研究論文

日本語教育のための文章難易度に関する研究

李　在鎬

要　旨

日本語教育の読解クラスを支援する目的で文章の難易度を判定する「日本語教育のリーダビリティ公式（X={平均文長*-0.056}+{漢語率*-0.126}+{和語率*-0.042}+{動詞率*-0.145}+{助詞率*-0.044}+11.724）」を作成した。本公式の予測精度を示す R^2 値は「0. 896」であり、高精度で文章の難易度を予測しうることが明らかになった。さらに、公式の妥当性を検証するため、旧日本語能力試験で出題された 25 年分の読解テキストに対する難易度判定を行ったところ、1 級から 4 級までの読解テキストのリーダビリティ値に有意な差が確認できた（$F(3,168)=141.035$, $p<.001$）。

リーダビリティ公式を利用した応用研究として「現代日本語書き言葉均衡コーパス」および「読売新聞記事」から無作為に 1,949 個のテキストサンプルを抽出し、文章の難易度に関する大規模な調査を行った。調査の結果、ウェブ系テキストの代表例である「Yahoo!知恵袋コーパス」は初級後半から中級後半レベルに相当、書き言葉の代表例である「書籍コーパス」は中級前半から上級前半レベルに相当、公的文章の代表例である「白書コーパス」は上級前半から超級レベルに相当、メディア系テキストの代表例である「新聞記事」は上級前半レベルに相当することが明らかになった。

キーワード

リーダビリティ　難易度　重回帰分析　やさしい日本語　日本語コーパス

1. 背景と目的

本研究は、データ科学の手法を用いて、文章が持つ難易度の問題を定式化する。本研究の位置づけおよび研究の意義としては、次の 2 つが考えられる。1 つ目は、読解クラスの教室支援および教材開発への貢献を目指す研究であること、2 つ目は、「やさしい日本語」における文章評価への貢献を目指す研究であることが挙げられる。

まず、1 つ目の位置づけに関しては、次のような問題意識がある。これまでの日本語教

育分野では、語彙や文法項目の難易度に関しては、（適切かどうかはさておき）旧日本語能力試験の「出題基準」のような資料が存在する一方、文章の難易度に関しては、共通認識となる資料および見方は存在しない。しかし、学習効果という観点から考えた場合、教師は学習者の理解度に応じて、教材の難易度を統制する必要がある。従って、文章の難易度に関する共通理解の不在問題は、日本語教育全体における課題と言わざるを得ない。こうした問題の根本原因として、語彙や文法項目に比べ、文章という単位は、情報量が多く、一貫した分析が難しいということが考えられる。本研究では、こうした問題を解決するために、自然言語処理で利用されるデータ解析の手法を用いて、難易度の公式化を行う。

次に、2 つ目の位置づけに関しては、次のような問題意識がある。近年、日本語教育の社会的役割を議論する場面において「やさしい日本語」の存在が強調されてきている。この「やさしい日本語」をめぐっては、庵（他）（編）（2012）で全体像が示されてから、NHK の「NEWS WEB EASY」（http://www3.nhk.or.jp/news/easy/（2016.7.27.閲覧））によるニュース配信、さらには、岩田（2016）のような実用書が刊行されるなど、様々な形で研究成果が活用されている。こうした動きの背景には、（日本が今後向かっていくと予想される）多文化共生社会におけるユニバーサルな日本語コミュニケーションの実現という課題にとって、「やさしい日本語」が必要不可欠だという認識がある。

このように「やさしい日本語」の必要性は認識されてきている一方、いざ「やさしい日本語」を書く（または話す）ということにおいては、様々な課題が指摘されている。例えば、田中（他）（2012）や岩田（2014）などでは、旧日本語能力試験の「出題基準」を利用し、「やさしい日本語」を実現しようと試みている。具体的には、3 級や 4 級の語彙や文法項目を使用することを一つの基準として提案しているが、これには 2 つの問題が考えられる。1 つに、テスト作成のために作られた「出題基準」のような項目表を、「やさしい日本語」のような到達目標が定まっていないコンテンツにおいて使用することの適切性の問題、2 つに、文章の構成要素を変えることで、文章全体がやさしくなるということが担保されないという問題である。本研究の立場としては、易しい語彙や文法項目で書いた結果、文章全体が易しくなっているかということが確認できてこそ、「やさしい日本語」が実現されると考える。以上の問題を解決するために、本研究では、計量文体論やリーダビリティの観点から文章を捉え、システム化を行う。

以上の問題意識から、次の手順で考察を行う。まず、２節では、文章難易度の研究であるリーダビリティ（readability）と計量文体論の問題意識について、具体的なアプローチや分析例を示す。これを踏まえ、3 節では、「日本語教材による、日本語教育のための、日本語教育のリーダビリティ公式作成」の試みについて述べる。次に、４節では、リーダビリティ公式を使った応用研究として、李・長谷部・久保（2016）が行ったコーパスデータに対する大規模な難易度調査について紹介する。そして、５節では、コーパス研究とリーダビリティ研究の関連性について述べる。最後に、６節では、文章の難易度を測ることの意味について改めて本研究の立場を確認する。

