論文

やさしい日本語のリーダビリティ研究
―想定読者が異なるニュース報道文の比較から―

言語教育研究科言語教育学専攻
博士後期課程 1 年
蓮田善郎

要　旨

本研究は、日本語文章を「やさしい日本語」で書き換えることで文章難易度がどのように変化するかを、文章を構成する形態的要素の出現頻度の違いから分析したリーダビリティ研究である。まず日本語リーダビリティの先行研究を概観した後、想定読者の違いから文章難易度に差があると推定できる 3 種類のニュース報道文（一般向け、やさしい日本語書き換え、小学生向け）を対象に、27 種類の形態的要素が文章に現れる頻度を比較・考察した。その結果、文章難易度に影響を与えるのは 18 要素で、この 18 要素は「語彙・文法要因」と「長さ要因」の 2 つの要因に分類できること、「やさしい日本語」の書き換えルールに示されたもの以外にも文章難易度に影響する要因があることを明らかにした。

【キーワード】やさしい日本語、リーダビリティ、文章難易度、ニュース報道文

１．研究目的

災害時の緊急支援情報を外国人に伝達するために佐藤(2004)が提案した「やさしい日本語」が、日常生活を支援するためにも活用され始めている。観光案内や役所のお知らせ、ニュース報道文などである。やさしい日本語への書き換えには 2 つの方法がとられると考えている。1 つは形態面からやさしさを追求する方法で、「漢語を和語にする」「受身文を使わない」など語彙や文法の制限で、書き換えルールが公表されている。もう 1 つは内容面からやさしさを追求する方法で、元の情報から理解するのが難しそうな部分を削除してしまう手法である。本研究は 1 つ目の制限である形態面からのやさしさ追及に着目したリーダビリティ研究である。想定読者が違う 3 種類のニュース報道文を分析し、文章を構成する形態的な要素の組み合わせ方が異なると文章難易度が変わることを統計分析で示したうえで、各要素の重み付けの違いから文章難易度を測る指標を導く基礎的要素を探るのが

目的である。さらに、やさしい日本語の書き換えルールとの比較から、本研究で特定した形態的要素の有効性も検討する。

なお本研究では、文章のわかりやすさとは、その文章が表現しようとする内容を読み手が容易に理解できる状態や程度ととらえている。このため、本研究で扱う形態面からのやさしさ追及だけでなく、内容面からのやさしさ追及も実現できたときに、わかりやすい日本語文章になると考えている。

2．日本語のリーダビリティ研究

リーダビリティ研究は、文章の読みやすさを示す指標を探る研究領域である。言語学だけでなく、教育学や自然言語処理など複数の研究分野にまたがる境界領域のテーマとされる。日本語を対象にした研究は 1950 年ごろから始まり、文章の測定可能な形態的要素（文の長さ、高頻度語彙の割合など）から、文章難易度を判定する計算式が開発されてきた。

森岡(1952:91)は小・中学校の教科書と専門雑誌、総合雑誌、大衆雑誌、児童雑誌の調査から、文の長さが平均 60 字以上の文章を「難しい」、35 字以下の文章を「やさしい」とし、また漢字の割合が 35%以上の文章を「難しい」、20%以下の文章を「やさしい」と評価する基準を示した。阪本(1971:1)は読みやすさのレベル分けを学校教育における学年に置き、読みやすさに影響を及ぼす要因は「漢字率」「基本語彙率」「長文率」であると論じた。

こうした先行研究の影響から、どういった文章構成要素が難易度に影響するかを測るリーダビリティ研究は、学校教科書の文章を分析対象とする形で進んでいった。柴崎・原（2010:215）は「文の長さ」「文字種の割合」「語種の割合」「文法構造の複雑さ」を要因とする重回帰分析による 3 種類の計算式を提案し、その計算結果から文章難易度を小学校 1 年から高校 3 年までの 12 段階に分類した。

語彙の難易度に着目した川村・北村（2013:18）は、旧日本語能力試験で公表されていた級別語彙表を活用し、「一文当たりの平均単語数」「出題基準 1 級の単語数/総単語数」「出題基準 2 級の単語数/総単語数」「出題基準 3 級の単語数/総単語数」「出題基準 4 級の単語数/総単語数」の 5 つの要素で、文章難易度を示す重回帰式を提案した。李（2016:1）は、難易度判定に影響する要素は「平均文長」「漢語率」「和語率」「動詞率」「助詞率」の 5 項目だとする重回帰式を提案した。

ただ、こうした重回帰式による難易度評価には限界があるという指摘もある。山村(2014:1063)は、「高学年における推定性能が悪いという問題がある」とし、学習者のレベ

ルに応じて判断項目を変えると判定結果の精度が高まることを統計的に示した。あらゆる文章の難易度を 3、4 個ほどの要素で評価するのが難しいことを指摘している。

３．研究の流れ

先行研究は、表記文字や使用語彙、長さ、品詞など、文章を構成する形態的な要素の組み合わせ方が異なると文章の難易度が変わることを、ある程度の精度で割り出せることを示している。一方で、問題点も確認できた。⑴形態的要素は分析者が主観で選定した 5 個程度しか使っていない、⑵文章難易度の推測は回帰式で示されるため判定誤差が大きくなる、⑶判定結果を「×年生レベル」と表現するために分析対象は学校教科書を基本にしている、の 3 点である。本研究はこれらを課題点ととらえ、修正を試みる。

具体的には、学年レベルでは判定できないニュース報道文を分析対象とし、文章難易度に影響しそうな構成要素を多数リストアップしたうえで絞り込みをおこなう。統計処理は回帰分析ではなく、多数の構成要素の影響を考慮できる主成分分析を採用し、各ニュース報道文の形態的な違いを 2 つ程度の要因に集約させる方法をとる。そのうえで、想定読者の違いに起因する文章難易度の差にはこれらの要因が影響していることを示す。また、やさしい日本語の書き換えルールには示されていない要素の影響についても考察する。

