読解教育支援のためのリーダビリティー測定ツールについて

李 在鎬[1]　長谷部 陽一郎[1]　柴崎 秀子[2]
（[1] 情報通信研究機構　[2] 長岡技術科学大学）

1. 研究背景

読解教育にとって重要なのは、何を基準にし、児童生徒に本を与えるのかということである。これには、児童生徒当人の内容面の好みといった主観的な要素も大きく関わるが、当人の学年といった客観的な情報を基にし、適切な本を推定する方法も有効と言える。その具体策として、文章の難易度を機械的に測定し、（その文章を読んで理解できるための）適正学年を推定する方法があり、我々が開発した「日本語リーダビリティー測定ツール(http://readability.nagaokaut.ac.jp/)」もその一つである。

リーダビリティー(readability) とは文章の読み易さのことを指すが、米国を中心とする欧米諸国では 1920 年台に研究が始められ、様々な公式が提案されてきた。これらは、いずれも読み易さを尺度化することで、数値として明示化しようとする試みである。近年、こうした試みは、図書選択における有用な指標として実用化されつつあり、様々な分野で期待されている。教育分野のみならず、商業分野でも注目されている。例えば、Amazon.com などでは、利用者へのサービスとして Flesch-Kincaid Index で難易度を表示している[1]。また、韓国などでも Kyobo-books (http://www.kyobobook.co.kr/) を中心とする大手書店で同様の試みがなされている。

さて、日本語に関するリーダビリティー研究としては、坂本(1971)を始めとして、「読書科学」の分野で注目すべき研究があった。ただ、その多くが手分析によるものであったため、大量のテキストを扱うには限界があった。しかし、近年、計算機の普及や Web の進化に伴い、ネットサービスの一つとして一般に公開されているものもあるなど、リータビリティー研究は多様化しつつある[2]。

以下では柴崎・沢井(2007)の公式を実装した「日本語リーダビリティー測定ツール」の開発手順や開発において直面した問題点を中心に報告する。

2. 先行研究

測定ツールが公開されている日本語リーダビリティーの主要な研究として、建石(他)（1998）、柴崎・沢井(2007)、Sato et al.（2008）がある。

まず、建石(他)（1988）は、文字種、文章の長さ、同種の文字の連続性を変数とし、ゼロから 100 までのポイント数で文章の難易度を出力する。次に、柴崎・沢井(2007)は、文章中の平仮名の割合、1 文における平均述語数や平均文字数、そして平均文節数を変数とし、適切学年を出力する。最後に、Sato et al. (2008) は、小学 1 年から大学までの 13 レベルの文字 unigram モデルが構築されており、その 13 の難易度から当該テキストに最適なレベルが出力される。いずれの研究でも、Web インターフェイスとして測定ツールが一般に公開されている[3]。

Sato et al.(2008)や建石(他)(1988)の計算方法は、主として文字列を変数とし、リーダビリティーを測定している。しかし、柴崎・沢井(2007)は文字列の他に述語数や文節といった、より実質的な言語情報を組み合わせ、リーダビリティーを測定している。なお、言語情報の測定には形態素解析や係り受け解析を利用している。

[1] この公式で計算された数値はゼロから 100 までで，ゼロに近いほど難しい。Flesch-Kincaid Index の詳細は、http://en.wikipedia.org/で詳しく解説されている。

[2] ネットサービスとして商用化されたツールとして popIn などがある。詳細は http://popin.cc/ja/home.html を参照してほしい。

[3] 建石(他)（1988）の詳細は http://www.utexas.edu/research/accessibility/index.html を参照してほしい。Sato et al.(2008)の詳細は http://kotoba.nuee.nagoya-u.ac.jp/sc/readability/を参照してほしい。

さて、文字列メインの方法と文字列と言語情報を組み合わせた方法では次のような一長一短がある。文字列メインの測定方法では、形態素解析の精度に依存しないため、測定結果は安定するが、文構造の複雑さといった要素に対応できないといった問題がある。一方の文字列と言語情報を組み合わせた方法では形態素解析の精度に依存するため、測定結果には部分的に誤りが含まれる。しかし、この短所を補う以上の長所として、複文のように複雑な述語構造を持つテキストとそうでないものに異なるスコアを与えることができる。