2. リーダビリティ研究とは

自然言語の文章が持つ潜在的な難しさを測定する研究領域として、リーダビリティ研究がある。リーダビリティとは、文章の読みやすさのことであり、リーダビリティ研究では、一文あたりの文字数や語数といった表層的情報をもとに、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研究は、Flesch（1948）や Smith&Kinkaid（1970）など、古くからの先行研究があり、リーダビリティを計算する目的で様々な計算式が提案されてきている。日本語においても、建石（他）（1988）、佐藤（2011）、酒井（2011）、柴崎・原（2010）、Hasebe&Lee（2015）、Lee&Hasebe（2016 forthcoming）などの研究があり、佐藤（2011）、柴崎・原（2010）、Hasebe&Lee（2015）においては、ウェブサービスとしてリーダビリティの計算式を提供しており、研究成果の共有がなされている[1]。

さて、リーダビリティ研究そのものは、20 世紀半ばにアメリカで盛んに研究され、世界に広まった研究枠組みであるが、元々の分析枠組みとしては、計量文体論の流れをくむものである。

計量文体論とは、簡単に言えば、文章の特徴を数量的に考察しようとする学問である（陳 2012）。計量文体論において、「文体」は「文章上の個人的な体臭、あるいは個人的な習性を意味するもの」（前川 1995）と捉えられており、代表例としては、文の長さに関する調査例がある。

表 1　作家の文の長さ

作品	作家	文の長さの平均
吾輩は猫である	夏目漱石	29.8
坊っちゃん	夏目漱石	30.9
城の崎にて	志賀直哉	28.8
暗夜行路	志賀直哉	25.6
細雪	谷崎潤一郎	170.1
雪国	川端康成	55.5
伊豆の踊子	川端康成	30.2
楼蘭	井上靖	47.8
斜陽	太宰治	71.4
人間失格	太宰治	48.7
万延元年のフットボール	大江健三郎	43.0
羊をめぐる冒険	村上春樹	36.7
五分後の世界	村上龍	46.2
うたかた	吉本ばなな	41.2
キッチン	吉本ばなな	35.7

表 1 は、前川（1995）による調査で、15 編の小説における文の長さを示している。この場合の長さとは句点から句点までの文字数のことであり、表 1 は、その平均値である。一般に、谷崎潤一郎は文が長く、志賀直哉は文が短いと言われているが、表 1 からもこのことが確認できる。ところで、こうした 1 文の長さを測ることに、どんな意味があるのだろうか。この問に答えるためには、計量文体論の理論的前提について確認する必要がある。

計量文体論では、次のことを前提にしている。文体とは、書き手による文章の指紋のようなものであり、データ科学の方法を用いることで客観的に捉えることができる、ということである。データ科学の方法の具体例としては、表 1 に示した文長などが代表的な事例である。そのほかに品詞の分布や語の長さ、語彙の特性値、n-gram[2]、語種の分布、句読点の頻度や位置なども有効な指標とされている（計量国語学会 2010）。こうした指標を使った分析を通して、計量文体論では著者の推測や執筆時期の推測を行うなどの研究が行われてきた。リーダビリティ研究においても、こうした計量文体論で用いられている分析指標を使って、文章の難易度を推定している。

それでは、リーダビリティ研究における具体的な問題意識について確認しておく。リーダビリティ研究では、次の 3 つの研究課題が盛んに議論されてきた。1）文章の難易度を決定する要因は何か、2）文章の難易度を決定する複数の要因をどのように重み付けし、公式化するか、3）どのような難易度のスケールを使うかである。

まず、1）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、どのような話題かという問題や文章としてのまとまり具合などが考えられる。ミクロな要素としては、前述の田中（他）（2012）や岩田（2012）が指摘する語彙の難しさ、文法構造の難しさ、さらには、計量文体論で問題視されてきた語の長さ、文の長さなどが考えられる。

次に、2）の問題として、文章の難易度を決める要素が複数であることが明らかになった場合、個々の要素が持つ強さの度合いをどう表現するのかということが考えられる。つまり、文章に含まれる語彙の難しさの要因、文法項目の難しさの要因、文長などの長さの要因などを同等に扱ってよいかという問題に帰結する。当然ながら、これらの要因の強さは異なるものであり、その異なり具合は指標の重みとして明らかにする必要がある。

最後に、3）として、文章の難しさを表現するスケールをどう設定するかの問題が考えられる。日本語教育の文脈で言えば、日本語能力試験の 1 級から 4 級、または N1 から N5 が代表的な難易度のスケールになるであろう。国語教育の文脈で言えば、小 1〜高 3 までの学年が代表的なスケールになるであろう。

上述の 3 つの問題は、内省で明らかにできるものでもなければ、個々の事例をもとに短編的な考察を行ったところで、明らかになる問題でもない。こうした理由からリーダビリティ研究では、大規模なデータ（基準コーパス）を用いて、それを計算論的な手法で分析し、公式化するというアプローチが採用されている。具体的な研究例として 3 つの研究をとりあげる。まず、柴崎・原（2010）は小学校 1 年から高校 3 年を難易度のスケールとして設定し、重回帰分析によるリーダビリティ公式を提案している。難易度を決定する要因としては、①文章中の平仮名の割合、②1 文の平均述語数、③1 文の平均文字数、④文の平均文節数の 4 つの要素をとりあげている。次に、佐藤（2011）では小学校から大学まで