４．分析対象

４．１　分析対象の選定

分析対象はニュース報道文とする。ニュース報道文は、⑴5W1H を基本とする文章構成、⑵十分な推敲、校閲がおこなわれ表現上の不適切な部分がない、⑶あいまいな解釈を避けるため文構造がシンプル、⑷表記ルールが統一されているなどの特徴があり、表現方法のばらつきが抑えられているためである。具体的には「NHK NEWS WEB」（以下、[一般]と表記）、「NEWS WEB EASY」（以下、[やさしい]と表記）、「毎日小学生新聞」（以下、[小学生]と表記）の 3 つのニュース報道文を対象とする。

[一般]は NHK が放送するニュース原稿を WEB 上で公開したものである。[やさしい]は、やさしい日本語を使って[一般]を書き換えたもので、想定読者は日本語学習者や日本の小学生である。[小学生]は、やさしい日本語を使わずに内容を理解しやすく書いたもので、想定読者は日本の小学生である。[やさしい]と[小学生]はどちらも日本の小学生が理解できるレベルである。また、それぞれが使用する漢字は、[一般]は義務教育終了までに学習

を終える常用漢字を使い、[小学生]は小学校学年配当漢字を使用している。[やさしい]は明確な記載はないが、使用語彙の制限から旧日本語能力試験の 3 級、4 級レベルと推定できる。こうした想定読者や使用漢字の違いは、3 種類のニュース報道文の文章難易度の違いとして現れると考えられる。

分析対象は 2021 年 12 月の 1 カ月間に公表されたすべての記事で、その数は[やさしい]が 74、[小学生]が 78 である。ただし、[一般]は書き換えの元になった記事だけを選んだため[やさしい]と同じ 74 である。また[やさしい]に書き換えた記事は[一般]から数日遅れて掲載されることがあるため、[一般]には 21 年 11 月に公表されたものが含まれる（表 1 参照）。

表１　分析対象データ

	NHK NEWS WEB [一般ニュース]	NEWS WEB EASY [やさしいニュース]	毎日小学生新聞 [小学生ニュース]
想定読者	義務教育終了	日本語学習者、小学生	小学生
記事本数	74	74	78

※ 2021年12月に公表された記事
※ NHK NEWS WEBはNEWS WEB EASYの元となった記事だけを抜粋

さらにここから 3 種類のニュース報道文が共通して同じ事案を扱った記事に絞り込んだところ、該当するものは 9 組だった。だが 9 組のデータでは、量的分析をするには十分なデータ量とは考えられない。このため本研究では、扱う記事内容は完全に一致していないが、同一期間に公表された記事という条件をつけてデータ収集に客観性をもたせた。

４．２　形態的要素 27 要素の選定

文章難易度は形態的要素の現れ方の違いで判定できることを先行研究が示している。そこで本研究では、3 種類のニュース報道文を構成する形態的要素の出現頻度の違いが、文章難易度の違いに反映されると考えた。他のニュース報道文に比べて明らかに出現頻度が異なっている形態的要素を特定するために、まず分析候補を選定する。

文章の特徴を計量する指標を整理した浅井(2017:159)は、テキストの難易度推定には「文字種別」「品詞別」「語種別」「単語長」「文長」が使われると論じている。また『計量国語学事典』は、文体的特徴を抽出する要素として「文字数を単位とする単語の長さ」「文の長さ」「品詞の分布」「書き手のクセでもある識別語」「文法的機能が高い接続詞や助動詞などの機能語」「n 個の連続文字を切り出した n-gram」「漢字・仮名の比率」「読点の打ち方」のほか、「文頭・文末のパターン」「段落の長さ」などを列挙している。このほか、先行研究が採用した指標なども加味して 27 の形態的要素を分析候補とした。以下に、大まかな要素特徴に分類する形で示す。

１．長さに関する要素

総文字数、延べ語数、異なり語数、延べ-異なり係数、平均文長、1 文あたりの延べ語数、1 文あたりの異なり語数

2．品詞に関する要素

助詞率、助動詞率、動詞率、名詞率、名詞-動詞の比率、代名詞率、接続詞率、相の類率、MVR(相の類と動詞の比率)

3．語種に関する要素

和語率、漢語率、外来語率

4．表記文字に関する要素

漢字率、ひらがな率、カタカナ率、英文字率、数字･記号率

5．語彙レベルに関する要素

上級語彙率、中級語彙率、初級語彙率

この 27 要素は出現数または比率で計量できる比例尺度であるため、3 種類のニュース報道文の間で統計的に意味のある有意差があるかどうかを確認できる。まず Levene 検定により 3 つのデータ同士の等分散性の有無を確認する。等分散性がある場合は一元配置分散分析で、等分散性がない場合は Welch 検定で 3 群データの平均値を比較する。ここで平均値に有意差が認められた場合は、多重比較により、どのデータ間に違いがあるのかを特定する。多重比較の方法も等分散性の有無で検定方法が異なり、一元配置分散分析による場合は Bonferroni 法、Welch 検定による場合は Games-Howell 法でおこなう（図 1 参照)。

