp. 215–232, 計量国語学会. 2010.

[7] 建石由佳・小野芳彦・山田尚勇. 日本文の読みやすさの評価式. 文書処理とニューマンインターフェース. 18–4, pp.1–8, 情報処理学会. 1988.

[8] Flesch, R. A new readability yardstick, *Journal of Applied Psychology*, Vol. 32, 221–233. 1948.

[9] Smith A. Edgar and Kincaid J. Peter. Derivation and Validation of the Automated Readability Index for Use with Technical Materials. Human Factors: *The Journal of the Human Factors and Ergonomics Society*, 12–5, pp. 457–464. 1970.

[10] 李 在鎬・柴崎秀子. 日本語リーダビリティー公式構築のための国語教科書語彙の分析, 計量国語学会 2008 年度年次大会予稿集, pp. 18–19. 2008.

言語資源

[1] 国立国語研究所「分類語彙表-増補改訂版」: http://www.kokken.go.jp/kanko/goihyo/syokai/

[2] 国立国語研究所「現代日本語書き言葉均衡コーパス」: http://www.tokuteicorpus.jp/

(李 在鎬, 柴崎秀子)