の全教科で使用されるテキストを用いて、bigram という文字の連続をもとに言語処理の方法で難易度を予測している。難易度スケールとしては、9 段階のもの（とてもやさしい、やさしい、かなりやさしめ、ややややさしめ、ふつう、ややむずかしめ、かなりむずかしめ、むずかしい、とてもむずかしい）を設定している。最後に、Hasebe&Lee（2015）は、日本語教科書を用いて、重回帰分析によるリーダビリティ公式を提案している。難易度を決定する要因としては、①平均文長、②漢語率、③和語率、④動詞率、⑤助詞率の 5 つの要素をとりあげている。難易度のスケールとしては、6 段階のもの（初級前半、初級後半、中級前半、中級後半、上級前半、上級後半）を設定している。

最後に、リーダビリティ公式と公式を作成する際に使用する「基準コーパス」について述べる。リーダビリティ研究では、分析に使用する「基準コーパス」によって得られる公式が決まるため、どのようなテキストをどれだけ用いるかが研究の要になる。柴崎・原（2010）の場合、読解教育に役立つリーダビリティシステム構築を目標にしていたため、国語の教科書を使用している。佐藤（2011）は、「平易な日本語表現への工学的アプローチ」という科研研究として行われたもので、汎用性の高い解析システムを作ることを目標にしていたため、日本国内の公教育で使用される全教科の教科書を「基準コーパス」として使用している。Hasebe&Lee（2015）では、日本語教育のためのリーダビリティシステム構築を目標にしていたため、基本的には日本語教科書を使用しているが、上級前半と上級後半レベルを定義づけるために、例外的に「現代日本語書き言葉均衡コーパス」を使用している。

3. 日本語教育のためのリーダビリティ

Hasebe&Lee（2015）では、日本語教育のためのリーダビリティ構築のために、2 種類のデータセットを構築している。1 つ目は、初級から上級までの日本語教科書 83 冊と李（2011）で使用した「現代日本語書き言葉均衡コーパス」のデータで構成した「基本データ」、2 つ目は、旧日本語能力試験の 25 年分の読解テキストで構成した「評価データ」である。「基本データ」はリーダビリティ公式を開発するためであり、「評価データ」はリーダビリティ公式の妥当性を確認するために使用している。

分析は、次の 3 ステップで行われた。第 1 ステップとして「基本データ」をもとに「基準コーパス」を構築する作業、第 2 ステップとして「基準コーパス」をもとにリーダビリティ公式を作成する作業、第 3 ステップとして「評価データ」をもとにリーダビリティ公式の妥当性を確認する作業である。

3.1 第 1 ステップ：基準コーパス構築

リーダビリティ公式を作成するためには「基準コーパス」が必要である。この「基準コーパス」が満たすべき条件として、次の 2 点が考えられる。1 点目として、「初級前半、初級後半、中級前半、中級後半、上級前半、上級後半」の各レベルにおける言語的特徴を明確に持っていること、2 点目として、一定規模のデータサイズであることが求められる。この条件を満たすコーパスを作る作業として、「基本データ」に 2 つの作業を行った。1）す

べてのテキストファイルを同じ長さ（おおよそ 1000 文字）に分割したあと、2）各テキストファイルに対して主観判定と統計分析を実行し、「初級前半、初級後半、中級前半、中級後半、上級前半、上級後半」の 6 段階のレベルをつけた。6 段階のレベルイメージは表 2 のとおりであり、「基準コーパス」のデータサイズは表 3 のとおりである。

表 2　6 段階のレベルイメージ

レベル	能力記述文
初級前半	単文を中心とする基礎的日本語表現に関して理解できる。複文や連体修飾構造などの複雑な文構造は理解できない。
初級後半	基本的な語彙や文法項目について理解できる。テ形による基本的な複文なども理解できる。
中級前半	比較的平易な文章に対する理解力があり、ある程度まとまった文章でも内容が把握できる。
中級後半	やや専門的な文章でも大まかな内容理解ができ、日常生活レベルの文章理解においてはほぼ不自由がなく遂行できる。
上級前半	専門的な文章に関してもほぼ理解できる。文芸作品などに見られる複雑な構造についても理解できる。
上級後半	高度に専門的な文章に関しても不自由なく、理解できる。日本語のあらゆるテキストに対して困難を感じない。

表 3　「基準コーパス」のデータサイズ

	初級前半	初級後半	中級前半	中級後半	上級前半	上級後半
異なり語数	3,178	2,858	5,156	10,291	6,833	4,712
延べ語数	72,691	68,746	87,433	174,953	69,268	122,269

各レベルの具体的な文章例を以下に示す。

① 初級前半の文章例：音楽がすきですから、よく CD を聞きます。日本がすきですから、日本語を勉強します。安かったですから、買いました。ディズニーランドは楽しかったです。教室は静かでした。わたしはラーメンがすきです。わたしはたばこがきらいです。ワンさんは日本語が上手です。わたしは料理が下手です。

② 初級後半の文章例：むかしむかし、金が大好きな一人の王様がいました。ある日王様の家に一人の老人がやって来ました。その老人は有名な学者でしたが、お酒がたいへん好きでした。そこで、王様は、老人のためにたくさんの酒とおいしい料理を用意しました。10 日間、老人は飲んだり食べたりしました。そして、10 日目に満足して帰って行きました。この話を、酒の神が聞きました。酒の神はこの老人が好きだったので、王様にお礼をしたいと思いました。