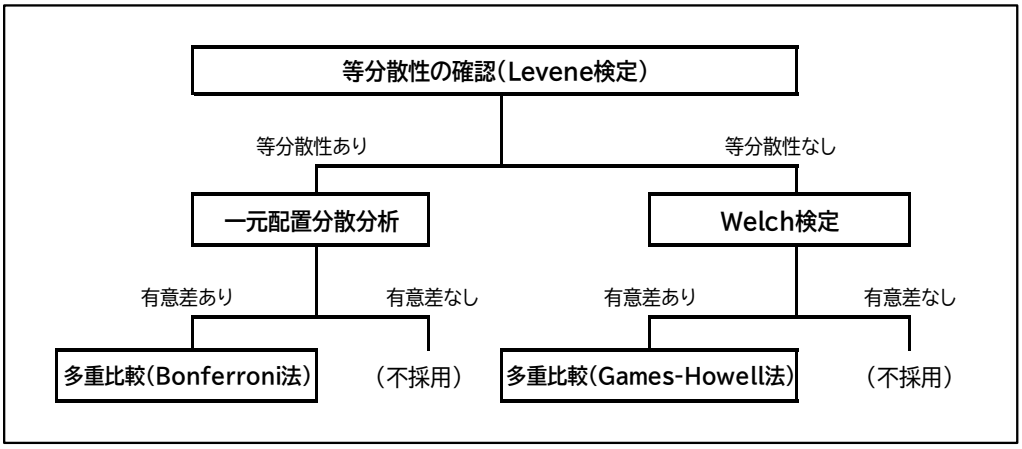

図１　統計分析の流れ

有意差の判断は 5%基準でおこなう。サンプル数は 74、74、78 でほぼ等しく、データ数

の偏りによる分析結果への影響は生じない。またデータはいずれも正規分布にあると想定する。統計処理は IBM 社の SPSS(ver.26)で実施した。

なお本研究では、形態素への分解、品詞分類、語種判定などは形態素解析エンジン『MeCab』、形態素解析辞書『UniDic 2.2.0』、『日本語教育語彙表』に基づいて実施する。実際の文字数や品詞数などの計測には、これらの形態素解析エンジンと辞書を組み込んだ日本語文章難易度判定システム『ｊ Readability』を、級別語彙数の計測には日本語読解学習支援システム『リーディング・チュウ太』『語彙チェッカー』を使った。

４.３　19 要素への集約

27 要素を多重比較した結果、3 種類のニュース報道文のうち、いずれかの 2 群以上で 5%水準の有意差が確認できたのは 19 要素だった。19 要素の統計分析結果を表 2 に示す。

表2　19要素に絞った統計分析結果

	等分散性	平均値比較	多重比較				
	Levene	分散分析／Welch 検定	Bonferroni／Games-Howell		[一般]／[やさしい]	[一般]／[小学生]	[やさしい]／[小学生]
総文字数	×	Welch		Games	◆	◆	
延べ語数	×	Welch		Games	◆	◆	
異なり語数	×	Welch		Games	◆	◆	◆
延べ･異なり係数	×	Welch		Games	◆	◆	◆
平均文長	×	Welch		Games	◆	◆	◆
助詞率	×	Welch		Games		◆	◆
助動詞率	○	分散分析	Bon		◆	◆	◆
動詞率	○	分散分析	Bon		◆	◆	◆
名詞率	○	分散分析	Bon		◆		◆
名詞･動詞の比率	×	Welch		Games	◆	◆	◆
接続詞率	×	Welch		Games		◆	◆
相の類率	○	分散分析	Bon		◆	◆	◆
和語率	×	Welch		Games	◆	◆	◆
漢語率	○	分散分析	Bon		◆	◆	◆
ひらがな率	×	Welch		Games	◆	◆	◆

漢字率	○	分散分析	Bon		◆		◆
上級語彙率	×	Welch		Games	◆		◆
中級語彙率	×	Welch		Games	◆		◆
初級語彙率	×	Welch		Games	◆	◆	◆

（○/× は等分散性の有無、 ◆ は 5%水準で有意差が確認できた箇所）

５．分析結果の評価

本研究では、有意差が確認できるのは文章難易度に差があるからだと考えている。つまり、有意差が確認できた 19 要素が文章難易度の違いに影響している可能性がある。そこで各要素について、統計結果から推定される影響を考察する。

５．１　長さに関する要素

「総文字数」「延べ語数」は 2 群で、「異なり語数」「延べ-異なり係数」「平均文長」は 3 群で、5%水準で有意差を確認できた。

「総文字数」

人間が短期記憶で保持できる情報量には限界があるため、初学者ほど語彙や文法に短期記憶が使われ、文章内容の理解に使う容量が減って読解が進まないとされる。総文字数は文章量を示す指標となり、理解過程にも影響を及ぼす可能性がある。統計分析では[一般]の文章だけが文字数が多い（長い）ことがわかった。[一般]には字数制限がないが、[やさしい]と[小学生]はある程度の字数制限をしていることが推測でき、これが文章難易度に影響していると考えられる。

「延べ語数」「異なり語数」

延べ語数は、文章中に同じ形態素が何回現れても出現総数を数えたものである。異なり語数は、同じ形態素が何回出てきても 1 と数える。一般に延べ語数はテキストの長さを、異なり語数は話題や表現の多様性を、捉える指標とされる。延べ語数の統計分析からは、[やさしい]と[小学生]には字数制限があると推測できるが、[一般]には制限がなく総語数が多くなる傾向が確認できた。異なり語数の統計分析結果は、[やさしい]は使われる語の種類が少いことから話題が集約されている可能性を、[小学生][一般]は語の種類が多くなり表現が多様化するか、扱う話題が拡大している可能性を示した。なお本研究では、形態素を語と読み替えている。

「延べ-異なり係数」

一般に文章中では、中心的に使われるいくつかの語の出現頻度が高く、補足的に使われる語の出現頻度は低くなる。この違いは延べ語数と異なり語数の違いで測れる可能性があるため、[延べ-異なり係数＝延べ語数／異なり語数]と定義した。樺島・寿岳(1965:62)は、延べ語数と異なり語数の関係は「濃縮度の大小を図る一つのものさし」になると論じ、延べ語数は減るが異なり語数が減らないケースは濃縮化であり、この場合は文章の意味は削られないと説明している。この濃縮化は、「延べ-異なり係数」が小さいケースに一致する。統計分析では、[小学生][やさしい][一般]の順で「延べ-異なり係数」が大きくなることを確認した。つまり、[小学生]では濃縮化が起きて扱う話題が絞り込まれている一方、[一般]は話題が多岐にわたって補足説明が多いと推察できる。

「平均文長」

文の長さが長いほど構文が複雑になりやすいために、読みにくく、認知処理に負荷がかかるようになると考えられる。阪本（1964:2）は、「文の長さを字数で計算した結果と、語数で計算した結果とは、統計上は完全に積極的に相関する」とする。そこで本研究では字数を採用し、[平均文長＝総文字数／文章中の文の数] と定義した。統計分析では、[やさしい][小学生][一般]の順で平均文長が大きくなり、修飾語が増えたり、文構造が複雑化していく可能性を示した。[やさしい]と[小学生]の標準偏差は小さくなっていることから、1 文を構成する文字数に目標値があることも推察できる。