3. 測定ツールの開発

柴崎・沢井 (2007) の公式に基づく測定ツールについて述べる。まず、用いた公式は以下の通りである[4]。

$$Y=-0.148X_1+1.585X_2-0.117X_3-0.126X_4+15.581$$

***Y=学年, X_1＝文章中の平仮名の割合, X_2＝1 文の平均述語数, X_3＝1 文の平均文字数, X_4＝1 文の平均文節数**

平均述語数は形態素解析で、平均文節数は係り受け解析で計算している。形態素解析には、開発当初は ChaSen を使っていた。しかし、小学校低学年の教材は平仮名が連続していることが多く、方言や口語表現も多いため、形態素解析がうまくできなかった。そこで、形態素解析の辞書を標準の IPA 辞書 から UniDic(http://www.tokuteicorpus.jp/dist/)に変更し、形態素解析器も MeCab に切り替えた。その結果、全体の解析精度も大きく向上した。

アプリケーションは、Ruby を用いて、Web アプリケーションとして製作した。具体的には図 1 の初期画面からテキストを貼り付け、テキストのリータビリティーを測定することができる。測定の結果が、図 2 であり、文字種などの情報がグラフとして表示され、推奨学年となるリーダビリティー値が表示される仕組みとなっている。

図 1. 文章入力画面

図 2. 結果の出力画面

3.1 文の認定

Web インターフェイス画面を通じて与えられた日本語テキストは、まず複数の文へと分割される。これには主に句点(。)をマーカーとして用いるが、単純に句点だけを用いた文分割には3つの問題があった。

1. **句点以外が文区切りマーカーとして使用される問題**
2. **カギ括弧内に埋め込まれた文を検出する問題**
3. **改行に関する問題**

一つ目に、現実世界のテキストでは、句点だけが文区切りマーカーとして用いられているわけではない。句点以外のマーカーとして最もよく現れるのはクエスチョンマーク(？)とエクスクラメーションマーク(！)であるが、教科書以外のテ

[4] 公式開発においては、48 冊の小中高の国語科教科書を利用し、重回帰分析で学年への予測力のもっとも高い変数を選別した。詳細は柴崎・沢井(2007)を参照してほしい。

キストに目を向けると、これらを組み合わせたもの(！？，！！)や、三点リーダー(・・・)などの記号群が文区切りマーカーとなっている場合が少なくない。また、いわゆる「ケータイ小説」のように現代的なテキストの中には、句点を複数並べたもの(。。。)など、例外的なマーカーも観察される。このような例にも対応できるよう、本システムでは、句点マーカーとして使用される可能性のある記号をリストとして持つだけではなく、それらを組み合わせた記号列を単一のトークンとして検出する仕組みを実装した。

二つ目に、直接話法による被引用部など、カギ括弧内(「」,『』)に埋め込まれた文を適切に検出しなければならないという問題がある。これについては、単純なように見えて、クリアすべき事項が多い。被引用部では、テキストの性質や筆者のスタイルによって、閉じ括弧の前後に句点(やその他の文マーカー)が付加される場合と、そうでない場合がある。また、句点が付加される場合、閉じ括弧の前に置かれる場合と後に置かれる場合とがある。一方、句点が付加されない場合には、括弧内の文字列が「文」と認定すべきものかどうか曖昧であることが多い。さらに、カギ括弧が二重以上の入れ子になっている部分を文として認定しないように処理する仕組みも必要である。これらのことから、本システムではテキストを文字ごとにスキャンして確実に括弧の対応をとった。そして、いわゆる「地の文」と被引用部とを厳密に区別した上で、被引用部に下に挙げる記号および語が含まれている場合(のみ)、独立した文としてカウントすることにした。

1.「。」 2.「、」 3.「？」 4.「！」 5.「が」 6.「に」 7.「を」 8.「は」 9.「で」

このような方針をとることで、閉じ括弧付近の句点についての処理や、入れ子状に深く埋め込まれた部分を文と見なさない処理も容易になる。その結果、100%ではないものの、かなりの精度での文分割が可能になった。なお、関連した問題として、いわゆる地の文が、被引用部に先行する文字列と後続する文字列の2つに分離してしまう場合がある。これについて、本システムでは、前者を被引用部と結合し、後者は被引用部から独立させるという方針をとっている。それにより、「被引用部に対応する地の文は常に1つ」という原則のもと、少なくとも文数のカウントに関しては確実な値を得ることができる。具体例を示す。