③ 中級前半の文章例：毎週 1 回は祖母の家に子どもたちが孫たちをつれて集まります。とてもにぎやかです。祖母の 80 さいの誕生日には、マニラで一番大きなホテルを借りて、大家族の全員と親しい友人が、全部で 500 人以上集まりました。ごちそうを食べたり、ダンスをしたり、歌をうたったりして、とてもにぎやかでした。祖母

もワルツやチャチャチャをおどりました。それから子どもと孫の全員が花をプレゼントしました。

④　中級後半の文章例：いまでいうリフォーム、リサイクルをごく当たり前のこととしてやっていました。日本は、1950 年代後半から経済の成長がいちじるしく、供給がどんどん増加し、国民一人あたりの所得も上がってきました。この時代を境にして、需要と供給のバランスが逆転しました。現在の日本は完全に供給が過剰、需要が不足している時代です。ものをつくる企業はこういうときにどうするでしょうか。

⑤　上級前半の文章例：動物の動きにしてもそうで、ネズミはちょこまかしているし、ゾウはゆっくりと足を運んでいく。体のサイズと時間との間に、何か関係があるのではないかと、古来、いろいろな人が調べてきた。例えば、心臓がドキン、ドキンと打つ時間間隔を、ネズミで測り、ネコで測り、イヌで測り、ウマで測り、ゾウで測り、と計測して、おのおのの動物の体重と時間との関係を求めてみたのである。サイズを体重で表わすのは、体重なら、はかりにポイと載せればすぐ測れるが、体長でサイズを表わすと、しっぽは計測値に入れるのか、背伸びした長さか丸まったときの長さかなどと、難しい問題がいろいろ出てくるからだ。

⑥　上級後半の文章例：数学は，科学を記述する普遍的な言語であるという基本的な性格を持つ。また「自然は数学の言葉で書かれた書物である」とはガリレイの言である。

ニュートン以後１９世紀まで，古典物理学と数学とは，微分方程式と特殊関数の研究を“かなめ”として即かず離れずの関係で発展してきたが，今世紀に至り，場の量子論・統計力学と現代数学が結合し，数理物理学の新しい発展を遂げることになった。この展開によって，解析学のみでなくトポロジー，多様体論，代数幾何学，整数論にまでわたる，現代数学の先端諸分野を横断する新しい視点と手段がもたらされ，重要な問題の解決や新しい理論の展開にまで導かれることになった。数学の分野において我が国は多数の優れた研究者を擁し，世界のこの分野の発展に大きく貢献した業績は特筆すべきものがある。

3.2　第 2 ステップ：リーダビリティ公式の作成

第 2 ステップとして、基準コーパスに対して、自然言語処理のツールを利用してテキスト処理を行った。そして、ファイル単位で文字種別の使用頻度や品詞類の使用頻度を計算し、テキスト特徴量を抽出した。そして、統計分析として重回帰分析を行い、難易度を予測するリーダビリティ公式を作成した。なお、重回帰分析とは多変量解析の一種であるが、単回帰分析が一つの独立変数で分析するのに対して、重回帰分析では 2 つ以上の独立変数で分析を行う。回帰分析を行うことで、一方の値が与えられた時、他方の値を予測することができる。

統計分析は、「SPSS Statistics」を使って行った。重回帰分析の分析オプションとして、ステップワイズ法を使用し、5 つのモデルを生成した。各モデルの詳細は表 4 のとおりである。

表 4　重回帰分析の結果

		係数	決定係数（R^2）
モデル 1	（定数）	5.938	0.787
	平均文長	-.099	
モデル 2	（定数）	6.691	0.839
	平均文長	-.082	
	漢語率	-.073	
モデル 3	（定数）	13.195	0.878
	平均文長	-.063	
	漢語率	-.153	
	和語率	-.086	
モデル 4	（定数）	12.128	0.893
	平均文長	-.057	
	漢語率	-.142	
	和語率	-.061	
	動詞率	-.159	
モデル 5	（定数）	11.724	0.896
	平均文長	-.056	
	漢語率	-.126	
	和語率	-.042	
	動詞率	-.145	
	助詞率	-.044	

表 4 のモデル 1 は、定数と平均文長のみで構成されたモデルであり、予測精度を示す R^2 値は 0.787 である。モデル 2 からモデル 5 も同様の観点で捉え、決定係数の推移を確認したところ、モデル 5 が（R^2 値が高いため）もっとも予測精度が高いと判断し、モデル 5 を採用した。

この結果に従った場合、日本語教育のためのリーダビリティにおいては、文の長さを示す平均文長、語種に関連するものとして、和語や漢語の含有率、文法的特徴を示す動詞率と助詞率がもっとも重要な変量であるということになる。具体例として、3.1 節で初級前半の文章例として示した文章を解析した場合、以下のように計算する。

{8.56×-0.056}+{0.12×-0.126}+{0.83×-0.042}+{0.05×-0.145}+{0.22×-0.044}+11.724=6.08

リーダビリティ公式によって算出されたリーダビリティ値の「6.08」という数値は、表 5 の対応表に基づいて解釈する。この場合、5.5〜6.4 の間に入るため、初級前半のテキストであると判定される。

表 5　リーダビリティ値の解釈基準

レベル	リーダビリティ値	
	上限	下限
初級前半	5.5	6.4
初級後半	4.5	5.4
中級前半	3.5	4.4
中級後半	2.5	3.4
上級前半	1.5	2.4
上級後半	0.5	1.4