5.2　品詞に関する要素

「助詞率」「名詞率」「接続詞率」は 2 群で、「助動詞率」「動詞率」「名詞-動詞の比率」「相の類率」は 3 群で、5%水準で有意差が確認できた。

「助詞率」

付属語である助詞は単独では使われず、活用もない。機能により格助詞、副助詞、係助詞、接続助詞、終助詞、準体助詞に分類できるが、いずれも、読み手が未知の語彙に遭遇した場合に文脈を推測する目印になる。[助詞率＝（助詞の数／延べ語数）×100]で算出した。統計分析では、[小学生]の助詞率だけが低かった。これは文字数制限の影響から補語・目的語の出現が少なく、文構造が単純化している影響と推測できる（「名詞率」の項参照）。

「助動詞率」

付属語である助動詞は単独では使えないが、活用があり、主に述語の一部として働く。ヴォイスやテンス、否定などのほか、モダリティとしても働く。特にテンスと否定は、ニュース報道文が伝える状況を把握するうえで重要な要素になる。[助動詞率＝（助動詞の数

／延べ語数）×100]で算出した。統計分析では、[一般][小学生][やさしい]の順で助動詞率が増えるのを確認した。主に動詞の後に続くことから、動詞率の影響を受けるとも考えられる。助動詞の出現率が高いのは、[やさしい]の特徴になる可能性がある。

「動詞率」

動詞は、ニュース報道文が伝えようとする事態を描き出す役割を持つ、文の中心的な要素である。[動詞率＝（動詞の数／延べ語数）×100]で算出した。動詞率の平均値は[一般]が 5.26、[やさしい]が 6.11、[小学生]が 4.54 で、統計的に意味のある差が確認できた。動詞は文の骨格要素であるため、文構造や文章難易度に関係する可能性がある。

「名詞率」

名詞は主語、目的語、述語になれることから、名詞の理解が、文章内容を把握するうえで重要になる。本研究では数詞は分析対象としていないため、[名詞率＝（(普通名詞の数＋固有名詞の数）／延べ語数｝×100]で算出する。統計分析では、[やさしい]の名詞率だけが低くなった。これは、補語や目的語になる名詞が少なく文構造が単純化していたり、話題が絞り込まれている可能性が考えられる。

「名詞-動詞の比率」

動詞 1 つ当たりの名詞の数を示している。主語、目的語、補語の多さを把握する手掛かりになることから、文構造の複雑さを測る指標になる。[名詞-動詞の比率＝名詞の出現数／動詞の出現数] で算出する。統計分析では、[やさしい][一般][小学生]の順で名詞-動詞の比率が増えるのが確認できる。[やさしい]は 1 つの動詞に対して出現する名詞数が少ないため、たとえば名詞や名詞句が並列するような文構造が少ないとも考えられる。つまり[やさしい]の文構造は単純化している可能性がある。

「接続詞率」

接続詞は文や段落の冒頭に現れて前後のつながりを示し、文章理解の目印になるとされる。[接続詞率＝（接続詞の数／延べ語数）×100]で算出する。統計分析では、[小学生]の文章の接続詞率だけが低い。掲載スペースが限られるため必要性の低い接続詞を優先的に落としたためか、文章の論理構成が単純化しており接続詞を示す必要がなかった可能性がある。[やさしい]はばらつきを示す標準偏差が大きく、書き換え時に「接続詞」の使用については考慮されていないと考えられる。

「相の類率」

『分類語彙表』は形容詞類を「相の類」として区分けしており、形状詞（学校文法でい

う「形容動詞」の語幹)、副詞、形容詞、連体詞が含まれる。樺島・寿岳(1965:30)は、相の類は様態を表す語で、読み手の大まかな理解を促進させる働きが想定できるとしている。[相の類率＝（相の類（形状詞、副詞、形容詞、連体詞）の数／延べ語数）×100]で算出する。統計分析では、[やさしい]が最も高くなったことから、 [やさしい]は情報の概要を示して、大まかな理解を優先させる傾向があると推測できる。

5.3　語種に関する要素

「和語率」「漢語率」は、3 群すべての組み合わせで 5%水準で有意差が確認できた。

「和語率」「漢語率」

日本語の大半は和語と漢語で、一方が増えると他方が減る反相関の関係がある。国立国語研究所(1972)は、同じ意味内容を表す動詞を和語と漢語で比較し、その語が表現する意味特徴の違いから、漢語の方が大規模／公的／抽象的な現象を示すことが多いとしている。つまり、意味内容の抽象度という点で文章の難易度に影響すると考えられる。[（和語/漢語）率＝（(和語/漢語）の数／延べ語数）×100] で算出する。統計分析では、和語率は[やさしい]が最も高く、漢語の使用を避ける傾向がある。また、アナウンサーが音声で伝達する情報でもある[一般]は、[小学生]よりも和語を使う傾向が強く現れている。

5.4　文字種に関する要素

「ひらがな率」は 3 群すべての組み合わせで 5%有意となった。漢字率は[やさしい]がからむ 2 つの組み合わせで 5%水準で有意差があった。

「漢字率」「ひらがな率」

本研究では文字種のうち漢字、ひらがなに着目する。日本語母語話者の学校教育には学年配当漢字表があり、学年が進むほど教科書の「ひらがな率」が減り「漢字率」が増えるため、文章難易度と語種の出現率に相関があると推測できる。[(漢字/ひらがな)率＝((漢字/ひらがな）数／総文字数）×100] で算出する。統計分析では、[小学生][一般][やさしい]の順でひらがな率が増え、漢字率は[やさしい]だけが低いことを示した。[やさしい]は、使用できる漢字数を[一般]や[小学生]よりも少なく制限している影響と考えられる。

