# 難しい日本語文の自動検出のための基礎調査

伊藤 美咲姫　　佐藤 理史　　駒谷 和範

名古屋大学 大学院工学研究科 電子情報システム専攻
{misaki_i, ssato, komatani}@nuee.nagoya-u.ac.jp

## 1　はじめに

英語のテキストの難易度を推定する研究は長い歴史を持ち [1,2]，以前から，Unix コマンドの style など，すぐに使えるツールが存在した．日本語テキストに対しては，長らくこのようなツールが存在しなかったが，難易度推定システム『帯 2』[3,4] の出現により，この状況は変わりつつある.『帯 2』は，あらかじめ用意した難易度の規準となるコーパスに基づき，与えられたテキストの難易度を推定する．現在公開されている『帯 2』(http://kotoba.nuee.nagoya-u.ac.jp) では，教科書コーパスに基づく 13 段階の難易度，および，現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese，以下 BCCWJ）に基づく 9 段階の難易度を推定することができる．

『帯 2』が推定するのは，与えられたテキスト全体の難易度である（テキストは，1000 字以上であることが望ましい）．このツールを用いて，ある文書の難易度が，想定する（望ましい）難易度より高いということがわかったとしよう．その場合，その文書をよりやさしく書き換える必要が生じるが，このツールは，その文書のどの部分を書き換えるべきかについて，何も情報を提供しない．このような情報を出力するためには，テキストを構成する部分（段落や文）の難易度を推定する必要がある．

我々は，現在，このような機能を実現するために，文の難易度を推定する方法について検討している．その一環として,『帯 2』によって 9 段階の難易度を付与した BCCWJ のサンプル [5] を対象に，51 種類の指標と難易度の相関を調査した．本稿では，この調査結果について報告する．

## 2　関連研究

日本語テキストの難易度を推定する他の研究として，柴崎らの研究 [6] がある．この研究では，国語教科書コーパスを用いて各学年のテキストに対するいくつかの指標の値を求め，学年による増加または減少が見られた指標を用いて重回帰分析を行い，9 段階の学年の推定を行う公式を作成した．この公式が使用する指標は，文章の総文字数に対するひらがなの割合と，1 文あたり平均述語数の 2 つである．この式は，テキスト全体の難易度を推定するための式であるが，文に対しても適用可能であると考えられる．

小島 [7] は，全教科を網羅した教科書コーパス [8] を利用し，各学年のテキストに対するいくつかの指標の値を求め，13 段階の学年と相関の高い指標として，「(1 文あたりの）基本語彙約 2,500 語以外の語を含む文節数」を得，この指標を用いて文の難易度を推定する方法を提案している．

これらの研究は，いずれも，難易度の規準として，教科書のテキストを用いている．しかしながら，教科書のテキストは，日本語としての代表性を有しない．これに対して本研究では，代表性を有する BCCWJ を使用することにより，より一般性を持った結果を得ることを目指す．

## 3　調査方法とその結果

### 3.1　調査対象コーパス

本調査では，BCCWJ の一部を調査対象とした．

BCCWJ を構成する 3 つのサブコーパスのうち，厳密にサンプリングされているのは出版サブコーパスと図書館サブコーパスであり，その両者に共通するメディアは，書籍である．本調査では，これらのサンプル，すなわち，レジスタが PB（出版サブコーパスの書籍）と LB（図書館サブコーパスの書籍）を調査対象とした．

BCCWJ には，約 1000 字から構成される固定長サンプルと，長さが揃っていない可変長サンプルがあるが，各種の統計値を求めるという観点から，長さが揃っている固定長サンプルを選択した．最終的に，有効文

字 bigram[1] が 300 以上存在する，合計 20,664 件のサンプルを対象とした．なお，これらのサンプルには，『帯 2』により，9 段階の難易度が付与されている [5].

### 3.2 文の抽出

対象とするサンプルの固定長 XML 文書ファイルから，文を抽出した．固定長 XML 文書ファイルは，テキストに文書構造タグが付与されている (図 1).