表 4 のリーダビリティ公式と表 5 の値の解釈基準は Excel などの表計算ソフトを使えば、簡単に計算できるが、前段階の平均文長や漢語率などの特徴量を計算するには、自然言語処理が用いられる形態素解析という技術を使わなければならず、一般のユーザーには実行することが難しい。このことを踏まえ、ウェブブラウザによるウェブサービスの 1 つとして、本研究の成果を組み込んだ。

図 1　リーダビリティ計算例

図 1 は、「日本語文章難易度判別システム」（http://jreadability.net/）によるデータ分析例である。本論の 2 節における一部のテキストを貼り付けてみたところ、リーダビリティ値としては、「2.45」という値が出力されており、表 5 に基づいて解釈すると、「上級前半」

と判定される。なお、上級後半の上限にあたる 0.5 を超える値、もしくは、初級前半の下限にあたる 6.4 を下回る値に関しては、システム上では「判定不能」という結果をかえすように設計されている。

3.3　第 3 ステップ：外部基準による公式の検証

第 3 ステップとして、旧日本語能力試験の読解領域の 172 テキストを利用し、リーダビリティ公式の評価を行った。この評価の趣旨としては、リーダビリティ公式を作った際に使用したデータ以外のものを使い、リーダビリティ値を計算してみることで、難易度の差が再現できるか確認するというものである。すなわち、リーダビリティ公式を作成する際に使用したデータ以外のもので解析しても、その難易度の差が捉えられるならば、リーダビリティ公式は妥当な公式だと言えるということである。

表 6　旧日本語能力試験の読解テキストのレベル×リーダビリティレベルのクロス集計

			リーダビリティレベル					
			初級前半	初級後半	中級前半	中級後半	上級前半	合計
旧日本語能力試験読解テキストレベル	1 級	度数	0	0	6	47	25	78
		JLPT の %	0.0%	0.0%	7.7%	60.3%	32.1%	100.0%
	2 級	度数	0	1	19	44	2	66
		JLPT の %	0.0%	1.5%	28.8%	66.7%	3.0%	100.0%
	3 級	度数	0	7	10	0	0	17
		JLPT の %	0.0%	41.2%	58.8%	0.0%	0.0%	100.0%
	4 級	度数	5	6	0	0	0	11
		JLPT の %	45.5%	54.5%	0.0%	0.0%	0.0%	100.0%
合計		度数	5	14	35	91	27	172
		JLPT の %	2.9%	8.1%	20.3%	52.9%	15.7%	100.0%

表 6 では、縦軸に旧日本語能力試験の読解テストに使用したテキストのレベル、横軸に Hasebe&Lee（2015）が提案するリーダビリティレベルを配置し、両者のクロス集計表を作成した。網かけの部分はデータが集中している箇所を示しているが、網かけ部分に注目した場合、旧日本語能力試験の読解テキストに関して、次の 4 点が言える。1）1 級の読解文章は、中級後半から上級前半のレベルに集中していること、2）2 級の読解文章は、中級前半から中級後半に集中していること。3）3 級の読解文章は初級後半から中級前半に集中していること、4）4 級の読解文章は初級前半から初級後半において分布していることが明らかになった。このことは、旧日本語能力試験の読解テキストの難易度の差は、本研究のリーダビリティ公式によっても、ある程度再現できていることを意味する。このことを確認する証拠として、旧日本語能力試験の読解テキストのレベルを因子、リーダビリティ値を従属変数にし、一元配置分散分析を行ったところ、読解テキストの難易度の差とリーダビリティ値には統計的に有意な差が確認された（$F(3,168)=141.035$, $p<.001$）。

以上の分析によって、本研究が提案するリーダビリティ公式は、日本語能力試験の読解テキストのように、日本語教育で信頼されているテキストデータにおける難易度の差も明確に捉えており、妥当性の高いものであると言える。

4. 均衡コーパスに対するリーダビリティ調査

李、長谷部、久保（2016）では、3 節のリーダビリティ公式を使い、「現代日本語書き言葉均衡コーパス DVD 版；以下、BCCWJ」および「読売新聞記事」から抽出した 1,949 個のテキストサンプルを分析し、日本語文章の（日本語教育のための）難易度に関する大規模な調査を行っている。

BCCWJ の完成以降、日本語教育においてコーパスを利用する試みは数多く出現しており、今後も加速化していくものと見られている。このようにコーパスを利用した研究が活発になっていく一方で、日本語教育に関わるものとしては、こうしたコーパスデータの教育コンテンツとしての妥当性について考えていかなければならないであろう。特に注目したい点として、コーパス内の文章が日本語学習者にとって、どの程度許容され、理解されるのかという問題意識を持たなければならない。こうした問題意識のもとで、李、長谷部、久保（2016）では無作為に抽出したテキストデータを対象に「日本語文章難易度判別システム」を使って文章難易度を調べた。

使用データは、BCCWJ の中から「書籍」、「Yahoo!知恵袋（web）」、「白書」、そして BCCWJ 以外のデータとして「日英新聞記事対応付けデータ（JENAAD; http://www2.nict.go.jp/astrec-att/member/mutiyama/jea/index-ja.html）（2016.7.27.閲覧）」の「読売新聞記事」の日本語部分を使用した。分析の結果として表 7 の内容が明らかになった。