5.5　語彙レベルに関する要素

「上級語彙率」「中級語彙率」は 2 群以上で、「初級語彙率」は 3 群で、5%水準で有意差を確認できた。

「級別（上級／中級／初級）語彙率」

知らない語彙が多いと、その文章の内容を理解できない。川村・北村（2023:18）は語彙

の難易度が、文章の難易度に影響すると推定している。旧日本語能力試験が公表していた級別に習得すべき語彙リストを参考に「級外」「旧 1 級」を「上級語彙」に、「旧 2 級」を「中級語彙」に、「旧 3 級」「旧 4 級」を「初級語彙」と分類し、[級別語彙率＝（当該級の延べ語彙数／総語彙数）×100]で算出した。

統計分析では、上級語彙と中級語彙の出現率は[やさしい]だけが低くなった。初級語彙率は[小学生][一般][やさしい]の順で増えている。[やさしい]が、使用する語彙を旧 3 級以下に制限している影響と考えられる。[一般]と[小学生]には使用語彙の制限はないと考えられるが、[小学生]よりも[一般]の方に初級語彙が多く現れるのは、アナウンサーが音声で伝える文章であることが関係していると推察する（「和語率」の項参照）。

6． 2 成分への集約

この 19 要素に、どのような重み付けをすれば 3 種類のニュース報道文の違いを示せるかを検討するため主成分分析をおこなう。主成分分析は、多数の量的な説明変数を、より少ない指標や合成変数（複数の変数が合体したもの）に集約する手法である。つまり、ニュース報道文の難易度の違いを 2～3 の要因にまとめ直すことができる。

表３　主成分分析の適正検定

KMO および Bartlett の検定		
Kaiser-Meyer-Olkin の標本妥当性の測度		0.764
Bartlett の球面性検定	近似カイ 2 乗	6881.840
	自由度	171
	有意確率	0.000

19 要素で構成したデータの適正性を調べた結果、KMO の測度が 0.764 で、基準となる 0.5 を上回っている。Bartlett 検定による有意確率は p <.05 で、基準となる 5%水準での有意を満たしている（表 3 参照）。これは、19 要素を基にした主成分分析が統計的に有効であることを意味している。

そこで、実際に主成分分析を実行してスクリープロットを確認すると、19 要素は 3 つの成分にグループ分けできることがわかる（図 2 参照）。ただ 3 成分に集約すると、影響力の小さい第 3 成分の構成要素が 1 つだけになってしまう。このため、本研究では 2 成分に集約させることにした。2 成分に集約しても、負荷量平方和の累積寄与率は 58.77 で、全体データの

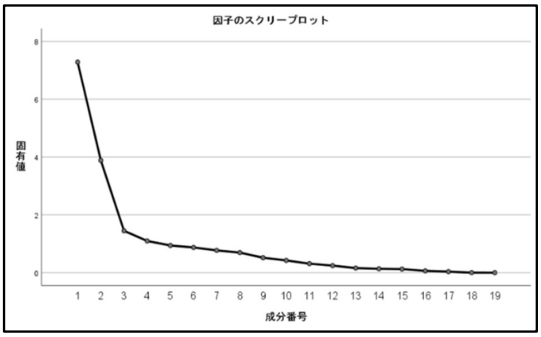

図２　19 要素によるスクリープロット

6 割程度を説明できるためである。

2 成分にグループ分けした際の成分行列を表 4 に示す。主成分分析では各要素の因子負荷量は 0.4 を超えるものを採用するため、第 1 成分は「和語率」「初級語彙率」「ひらがな率」「上級語彙率」「名詞率」「漢語率」「中級語彙率」「名詞－動詞の比率」「動詞率」「漢字率」「助詞率」「相の類率」「助動詞率」の 13 要素、第 2 成分は「総文字数」「延べ語数」「延べ-異なり係数」「異なり語数」「平均文長」の 5 要素に明確に分けられる。

「接続詞率」はどちらの成分でも因子負荷量が 0.4 を下回った。因子負荷量が 0.4 に満たない要素は統計上の意味がないため、構成要素から除外する。ここまでの結果、統計的には、⑴「接続詞率」を除く 18 要素が文章の難易度を判定する要素として意味があること、⑵この 18 要素は 2 つの成分に分類できること、を確認した。

表４　２成分の因子負荷量

	成分	
	1	2
和語	0.907	0.189
初級語彙	0.884	-0.161
ひらがな	0.875	0.176
上級語彙	-0.832	0.080
名詞	-0.814	0.016
漢語	-0.787	-0.016
中級語彙	-0.722	0.220
名詞/動詞	-0.676	-0.200
動詞	0.673	0.143
漢字	-0.648	0.162
助詞	0.641	0.354
相の類	0.559	0.161
助動詞	0.415	-0.210
接続詞	0.237	0.190
総文字数	-0.064	0.942
延べ語数	-0.065	0.942
延/異係数	0.052	0.837
異なり語数	-0.100	0.761
平均文長	-0.337	0.603
因子抽出法：主成分分析		

ところで、表 4 に示す各要素の因子負荷量は、最終的に求めたい変数（文章難易度）に与える各要素の影響の度合いである。－1～1 の範囲の値をとり、絶対値が大きいほど影響が強いと考える。また主成分分析は、各要素間に何らかの関連性があると判断されるときに、同じ成分として分類される。そこで本研究で分離した「第 1 成分」と「第 2 成分」には、それぞれどのような意味を示すのかを検討する。

６.１　第１成分の検討

第 1 成分は語彙や文法に関する要素で、学習者の理解しやすさに関わっていると判断した。各要素の因子負荷量はプラスとマイナスに分かれるが、全体としては数値が大きい要素ほどやさしく、数値が小さい要素ほど難しいと説明することができる。そこで以下の考察は、負荷量がプラスかマイナスかに分けておこなった。

プラスの要素

「和語率」「初級語彙率」「ひらがな率」「動詞率」「助詞率」「相の類率」「助動詞率」である。いずれも日本語学習の比較的早い段階から現れる要素と考えられる。負荷量が 0.9 前後と大きめに出た「和語」「初級語彙」「ひらがな」は、学習者がやさしいと感じる要素