この図に示すように，固定長 XML 文書ファイルは，基本的には 1 行 1 文形式であり，文は<sentence>タグで囲まれている．しかしながら，テキストには，他のテキストからの引用や，厳密な意味では文とは言えないタイトル等も含まれている．

そこで，次のような基準で，抽出する文を決定した．

1. <sentence>タグで囲まれた部分を文として抽出する
2. </sentence>タグの直後が改行以外の場合は，抽出しない
3. <sentence type="quasi">タグ内からは抽出しない
4. 複数行に渡る<quote>タグ内からは抽出しない
5. 句点（。）で終わっている文のみを抽出する
6. 1 行に，複数の句点が含まれている場合は，抽出しない
7. 句点のみの文は抽出しない

これらの基準により，たとえば，図 1 からは，9 行目の文のみが抽出される．

抽出した文を 9 段階の難易度のそれぞれで集計した数を，表 1 の上部に示す．

### 3.3 形態素解析辞書への基本語彙データの反映

各種の統計値を計算するために，各文を形態素・文節解析する．

小島の研究では，mecab 用の IPADIC に基本語彙データを付与した辞書を用いて形態素解析を行い，その後，cabocha を用いて，文節まとめ上げを行なっている．本研究でも，mecab と cabocha を用いて形態素・文節解析を行なう．ただし，mecab 用の辞書には，BCCWJ が準拠する UniDic-2.1.0 に，基本語彙データを付与したものを用いる．

基本語彙データには，JC1.6 [9] を用いた．この基本語彙データは，約 3.5 万語に以下の 5 つの基本語彙レベルが付与されている．

**A1** 基本語彙約 2,500 語 (に含まれる)
**A2** 基本語彙約 2,500 語から約 5,000 語
**B** 基本語彙約 5,000 語から約 10,000 語
**C** 基本語彙約 10,000 語から約 20,000 語
**F** 基本語彙約 20,000 語から約 35,000 語

一方，UniDic-2.1.0 には，語彙素・語形・書字形という階層が存在する．まず，JC1.6 の各語と UniDic-2.1.0 の語形との対応を取り[2]，語形に上記の基本語彙レベルを付与した．その後，それぞれの語彙素に，所属する語形の基本語彙レベルのうち，最も低いレベルを付与した．

こうして得られた辞書を形態素解析で用いることにより，それぞれの形態素に対して，上記のレベルのいずれか（または，レベルなし）が付与されることになる．形態素解析では，語形に付与された基本語彙レベルと，語彙素に付与されたレベルの両方が出力されるが，以降の調査では，語形に付与された基本語彙レベルを利用した．

### 3.4 調査項目とその結果

今回の調査では，51 種類の指標の値を調査した．これらはすべて，各難易度のテキストの，1 文当たりの平均値である．すなわち，本調査では，各難易度における，1 文当たりの平均値を，その難易度を代表する文（の値）と見なす方法を採用する．これは，それぞれの文に難易度が付与されたコーパスが存在しないための便宜的な措置である．

51 種類の指標は，

1. 文字に関する指標：9 件
2. 形態素に関する指標：21 件
3. 文節に関わる指標：21 件

の 3 種類に分類される．それらは更に，長さで正規化していない指標と，長さで正規化した指標の 2 種類に細分される．

調査した指標とその結果の一覧を表 1 に示す．

#### 3.4.1 文字に関する指標

文字に関する指標では，1 文当たりの文字数，それぞれの字種 (ひらがな，カタカナ，漢字，その他) の文

[1] ひらがな（83 文字），カタカナ（84 文字）,JIS 第一水準の漢字（2,965 文字）から構成される bigram.『帯 2』では，難易度判定において，有効文字 bigram が用いられている．

[2] 表記，読み，品詞などを用いて，確からしい対応関係を機械的に推定した．

```
01: <quotation>
02: <speech>
03: <paragraph>
04: <sentence type="quasi">  「そんうちい降るじゃろう」</sentence>
05: <br type="automatic_original" />
06: </paragraph>
07: </speech>
08: </quotation>
09: <sentence>ち，言いよるうちい，また田は干割れちしもうた．</sentence>
10: <br type="automatic_original" />
```

図 1: 固定長 XML 文書の例 (LBc3_00052.xml の一部)

字数，および，それらの割合 (1 文において，どれだけの割合を占めるか) を調べた．

#### 3.4.2 形態素に関する指標

形態素に関する指標では，1 文当たりの形態素数，形態素ランク別の内容語数，および，それらの割合について調べた．

形態素ランク (内容語のみに対して設定される) は，次のように設定した．

1. 名詞-数詞，および，名詞-固有名詞は，付与されている基本語彙レベルに関わらず，形態素ランクを 0 とする．
2. 基本語彙レベルが付与されている内容語は，そのレベルに応じたランクとする (A1 から順に，1 から 5 を用いる)．
3. 基本語彙レベルが付与されていない内容語は，ランクを 6 とする．