表 7　難易度調査の結果

区分	初級前半	初級後半	中級前半	中級後半	上級前半	上級後半	超級
書籍	2(0.3%)	36(5.5%)	269(41.2%)	209(32.0%)	106(16.2%)	30(4.6%)	1(0.2%)
web	31(5.8%)	173(32.1%)	72(13.4%)	261(48.4%)	2(0.4%)	0(0.0%)	0(0.0%)
白書	0(0.0%)	0(0.0%)	16(6.3%)	0(0.0%)	46(18.0%)	102(39.8%)	92(35.9%)
新聞	0(0.0%)	0(0.0%)	0(0.0%)	24(4.8%)	477(95.2%)	0(0.0%)	0(0.0%)

表 7 の通り、書籍は中級前半から上級前半レベル、web のデータは初級後半から中級後半レベル、白書は上級後半から超級レベル、新聞は上級前半レベルに相当することが明らかになった。なお、超級レベルは、「日本語文章難易度判別システム」では（リーダビリティ値が 0. 49 を超えるため）「測定不可」として出力されるものであるが、白書のデータに多く分布するテキストであることから、議論の便宜上、設定したものである。さらに、分布を確認すべく、箱ひげ図を描画してみたところ、図 2 の結果になった。

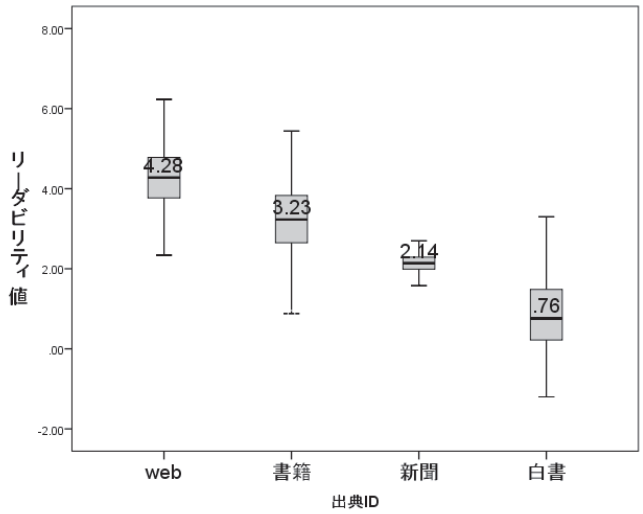

図 2　リーダビリティ値の分布

図 2 で注目すべき点は、1）web（Yahoo!知恵袋）が最も易しく、白書が最も難しいテキストであること、2）新聞は他のジャンルに比べ、リーダビリティのばらつきが少ないこと、3）白書は上級後半よりもさらに上のレベルに分布していることである。

李・長谷部・久保（2016）が行った調査は、コーパスを利用した語彙調査や文法項目の調査、さらには、表現の使用頻度に関する調査を行う際や、調査結果の利用方法を検討する際の基礎資料になると考えられる。

5. コーパス研究とリーダビリティ研究

リーダビリティ研究は、1）大規模データを利用する点、2）統計的な分析手法を利用する点で、コーパス研究に通じる側面がある。本節では、コーパス研究と言語教育の関係、さらには、コーパス利用におけるリーダビリティ研究の位置づけを検討する。

多くの先行研究が指摘することであるが、言語教育のコンテンツ作成においてコーパスは強力なツールになる（石川 2008、李・石川・砂川 2012、投野 2015）。特に英語教育においては、コーパス研究は量・質いずれにおいても充実しているが、中でも辞書編纂や学習者のための記述文法書の作成に関しては、コーパスは不可欠なツールであると認識されている [3]。

こうしたコーパス準拠の試みが行われている一番の理由は、個人単位では達成できない言語データの精緻な観察ができる点である。さらに、頻度情報をもとに言語研究の様々な課題に対して反証可能な形で考察を組み立てられる点も大きなメリットであると言える。日本語研究・日本語教育研究においても、国立国語研究所の BCCWJ の公開により、コーパス利用環境が整ったことや石川（2012）のような優れた入門書が刊行されたことを受け、多くの研究成果が公開されるようになった。

さて、コーパスと言語教育に関する研究を捉える視点について、投野（2015:7）は表 8 のようにまとめている。

表 8　コーパスと言語教育の応用を考える際の観点

観点	領域	具体例
利用モード	直接利用	・教室内の利用：データ駆動型学習 ・教員研修
	間接利用	・資料：学習語彙表 ・教材：辞書、文法書、教科書など ・シラバス・カリキュラム ・言語テスト ・CALL システム
	教育用コーパス作成	・学習者コーパス ・難易度調整済みコーパス
コーパス情報	語彙	語彙統計（頻度、分布）、コロケーション、分野別キーワード
	統語	品詞、構文解析（係り受け）、動詞の下位範疇化、名詞句の長さ
	談話	文の結束性、一貫性、談話標識など
学習者	外的	学習環境、教員の指導能力、IT スキル、学校のインフラ整備、学習形態
	内的	認知・学習スタイル、年齢、母語、外国語の習得レベル、適正、動機付け、性格、ニーズ