と考えられる。「動詞」は活用があったり、「する」の形で現れるサ変動詞のように学習者が難しく感じる要素があることから、負荷量がやや低くになったと考える。また「助詞」は、同じ形をしていても意味機能は多岐にわたるケースが多い。たとえば日本語教科書で格助詞を学習する場合、代表的な使い方は初級段階で習得するが、補足的な使い方は上級になってから学習するケースがある。このように、上級になっても新しい知識を学習する必要があることから難しさがあるとして、因子負荷量がより小さくなったと推察できる。

また「相の類」と「助動詞」の因子負荷量は、「助詞」よりもさらに小さいことから、文章難易度に影響は与えるが、その度合いは低くなると考えられる。

マイナスの要素

マイナスの要素は「上級語彙率」「名詞率」「漢語率」「中級語彙率」「名詞-動詞の比率」「漢字率」である。これらは統計的に、文章を難しくする要素と判断されたことになる。このマイナス要素を使わずに文章を構成することはできないが、必要以上に使いすぎると文章難易度を高めてしまうと考えられる。

このうち重み付けが比較的大きく出た「中級語彙」「上級語彙」は、日本語学習過程の中級以降に現れる。学習が進んでから登場するために、学習者にとって理解しにくい難しい項目と判定されたと推察する。

「名詞」も重み付けが大きかった。これは、⑴ニュース報道文では固有名詞が多い傾向があり、その多くが漢字で表記されていること、⑵一般に「動詞」よりも「名詞」の方が難しいとされること、が原因ではないかと推測する。⑴の固有名詞は知っているかどうかが重要で、表記された形から内容を推測することが難しい。また⑵については、たとえば動詞「歩く」は N5 相当に対して名詞「歩み」は N1 相当と判定されたり、動詞「輝く」は N2、N3 相当なのに対し名詞「輝き」は級外の超上級相当と判定されている。名詞化することで語の意味概念が拡大し、抽象的な概念までも含んでくることが関係すると考える。JF スタンダードのレベル表でも、抽象的な概念は学習がかなり進んだ段階での習得目標に設定されており、本研究の主張と矛盾しない。

これに関連する「名詞-動詞の比率」は単純な出現率ではないが、「名詞」の出現頻度の影響を受けやすいことからマイナス要素に分類されたと考えられる。

重み付けはやや下がるが、「漢語」と「漢字」は、特に非漢字文化圏の日本語学習者にとっては日本語学習を難しくする要素の 1 つに挙げられている。漢語には 1 字漢語も含まれるが、漢字が複数組み合わさった熟語の形で現れる方が多いため、「漢字」よりも「漢語」

の方が難しさへの影響が強いと判断されたと推察する。

以上のように第 1 成分はプラスとマイナスの成分に分かれているが、それぞれの要素は語彙や品詞に関連していることから「語彙・文法要因」と名付ける。ここに示される要素は、主に日本語学習の段階に応じて登場する時期の違いが重み付けの差として現れ、文章難易度の判断材料になっていると考えられる。マイナス側が難しい要素であることから、第 1 成分の算出結果が大きいほど文章はやさしくなると考えられる。

6.2　第 2 成分の検討

第 2 成分は文や文章の長さに関する要素であり、指標がすべてプラスの値をとるため、文章や文が長くなるほど算出結果が大きくなる。そこで「長さ要因」と名付ける。第 1 成分とは逆に、数値が大きいほど文章は難しくなる傾向を示すと考えられる。

6.3　2 つの成分の相互影響

「語彙・文法要因」が「長さ要因」から受ける影響、逆に「長さ要因」が「語彙・文法要因」から受ける影響を確認する。表 4 では、該当する部分の負荷因子量がかなり低いことがわかる。これは 2 つの要因は完全に独立ではないが、2 要因間の相関は非常に小さいことを示している。つまり、文章を形態的面からとらえた場合、やさしい文章であると判断するためには「語彙・文法要因」と「長さ要因」を共にやさしくなるように調整する必要があると考える。たとえば短く書かれた文章であっても、難解な語彙ばかりを使って書かれたものは読み手にとって難しいと判断するべきだという考えである。

一方、先行研究の多くは、文法関連の要素と長さ関連の要素を同列に扱っている。このため、たとえば難解な語彙ばかりを使っていても文章が短い場合には語彙による影響が薄められて、全体としては難しいと判定されない可能性が残ってしまう。そこで本研究では主成分分析の結果を根拠に、文章を構成する形態的要素の組み合わせから見た文章難易度には、「語彙・文法要因」と「長さ要因」の 2 つの要因の影響を別々に考慮する必要があると主張する。

6.4　考察

この 2 つの成分から算出した値で表したデータの分布状況を、ニュース報道文別に示す（図 3 参照）。横軸には第 1 成分である「語彙・文法要因」、縦軸には第 2 成分である「長さ要因」を設定する。座標軸の数字はデータを標準化したうえで算出した各成分の値である。3 つの分布状況の違いを直接比較できるよう、散布図の座標軸スケールは同じにした。

6.1、6.2 で述べた通り、データが左上に分布するほど文章は難しく、右下になるほど文章はやさしくなると解釈できる。

横軸の「語彙・文法要因」は[一般]と[小学生]は、ほぼ同じ範囲に分布している。これに比べて[やさしい]だけはプラス側、つまり文章難易度が低い（やさしい）側に集中して分布している。一方、縦軸の「長さ要因」は、[一般]だけが広い範囲に分布しているのが確認できる。[やさしい]と[小学生]はどちらも座標軸周辺に分布している。ただ[小学生]には多少のばらつきが確認でき、これは扱う情報によって文章量を加減している可能性を示している。