内容語の判定では，品詞大分類が，名詞，動詞，形容詞，副詞，形状詞，感動詞，代名詞，連体詞，接続詞の形態素を内容語と判定した．

#### 3.4.3 文節に関する指標

文節に関する指標では，1 文当たりの文節数，文節ランク別の文節数，および，それらの割合について調べた．

文節ランクは，文節に含まれる形態素のランクの中で，最も高いものを，その文節のランクとする．例えば，「解析する」という文節は，「解析（基本語彙レベル F，ランク 5 の形態素)」と「する（基本語彙レベル A1，ランク 1 の形態素)」を含むので，この文節は，「ランク 5 の文節」となる．

### 3.5 検討

表 1 より，次のことが観察される．

- 1 文あたりの文字数，形態素数，文節数といった，長さに関する指標は，全て 9 段階の難易度と高い相関をもつ．
- 文字に関する指標では，ひらがなの割合が負の相関を，漢字の割合が正の相関を示す．
- 形態素に関する指標では，ランクの高い形態素数が高い相関を示す．
- 文節に関する指標では，ランク $R(R \geq 2)$ 以上の文節数が高い相関を示す．それと同時に，ランク 1 の文節の割合が，高い負の相関を示す．

最後の観察結果は，小島 [7] の研究と一致する．ランク 1 の文節数は，難易度の増加とともに増加する (正の相関を持つ) のに対し，ランク 1 の文節の割合は，難易度の増加とともに減少する (負の相関を持つ)．これは，興味深い事実であり，この点について，さらに調査を進める必要がある．

## 参考文献


[1] W.H. DuBay. *Smart Language*. Impact Information, 2007.

[2] W. H. DuBay. *Unlocking Language*. Impact Information, 2007.

[3] Satoshi Sato, Suguru Matsuyoshi and Yohsuke Kondoh. Automatic assessment of Japanese text readability based on a textbook corpus. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 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, Vol. 52, No. 4, pp. 1777–1789, 2011.

[5] 佐藤理史. 現代日本語書き言葉均衡コーパスに対する難易度付与. 第 2 回 コーパス日本語学ワークショップ予稿集, pp. 174–184, 2012.

[6] 柴崎秀子, 玉岡賀津雄. 国語科教科書を基にした小・中学校の文章難易学年判定式の構築. 日本教育工学会論文誌, Vol. 33, No. 4, pp. 449–458, 2010.

[7] 小島健輔. 日本語テキストの難易度推定. 名古屋大学大学院 工学研究科 修士論文, 2011.

[8] 松吉俊, 近藤陽介, 橋口千尋, 佐藤理史. 全教科を収録対象とした日本語教科書コーパスの構築. 言語処理学会第 14 回年次大会発表論文集, pp. 520–523, 2008.

[9] 佐藤理史. 異表記同義語認定のための辞書編纂. 情報処理学会研究報告 2004-NL-161, pp. 97–104, 2004.