表 8 は、Leech（1997）が提唱した区分と投野（2003）のコーパス情報や学習者要因に関する区分を融合したものである。一言でコーパス基盤の言語教育といっても、その中身は様々であり、教室内で直接コーパスを使い、教育活動を行うパターンもあれば、資料作成目的で、間接的にコーパスを利用する立場もある。さらに、学習者コーパスということで、学習者の産出データを集め、第二言語習得研究の仮説検証のために使う研究もある。また、コーパスから利用する情報も、語彙、統語、談話など多種多様である。最後に、言語教育のためのコーパスの利用においては、学習者の外的・内的な要因との相互関係についても検討しなければならない。

投野（2015）は、コーパスの英語教育への応用研究はコーパスの教育利用に関する提案やシステム・教材の教具案を示すだけのものが大多数であるが、効果研究まで視野に入れて行っている研究は多くないと指摘している。日本語教育におけるコーパス研究の現状を見ても、基本的には同じ状況であると言える。間接利用の観点から様々な教育コンテンツが作成できることを示す研究が主であり、文法項目や語彙項目といった学習素材に対して頻度調査までは行っているものの、効果検証に関する考察はほとんどされていない。また、コーパス作成という観点においては、近年、科研費による大規模プロジェクト（例えば、I-JAS；https://ninjal-sakoda.sakura.ne.jp/lsaj/）（2016.7.27.閲覧）として教育用コーパス作成も進んではきているものの、公開されている研究リソースは限られたものしかなく、十分とは言えない。そして、コーパス情報を利用したトピック別の傾向を見ても語彙に関

するものがもっとも多く、統語や談話に関する研究はあまり進んでおらず、未開拓な部分が多いと言えよう。以上の理由からコーパスの教育利用は、まだ発展途上の段階にあるとみることができる。

さて、本研究のリーダビリティに関する公式開発およびコーパスデータに対する大規模調査を、表 8 の記述に基づいて捉えた場合、２つの観点から評価することができる。１つ目は、語彙論や統語論を超える言語単位の分析を行ったという意味で、コーパス研究の可能性を広げたと言える。２つ目は、コーパスデータの「直接利用」を支援するものとして位置づけることができる。リーダビリティ研究は、BCCWJ のような巨大なデータベースと日本語の読解クラスをつなぐ役割が期待できる。BCCWJ は、日本語の縮図を作るという目的のため、開発されたものではあるが、日本語教育的な観点から見た場合、巨大な読解用データバンクと見ることもできる [4]。

BCCWJ のようなコーパスに収録されているテキストは、生テキストであるがゆえに難易度という観点から見た場合、多種多様である。日本語クラスへの導入ということにおいては、どのクラスに、どの程度、用いるべきかということは容易に判断できない。こうした課題に対して、本研究が提案するリーダビリティ研究およびそれを実装したシステムを使うことで、誰が、いつ、どこで測っても、文章の難易度がぶれることなく判定できる。従って、本研究が提案する「日本語教育のためのリーダビリティ公式」と図 1 の「日本語文章難易度判別システム」は、将来においてコーパスデータと読解クラスをつなぐ役割が果たせるのではないかと考えている。

6. 終わりに：文章の難易度を測ることについて

本稿では、語彙や文法項目に比べ、難易度という観点からの分析が難しい「文章」データに関して、リーダビリティの観点から考察を行った。特に、日本語教育のためのリーダビリティ研究として Hasebe&Lee（2015）が行った基礎研究と李・長谷部・久保（2016）が行った応用研究を中心に研究成果を紹介してきたが、本節では、本稿の締めくくりとして、リーダビリティ公式を使って難易度を推定することの限界についても述べておく。

本稿の基本的態度として、文章の難易度を決定するための唯一無二の公式が存在するとは考えていない。実際の難易度というのは、文章によっても異なるが、読み手の属性によっても異なる。それを示す具体的な成果として、柴崎・玉岡・沢井（2008）の研究があげられる。柴崎・玉岡・沢井（2008）では、仮名だけで表記されたテキストを小学生と大学生にそれぞれ読ませ、文正誤判断課題を実施し、反応時間を測定した場合、大学生のほうが反応時間が長かったと報告している。この実験結果は、「仮名で書いてある文章＝やさしい文章」という前提がすべての日本語話者にとって成立するわけではないことを示している。

柴崎・玉岡・沢井（2008）の結果を日本語学習の文脈で考えてみた場合、漢字で書いてあるテキストであれば、すべて難しいかというと、そうではないということを示唆しているのではないだろうか。漢字圏の学習者にとっては、漢字で書いてあるほうが読みやすいと考えられる反面、非漢字圏の学習者にとっては、仮名表記で書かれてあったほうが読みやすいと考えられる。こうしたことから考えた場合、リーダビリティ公式というのは、学

習者の属性の数だけ存在するものと見るべきであろう。なぜなら、漢字圏学習者にとってのリーダビリティと非漢字圏学習者にとってのリーダビリティ、教室学習者にとってのリーダビリティと自然習得者にとってのリーダビリティ、さらには、成人学習者にとってのリーダビリティと年少者日本語学習者にとってのリーダビリティが同じであるという保証はどこにもないからである。こうした問題から考えてみた場合、筆者の研究グループは、日本語教科書という比較的に安定した資料を利用することで、現状として可能な最大公約数としての難易度判定を試みたと見るべきであろう。

謝辞

本研究は JSPS 科研費 25370573 および 16K02794 の助成を受け、行ったものである。また、本稿で紹介したリーダビリティシステムは、長谷部陽一郎氏，久保圭氏との共同研究によって開発したものである。両氏には、草稿にも目を通してもらって、有益なコメントをいただいた。感謝申し上げたい。