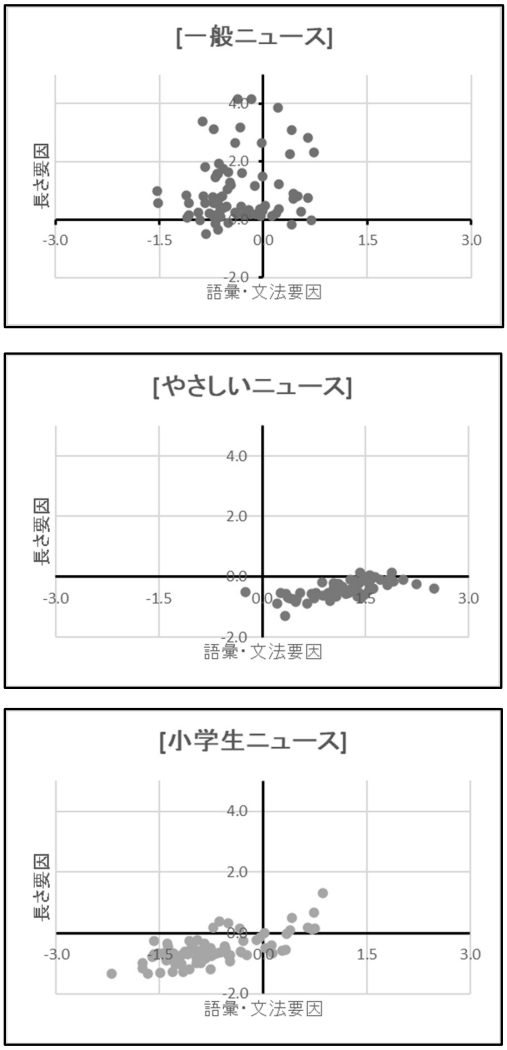

図３　データの散布図

（上から[一般][やさしい][小学生]）

ここから、[一般]を、わかりやすく書き換えるには、まず文章量を少なく抑える必要があることが、[やさしい]と[小学生]の「長さ要因」の分布から読み取れる。そのうえで「語彙・文法要因」の分布が[やさしい]はプラス側（さやしい側）に、[小学生]はマイナス側（難しい側）を含む範囲に分布するのは、書き換える際に「やさしい日本語」を使うかどうかの違いが影響していると考える。つまり、形態的な側面から文章のやさしさを求めるのであれば、「やさしい日本語」を使った書き換えは効果が大きいことが確認できる。

７．書き換えルールとの比較検討

ここまで 3 種類のニュース報道文を使い、文章難易度に影響する 18 の要素とその重み付けを一般化した。そこで、この 18 要素と、形態的な側面から最もやさしいと判断された[やさしい]の書き換えルールとの関連性を検討する。ルールと一致する要素、一致しない要素を分類することで、文章をやさしくするには「やさしい日本語」への書き換えルール

以外にも条件があることを示すことができる。

まず、「やさしい日本語」への書き換えルールを確認する。佐藤(2004)は災害発生時のような緊急事態の際に使う「やさしい日本語」を想定し、書き換えルールを示している。だが本研究で扱うニュース報道文は緊急時ではなく、平時に使われる日本語文章である。そこで、田中ら（2018:88-89）が示すルールを採用する。これは実際に[一般]から[やさしい]への書き換えの際に使われているもので、以下の 3 つが示されている。

① 語彙は初級（旧 3 級、旧 4 級）の範囲を使う

② 文法面では「1 文は 50 字以下」「受動態は能動態に書き換える」「ニュース特有表現は使わない」の 3 点

③ 重複した内容は削除する

7.1 明確に該当する要素

本研究で特定した 18 要素のうち、明確に書き換えルールと一致しているものを示す。

- ①の語彙の使用は、「初級語彙率」「中級語彙率」「上級語彙率」の各要素に相当する。「中級語彙」「上級語彙」の使用を避け、「初級語彙」を使用するよう指示していることが確認できる。
- ②の文法制限うちの「1 文は 50 字以下」としたルールは長さに関する制限で、これは 18 要素のうちの「平均文長」に該当する。
- ③の重複内容の削除は、同じ語の使い回し具合から、話題の絞り込みを測る指数である「延べ-異なり係数」が該当する。ただ統計分析では、[やさしい]が明確に小さいわけではなく、話題の絞り込みに成功していない。これに対し、[小学生]は同じ語の使い回しが少なく話題が絞り込まれている一方、[一般]は話題が多岐にわたり補足説明が多いと推察できる（表 5 参照）。

表５　延べ-異なり係数の統計量

延-異係数	度数	平均値	標準偏差
[一般]	74	3.06	0.76
[やさしい]	74	2.35	0.23
[小学生]	78	2.03	0.34

7.2 該当すると推測できる要素

ルールとして明文化されていないが、ある程度の制限を推測できる要素がある。

- 「初級語彙」は「漢字」より「ひらがな」の割合、「漢語」より「和語」の割合が高いこと、「上級語彙」ほど「漢字」や「漢語」の割合が高いことが経験的に知られている。つまり、文字種と語種に分類される「漢字」「ひらがな」「漢語」「和語」の 4 要素は、語彙の使用制限をすることで、間接的に制限がかかっていると解釈できる。

- [やさしい]の書き換えルールに字数や語数の制限は明記されていないが、[やさしい]のデータを解析したところ、データのばらつきを示す標準偏差が小さいことから、「総文字数」は 310 字程度、「延べ語数」は 200 語程度、「異なり語数」は 85 語程度の付近にデータが集中しているといえる（表 6 参照）。ここから、「総文字数」「延べ語数」「異なり語数」の要素には目標値があると推察できる。

表6　[やさしい]の統計量

要素	度数	平均値	標準偏差
総文字数	74	312.31	25.43
延べ語数	74	200.27	20.87
異なり語数	74	85.51	7.30

7.3 検討結果考察

文章難易度に影響があると考えられる 18 要素のうち、12 要素については書き換えルールで何らかの制限があることを確認した。別の言い方をすれば、「語彙・文法要因」を構成する 6 要素（「名詞率」「名詞-動詞の比率」「動詞率」「助詞率」「相の類率」「助動詞率」）はルール化されていない。

もちろん文章を書く際には、どの品詞を、どの程度使うかをあらかじめ決めておくことはできない。特に付属語である「助詞」「助動詞」は文法的な役割はあるが、自立語など他の要素と結びついて使うため、助詞や助動詞の出現頻度を調整することは難しいと考えられる。そこで、書き換えの際に使うかどかを考慮できる「名詞」「動詞」「相の類」と、ここから算出される「名詞-動詞の比率」の 4 要素の出現頻度から、書き換えの際にどのような調整がおこなわれた可能性があるかを考察した。この調整がおこなわれた結果として品詞の出現頻度に差が生じ、これが文章難易度の違いにつながったと考えるためである。