[分割前]
良一は「僕は言ってないよ。『君が犯人だ』なんて」と答えた。

[分割後]
S1: 良一は「僕は言ってないよ。
S2:『君が犯人だ』なんて」
S3: と答えた。

三つ目に、改行に関するもので、テキストによっては改行が文の切れ目と一致しておらず、文中の任意の場所で挿入されている。その一方で(詩などのように)句点によってではなく、改行によって文の切れ目を表現しているものも存在する。前者については、すべて改行が文の切れ目と一致しているテキストのみを容認すると規定する方法もあり得るが、後者についての解決にはならない。また、改行によって文の切れ目を表したテキストは例外として排除するという選択肢もあるが、この種のテキストは小学教科書などで散見されるため、本研究の目的に照らすと無視できない。そこで、本システムでは、オプションとして「改行も文区切りとみなす」モードを設け、これが指定された場合には、すべての改行が文区切り位置にあると仮定し、句点が行末になくとも文として容認するようにしている。

3.2 文節と文字種の測定

文節の認定は、係り受け解析器 CaboCha を用いている。CaboCha では、形態素解析の結果をもとにテキストを文節の単位で分割した上で文の係り受け構造を表示するため、これを利用して文節分割が実現できるのである。なお、本システムでは MeCab と UniDic を用いている

が、CaboCha が本来想定している IPA 辞書の品詞構造と UniDic の品詞構造との間に多少の相違があるため、両者の擦り合わせを行う必要があった。その際、本システムにとってできるだけ理想的な文節分割が行われるよう、一部独自の調整を行っている。

テキストが文節に分割されると、次に述語数のカウントが行われる。その際の認定基準は基本的に次の通りである。

1. **すべての動詞（２つ以上の動詞から成る複合語は１語と数える。例：入り込む，連れ出す，呼びつける，走り回る，教えてもらう，来てくれる，歩いていく）**
2. **叙述形容詞（例：空は高く，山は青い。父の手は大きい。）**
3. **形容動詞（例：その男は正直で，誠実だった。）**
4. **名詞＋助動詞（例：明日はよい天気でしょう。これは母の鏡だ。次は渋谷ですか。）**
5. **名詞＋句点（例：空からふる白いものは雪。）**
6. **非自立名詞＋助動詞（例：のだ，のです）**
7. **形容詞連用テ接続＋読点（日本の車は安くて，性能も良くて，デザインもいい。）**
8. **文末の体言**

ただし、これらの条件を満たす文字列でも、「形容動詞語幹＋な」や「形容詞連用テ接続」などの形式を持ち、かつ名詞句の修飾句となっている語句は非述語とみなして排除する。この処理にあたっては、MeCab と UniDic による形態素解析を用いて大まかに判断した後、CaboCha による係り受け解析の結果を参照することで精度の向上を図っている。

最後に、各文字種のカウントは、ごく単純に「漢字」「カタカナ」「ひらがな」「記号」といった文字種を数え上げ、それぞれの割合を算出している。

4. 最後に

本システムはオブジェクト指向的な設計に基づいている。各処理の過程を通じて、テキストの文法構造をできるだけ忠実に反映させたオブジェクトの構築を行っており、測定の完了時には、テキストを構成する文オブジェクトや文を構成する形態素オブジェクトといったものが、緊密に関連し合った形で内部的に展開される。つまり、当該のテキストに含まれる各種要素の属性と相互関係がすべて利用可能な状態になっているのである。処理コストを考えたならば、よりシンプルな実装も可能であるが、本システムでは、今後さらに精度を高めるための調整や機能拡張を行う可能性を鑑み、このような設計を採用している。

最後に今後の課題として、語彙の難易はリーダビリティーの大きな要因となるが、現在のシステムでは実装されていない。今後は単語親密度を変数化にし、公式に組み込んでいこうと考えている。

* 本研究におけるシステムの開発は、基盤研究(B)(課題番号: 1930277、研究代表: 柴崎秀子)および特定領域研究(課題番号: 19011003、研究代表: 柴崎秀子)からの研究助成を受けて行った。

〈参考文献〉

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

[2] Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh.(2008) Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus, LREC-08. pp.28-30.

[3] 柴崎秀子・沢井康孝（2007）「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.

[4] 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」『文書処理とヒューマンインターフェース』18-4, p1-8