注

1 柴崎・原（2010）の研究成果は、http://readability.nagaokaut.ac.jp/readability（2016.7.27.閲覧）、佐藤（2011）の研究成果は、http://kotoba.nuee.nagoya-u.ac.jp/sc/obi3/（2016.7.27.閲覧）、Hasebe&Lee（2015）の研究成果は、http://jreadability.net/（2016.7.27.閲覧）で見ることができる。

2 n-gram とは、長さ n を持つ文字列または単語列などの記号列である。文字や単語の出現は直前の文字列または単語に影響されるということに着目した言語モデルである（言語処理学会（編）(2010:122-124)）。

3 このことを示す表的な事例として英国の COBUILD プロジェクトがある。詳細は、http://www.collins.co.uk/page/The+History+of+COBUILD（2016.7.27.閲覧）を確認してほしい。

4 BCCWJ の中には、宮沢賢治、夏目漱石、井上ひさし、菊池寛、遠藤周作、芥川竜之介、村上春樹など、日本語を代表する「美文」が多数、収録されており、日本語教育的な利用価値という意味においても様々な可能性を秘めていると考えられよう。

参考文献

庵功雄・イヨンスク・森篤嗣（編）(2012)『「やさしい日本語」は何を目指すか: 多文化共生社会を実現するために』ココ出版

庵功雄・山内博之（編）(2015)『データに基づく文法シラバス（現場に役立つ日本語教育研究 1)』くろしお出版

石川慎一郎（2008)『英語コーパスと言語教育―データとしてのテクスト』大修館書店

石川慎一郎（2012)『ベーシックコーパス言語学』ひつじ書房

岩田一成（2014)「看護師国家試験対策と「やさしい日本語」」『日本語教育』158 号、pp.36-48

岩田一成（2016)『読み手に伝わる公用文: 〈やさしい日本語〉の視点から』大修館書店

計量国語学会（2010)『計量国語学事典』朝倉書店

言語処理学会（編）(2010)『デジタル言語処理学会事典』共立出版

坂本一郎（1964)「文の長さの比重の測定法-Readability 研究の試み」『読書科学』8-1、pp.2-6

酒井由紀子（2011)「健康医学情報を伝える日本語テキストのリーダビリティの改善とその評価：一

般市民向け疾病説明テキストの読みやすさと内容理解のしやすさの改善実験」『Library and Information Science』65、pp.1-35
佐藤理史（2011）「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
柴崎秀子・玉岡賀津雄・沢井康孝（2008）「漢字表記と平仮名表記が文正誤判断課題に与える影響」、『言語科学会 2008 年国際大会予稿集』、pp.18-19
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、pp.215-232, 計量国語学会
建石由佳・小野芳彦・山田尚勇（1988）「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
田中英輝・美野秀弥・越智慎司・柴田元也（2012）「やさしい日本語による情報提供-NHK の NEWS WEB EASY の場合」、庵 功雄・イ ヨンスク・森 篤嗣（編）（2012）『「やさしい日本語」は何を目指すか: 多文化共生社会を実現するために』ココ出版、pp.31-57
陳志文（2012）『現代日本語の計量文体論』くろしお出版
投野由紀夫（2003）「コーパスを英語教育に生かすと英語教育」『英語コーパス研究』10、pp.249-264
投野由紀夫（2015）『コーパスと英語教育』ひつじ書房
中俣尚己（2014）『日本語教育のための文法コロケーションハンドブック』くろしお出版
前川守（編）（1995）『文学編　文章を科学する』岩波書店
李在鎬（2011）「大規模テストの読解問題作成過程へのコーパス利用の可能性」『日本語教育』148、pp.84-98
李在鎬・石川慎一郎・砂川有里子（2012）『日本語教育のためのコーパス調査入門』くろしお出版
李在鎬・長谷部陽一郎・久保圭(2016) 「日本語コーパスの文章難易度に関する大規模調査の報告」『2016 年度日本語教育学会春季大会予稿集』、pp.152-157
Flesch, Rudolph (1948) “A new readability yardstick,” *Journal of Applied Psychology*, *Vol.* 32. pp.221-233
Hasebe, Yoichiro, Jae-Ho Lee (2015) “Introducing a Readability Evaluation System for Japanese Language Education,” (CASTEL/J 2015)
Lee, Jae-ho and Yoichiro Hasebe (2016 forthcoming) “Readability Measurement for Japanese Text Based on Leveled Corpora, ” *Papers on Japanese Language from Empirical Perspective.*
Leech, Geoffrey (1997) “Teaching and language corpora: a convergence,” in Wichmann A., Fligelstone S., McEnery T. and G. Knowles (eds.) *Teaching and Language Corpora*, pp.1-23. Longman
Smith A. Edgar and Peter Kincaid (1970) “Derivation and Validation of the Automated Readability Index for Use with Technical Materials,” *The Journal of the Human Factors and Ergonomics Society* 12-5, pp.457–464
Sunakawa, Yuriko and Lee, Jae-ho, and Takahara, Mari. (2012) “The Construction of a Database to Support the Compilation of Japanese Learners Dictionaries,” *Acta Linguistica Asiatica* 2-2, pp.97-115

（り　じぇほ　早稲田大学大学院日本語教育研究科）