調整の可能性として推察できるのは 2 点ある。1 つは、[やさしい]は 1 つの動詞に対する名詞数が少ないことから、目的語や補語になる名詞や名詞修飾節を減らして、文構造を単純化していることである（表 7 参照）。文構造の単純化は平均文長を抑えることにもつながり、文章難易度を下げる要因になる。

表7　名詞-動詞の統計量

名詞／動詞	度数	平均値	標準偏差
[一般]	74	6.32	1.93
[やさしい]	74	4.82	3.22
[小学生]	78	8.05	3.65

もう 1 つは[やさしい]の「相の類」の出現率が高い点である。主成分分析の結果では、「相の類率」は因子負荷量が低いため文章難易度に与える影響は少ないが、前述したように [やさしい]は情報の大まかな理解を優先させるような文章構成をとっていると推察できることである（表 8 参照）。

表8　相の類の統計量

相の類率	度数	平均値	標準偏差
[一般]	74	2.80	1.21
[やさしい]	74	3.67	1.77
[小学生]	78	2.11	1.44

これら 2 点は、いずれも文や文章の構造に結び

つくという共通項がある。だが形態面からの分析だけでは、これ以上の考察を進めるのは難しい。書き換えの際に具体的にどのような調整がおこなわれて文や文章を組み立てたのかを探るは、想定する読者水準がほぼ等しい[やさしい]と[小学生]で同じ内容のニュース同士を比べて表現の違いを調べるなど、文章を内容面から分析する必要があると考える。

8．結論

本研究では、ニュース報道文を対象に、やさしい日本語を使った場合に文章難易度がどう変化するのか、変化する原因は何かについて、文章を構成する要素の出現頻度から検討するリーダビリティ研究として分析した。その結果、以下の 3 点を明らかにした。

1．文章難易度は、「語彙・文法要因」と「長さ要因」の 2 つの成分で判定でき、それぞれの影響を個別に考慮する必要があることを統計的に明らかにした。

2．「語彙・文法要因」は「助詞率」「助動詞率」「動詞率」「名詞率」「名詞-動詞の比率」「相の類率」「和語率」「漢語率」「漢字率」「ひらがな率」「上級語彙率」「中級語彙率」「初級語彙率」の 13 要素、「長さ要因」は「総文字数」「延べ語数」「異なり語数」「延べ-異なり係数」「平均文長」の 5 要素で構成され、それぞれの要素の重み付けの程度も割り出した。この重み付けは、文章難易度を測るリーダビリティ算出式を完成させる際の基礎的要素になる。

3．「やさしい日本語」への書き換えルールには、日本語文章の形態をやさしくする効果があることを確認した。さらに書き換えルールには示されないが、文章難易度に影響する要素として、「文や文章構造の単純化」と「相の類の使用による概要把握」の 2 つがあることを特定した。

文章難易度に影響がある形態的要素のうち、やさしい要素に分類したものを増やした場合、リーダビリティの観点から文章の難易度は下がる。だが、これだけで日本語文章がわかりやすくなるとは考えていない。たとえば、ひらがなが連続した文章はリーダビリティ判定ではやさしくなるだろうが、読みにくく、内容を理解しにくくなると想像できる。このケースでは、[語彙・文法要因]を構成するプラス要素とマイナス要素に適正な分布範囲がある可能性がある。さらに結論で示したように、文構造の在り方などの要因が文章難易度に影響する可能性も考慮すると、「やさしい日本語」への書き換えルールに示すような形態的要素の使用条件を一律に制限するだけでは、日本語文章は完全にはわかりやすくなら

ないと考える。ただ本研究は、文章表面に現れる形態的な側面からの分析に特化したため、文章が伝達できる情報量や程度といった内容的な側面からの文章の難易度分析や、読み手による認知的な難易度判断については扱えていない。いずれも今後の課題としたい。

■参考文献

浅井卓真(2017)「テキストの特徴を計量する指標の概観」『日本図書館情報学会誌』63(3), 159-169.

樺島忠夫・寿岳章子(1965)『文体の科学』綜芸舎

川村よし子・北村達也(2013)「日本語学習者のための文章難易度判定システムの構築と運用実験」『scientific journal』14, 18-30.

阪本一郎(1964)「文の長さの比重の査定法－Readability の研究の試み－」『読書科学』8(1), 2-6.

阪本一郎(1971)「読みやすさの基準の一試案」『読書科学』14(1,2), 1-6.

佐藤和之(2004)「災害時の言語表現を考える:やさしい日本語・言語研究者たちの災害研究」『日本語学』23(10), 34-45.

柴崎秀子・原信一郎(2010)「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), 215-232.

田中英輝、熊野正、後藤功雄、美野秀弥(2018)「やさしい日本語ニュースの制作支援システム」『自然言語処理』25(1), 81-117.

森岡健二(1952)「『読みやすさ』の基礎的研究」『昭和 26 年度国立国語研究所年報』 91-108.

山村毅(2014)「複数の判断基準を用いた日本語文章の難易度判定」『電子情報通信学会論文誌 D』J-97(5), 1063-1066.

李在鎬(2016)「日本語教育のための文章難易度に関する研究」『早稲田日本語教育学』21, 1-16.

■参考図書

計量国語学会編(2009)『計量国語学事典』朝倉書店

国立国語研究所(1972)『動詞の意味・用法の記述的研究』秀英出版

■参考ホームページ

『NHK NEWS WEB』日本放送協会 https://www3.nhk.or.jp/news/

『NEWS WEB EASY』日本放送協会 https://www3.nhk.or.jp/news/easy/

『毎日小学生新聞』毎日新聞東京本社 https://mainichi.jp/maisho/

『リーディング・チュウ太』https://chuta.cegloc.tsukuba.ac.jp/

『j Readability』http://jreadability.net