表 1: 調査結果

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
|---|---|---|---|---|---|---|---|---|---|---|
| 使用した文の数 | 21,196 | 34,081 | 74,642 | 77,192 | 91,729 | 50,313 | 35,334 | 20,210 | 12,097 | |
| 文字に関する指標 | | | | | | | | | | |
| 指標 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 相関係数 |
| 文字数 | 30.72 | 27.50 | 31.13 | 36.51 | 42.69 | 50.60 | 53.39 | 54.10 | 65.21 | **0.97** |
| ひらがな文字数 | 18.79 | 16.26 | 18.03 | 20.30 | 21.94 | 24.07 | 24.27 | 21.61 | 28.09 | 0.87 |
| カタカナ文字数 | 2.52 | 1.90 | 1.82 | 2.22 | 3.03 | 3.43 | 3.83 | 5.71 | 1.42 | 0.42 |
| 漢字文字数 | 5.62 | 6.40 | 8.17 | 10.37 | 13.32 | 17.73 | 19.01 | 18.43 | 28.34 | **0.96** |
| その他文字数 | 3.79 | 2.95 | 3.11 | 3.62 | 4.40 | 5.37 | 6.29 | 8.35 | 7.36 | 0.91 |
| ひらがな割合 | 0.612 | 0.591 | 0.579 | 0.556 | 0.514 | 0.476 | 0.455 | 0.399 | 0.431 | **-0.97** |
| カタカナ割合 | 0.082 | 0.069 | 0.059 | 0.061 | 0.071 | 0.068 | 0.072 | 0.106 | 0.022 | -0.20 |
| 漢字割合 | 0.183 | 0.233 | 0.262 | 0.284 | 0.312 | 0.350 | 0.356 | 0.341 | 0.435 | **0.96** |
| その他割合 | 0.123 | 0.107 | 0.100 | 0.099 | 0.103 | 0.106 | 0.118 | 0.154 | 0.113 | 0.38 |
| 形態素に関する指標 | | | | | | | | | | |
| 指標 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 相関係数 |
| 形態素数 | 18.80 | 17.41 | 19.96 | 23.40 | 27.30 | 32.62 | 33.99 | 33.31 | 43.21 | **0.96** |
| ランク 0 の形態素数 | 0.68 | 0.57 | 0.68 | 0.99 | 1.51 | 2.13 | 1.73 | 1.43 | 2.10 | 0.86 |
| ランク 1 の形態素数 | 4.69 | 4.49 | 5.24 | 6.02 | 6.63 | 7.55 | 7.88 | 7.24 | 9.44 | **0.95** |
| ランク 2 の形態素数 | 0.83 | 0.74 | 0.83 | 1.01 | 1.33 | 1.87 | 2.34 | 2.53 | 3.44 | **0.95** |
| ランク 3 の形態素数 | 0.53 | 0.47 | 0.61 | 0.82 | 1.12 | 1.61 | 1.98 | 2.28 | 3.34 | **0.95** |
| ランク 4 の形態素数 | 0.35 | 0.33 | 0.40 | 0.50 | 0.62 | 0.73 | 0.71 | 0.75 | 1.60 | 0.84 |
| ランク 5 の形態素数 | 0.38 | 0.31 | 0.37 | 0.47 | 0.62 | 0.78 | 0.92 | 1.11 | 1.29 | **0.96** |
| ランク 6 の形態素数 | 1.01 | 0.75 | 0.83 | 1.02 | 1.28 | 1.50 | 1.67 | 2.07 | 1.60 | 0.88 |
| ランク 2 以上の形態素数 | 3.00 | 2.48 | 2.93 | 3.73 | 4.90 | 6.43 | 7.57 | 8.70 | 11.25 | **0.96** |
| ランク 3 以上の形態素数 | 2.28 | 1.86 | 2.21 | 2.82 | 3.64 | 4.62 | 5.28 | 6.21 | 7.83 | **0.96** |
| ランク 4 以上の形態素数 | 1.75 | 1.39 | 1.59 | 1.99 | 2.52 | 3.01 | 3.29 | 3.93 | 4.49 | **0.96** |
| ランク 5 以上の形態素数 | 1.40 | 1.06 | 1.20 | 1.50 | 1.90 | 2.28 | 2.59 | 3.18 | 2.89 | 0.94 |
| ランク 0 の形態素割合 | 0.036 | 0.033 | 0.034 | 0.042 | 0.055 | 0.065 | 0.051 | 0.043 | 0.049 | 0.58 |
| ランク 1 の形態素割合 | 0.250 | 0.258 | 0.262 | 0.257 | 0.243 | 0.232 | 0.232 | 0.217 | 0.219 | -0.89 |
| ランク 2 の形態素割合 | 0.044 | 0.042 | 0.042 | 0.043 | 0.049 | 0.057 | 0.069 | 0.076 | 0.080 | 0.93 |
| ランク 3 の形態素割合 | 0.028 | 0.027 | 0.031 | 0.035 | 0.041 | 0.049 | 0.058 | 0.068 | 0.077 | **0.97** |
| ランク 4 の形態素割合 | 0.019 | 0.019 | 0.020 | 0.021 | 0.023 | 0.022 | 0.021 | 0.022 | 0.037 | 0.72 |
| ランク 5 の形態素割合 | 0.020 | 0.018 | 0.018 | 0.020 | 0.023 | 0.024 | 0.027 | 0.033 | 0.030 | 0.90 |
| ランク 6 の形態素割合 | 0.054 | 0.043 | 0.042 | 0.044 | 0.047 | 0.046 | 0.049 | 0.062 | 0.037 | 0.04 |
| ランク 2 以上の形態素割合 | 0.160 | 0.142 | 0.147 | 0.160 | 0.180 | 0.197 | 0.223 | 0.261 | 0.260 | 0.93 |
| ランク 3 以上の形態素割合 | 0.121 | 0.107 | 0.111 | 0.120 | 0.133 | 0.142 | 0.155 | 0.186 | 0.181 | 0.92 |
| ランク 4 以上の形態素割合 | 0.093 | 0.080 | 0.080 | 0.085 | 0.092 | 0.092 | 0.097 | 0.118 | 0.104 | 0.76 |
| ランク 5 以上の形態素割合 | 0.074 | 0.061 | 0.060 | 0.064 | 0.070 | 0.070 | 0.076 | 0.096 | 0.067 | 0.47 |
| 文節に関する指標 | | | | | | | | | | |
| 指標 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 相関係数 |
| 文節数 | 7.04 | 6.54 | 7.43 | 8.57 | 9.82 | 11.33 | 11.65 | 11.07 | 14.58 | **0.95** |
| ランク 0 の文節数 | 0.37 | 0.52 | 0.51 | 0.52 | 0.58 | 0.67 | 0.40 | 0.24 | 0.26 | -0.42 |
| ランク 1 の文節数 | 3.80 | 3.58 | 4.09 | 4.56 | 4.83 | 5.07 | 4.99 | 4.17 | 5.63 | 0.78 |
| ランク 2 の文節数 | 0.76 | 0.69 | 0.76 | 0.89 | 1.14 | 1.53 | 1.80 | 1.73 | 2.25 | **0.96** |
| ランク 3 の文節数 | 0.48 | 0.43 | 0.56 | 0.73 | 0.96 | 1.35 | 1.59 | 1.70 | 2.45 | **0.96** |
| ランク 4 の文節数 | 0.32 | 0.31 | 0.37 | 0.46 | 0.56 | 0.64 | 0.60 | 0.60 | 1.35 | 0.81 |
| ランク 5 の文節数 | 0.36 | 0.29 | 0.34 | 0.44 | 0.56 | 0.70 | 0.80 | 0.91 | 1.17 | **0.96** |
| ランク 6 の文節数 | 0.96 | 0.73 | 0.79 | 0.97 | 1.19 | 1.38 | 1.48 | 1.71 | 1.46 | 0.90 |
| ランク 2 以上の文節数 | 2.77 | 2.32 | 2.73 | 3.40 | 4.35 | 5.54 | 6.22 | 6.62 | 8.67 | **0.96** |
| ランク 3 以上の文節数 | 2.12 | 1.76 | 2.07 | 2.60 | 3.27 | 4.07 | 4.47 | 4.92 | 6.44 | **0.96** |
| ランク 4 以上の文節数 | 1.64 | 1.32 | 1.51 | 1.86 | 2.31 | 2.72 | 2.87 | 3.22 | 3.99 | **0.96** |
| ランク 5 以上の文節数 | 1.32 | 1.01 | 1.14 | 1.41 | 1.75 | 2.08 | 2.28 | 2.62 | 2.63 | **0.95** |
| ランク 0 の文節割合 | 0.052 | 0.080 | 0.068 | 0.060 | 0.059 | 0.059 | 0.034 | 0.022 | 0.018 | -0.82 |
| ランク 1 の文節割合 | 0.539 | 0.547 | 0.551 | 0.533 | 0.491 | 0.447 | 0.428 | 0.377 | 0.386 | **-0.95** |
| ランク 2 の文節割合 | 0.108 | 0.105 | 0.103 | 0.104 | 0.117 | 0.135 | 0.154 | 0.156 | 0.154 | 0.91 |
| ランク 3 の文節割合 | 0.068 | 0.066 | 0.075 | 0.085 | 0.098 | 0.119 | 0.137 | 0.154 | 0.168 | **0.98** |
| ランク 4 の文節割合 | 0.046 | 0.047 | 0.050 | 0.053 | 0.057 | 0.057 | 0.051 | 0.054 | 0.093 | 0.69 |
| ランク 5 の文節割合 | 0.050 | 0.044 | 0.046 | 0.051 | 0.057 | 0.062 | 0.068 | 0.083 | 0.080 | 0.93 |
| ランク 6 の文節割合 | 0.137 | 0.111 | 0.107 | 0.113 | 0.121 | 0.122 | 0.127 | 0.155 | 0.100 | 0.09 |
| ランク 2 以上の文節割合 | 0.394 | 0.355 | 0.367 | 0.397 | 0.443 | 0.489 | 0.534 | 0.598 | 0.595 | 0.94 |
| ランク 3 以上の文節割合 | 0.301 | 0.268 | 0.279 | 0.303 | 0.333 | 0.359 | 0.383 | 0.445 | 0.442 | 0.94 |
| ランク 4 以上の文節割合 | 0.233 | 0.202 | 0.203 | 0.218 | 0.235 | 0.240 | 0.247 | 0.291 | 0.273 | 0.83 |
| ランク 5 以上の文節割合 | 0.187 | 0.155 | 0.153 | 0.164 | 0.179 | 0.184 | 0.195 | 0.237 | 0.181 | 0.58 |