# 均衡コーパスを規範とするテキスト難易度測定

名古屋大学大学院工学研究科教授 **佐藤　理史**

PROFILE

京都大学大学院工学研究科電気工学第二専攻博士課程研究指導認定退学。博士（工学）。北陸先端科学技術大学院大学、京都大学を経て、2005年６月より名古屋大学大学院工学研究科電子情報システム専攻教授。

## 1 はじめに

情報伝達を目的としたテキストは、平易でわかりやすいことが望ましい。そのようなテキストの作成には、テキストの難易度を簡単に測定できるツールが不可欠である。このような考えに立ち、我々は、日本語テキスト難易度測定システム「帯」を開発している [1,2]。

最新の公開バージョン obi-2.30 に含まれている難易度モデル T13 は、教科書コーパスに基づいて、13 段階の学年区分 ( 小学 1 ～ 6 年、中学 1 ～ 3 年、高校 1 ～ 3 年、大学 ) に対応する難易度を出力する。学年区分に対応する難易度は、教育応用においては直接的で都合が良い。しかし、それ以外の用途においては、難易度の直感的理解に欠けるという欠点がある。

この問題を解決するために、我々は、**均衡コーパスを規範とした難易度測定**という新たな方法を考案した [3]。本稿では、その方法について述べる。

**均衡コーパス**とは、「ある言語の特徴や性質を知るために，その言語の多様性をできるだけ忠実に反映するようにバランスよくサンプルを収集して構築される言語資源」を意味する [4]。国立国語研究所は、日本語の最初の均衡コーパスとなる、「現代日本語書き言葉均衡コーパス (BCCWJ)」を構築し、2011 年 10 月に正式に公開するとアナウンスしている。

均衡コーパスは、難易度という観点においても、日本語という母集団の多様性を反映し、かつ、代表性を有すると期待できる。言い換えるならば、均衡コーパスは、平易なテキストから難しいテキストまで、各種の難易度のテキストを、日本語テキスト全体（母集団）の分布と同じような分布で含んでいることが期待できるということである。このような均衡コーパスに対して、難易度を知りたいテキストを適切に位置付けることができれば、たとえば、「そのテキストは、日本語のテキストの平均的な難易度よりも、すこしばかり難しい」といった評価が可能となる。

## 2 難易度と難易度スケール

テキストの難易度は、実は、仮想的な概念である。それぞれのモノが固有の質量を持つように、それぞれのテキストは固有の難易度を持つわけではない。後者は、単なる仮定である。我々は、テキストを読んで「このテキストは難解だ」、「このテキストは平易だ」と感じるわけだが、そのような素朴な感想と「それぞれのテキストに固有の難易度が実在する」ことの間には、大きな隔たりがある。

テキストの難易度とは、やさしさ・難しさという軸において、テキストをいくつかのグループに分けるために導入した仮想的な概念であり、仮想的なスケールである。その実在は、仮定されるものであって証明されるものではない。当然のことながら、それぞれのテキストに対して、真の難易度という値が実在するわけではない。

では、いったい、なぜ難易度測定システムを実現しようとするのかというと、難易度という人工的なスケールを仮定し、その値に基づいてテキストをグループ化することが役に立つからである。難易度測定研究の目的は、

役に立つ難易度スケールとその測定法をセットとして提供することにある。

教科書コーパスに基づく測定システム obi2/T13 では、学年区分というスケールを難易度スケールとして借用した。これは、(1) 教科書という形で各難易度のテキストが入手可能なこと、および、(2) 教育応用では学年区分というスケールが直接的でわかりやすいこと、がその理由である。このようなスケール以外にも、たとえば、日本語能力試験の N1 から N5 に対応するスケールを、難易度スケールとして採用することもできる。

その一方で、テキストの分布に基づくスケールを考えることができる。すなわち、難易度の値を、対象とするテキスト集合（あるいは、日本語テキストの全集合）の難易度分布のどのあたりに位置するか、ということに対応させるのである。この方法を実現するためには、(1) 対象テキスト集合の縮図となる適切な部分集合が入手可能であり、かつ、(2) それらに何らかの難易度が付与されている（あるいは、難易度順にソートされている）ことが必要である。これら２つの条件のうち、前者は、均衡コーパスがその条件を満たす。後者の条件、つまり、均衡コーパスに難易度を付与することができれば、テキストの分布に基づくスケールを実現することが可能となる。

## 3 均衡コーパスを規範とする難易度測定

### 3.1 正規分布に基づく９段階の難易度

さて、初心に帰って、我々が感じる難易度とはどのようなものか考えてみよう。ありそうなシナリオは、次のようなものである。

(1) 我々は、毎日、多くのテキストに接している。

(2) そのような状況で、難易度の平均的な値というものが意識されるようになる。

(3) その値を規準として、そこから逸脱したテキストを、やさしい、あるいは、難しいと感じるようになる。

つまり、我々が感じる難易度は、我々の日々の言語経験に基づいて形成され、意識されるのではないか、ということである。

このような考えに基づくとき、均衡コーパスにおける難易度の値の分布は、正規分布に従うと仮定するのが自然である。すなわち、難易度の平均的な値を持つテキストは多数存在し、そこから逸脱する値を持つテキストは、その逸脱度に応じて急速にその数が少なくなる、ということである。これは、まったくの仮説にすぎないが、そう仮定することに不自然さはない。

次に考えるべきことは、難易度を何段階に区分するかということである。詳しい議論は文献 [3] にゆずるが、現実的にはそれほど区分数を増やすことができない。最終的に、我々が採用したのは、stanine に基づく９段階の難易度スケールである（表 1）。このスケールでは、最もやさしい 4% を難易度 1、最も難しい 4% を難易度 9、真ん中 20% を難易度 5 と定める。

### 3.2 均衡コーパスに対する難易度付与

次に、上記で定めた９段階の難易度を均衡コーパスの各テキストに付与することを考える。これには、次のような方法を採用した。

(1) なんらかの難易度付きコーパスを準備する（我々は、教科書コーパスを利用した）。

(2) 難易度付きコーパスを用いて、**比較器**を構成する。

(3) 構成した比較器を用いて均衡コーパスを難易度順にソートし [5]、stanine の比率に基づいて、各テキストに難易度を付与する。

(4) (2) へ戻る

ここで、比較器とは、２つのテキストを与えたとき、

表 1 Stanine に従う難易度スケール

| 難易度 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 割合 | 0.04 | 0.07 | 0.12 | 0.17 | 0.20 | 0.17 | 0.12 | 0.07 | 0.04 |
| 下限 | 0.00 | 0.04 | 0.11 | 0.23 | 0.40 | 0.60 | 0.77 | 0.89 | 0.96 |
| 上限 | 0.04 | 0.11 | 0.23 | 0.40 | 0.60 | 0.77 | 0.89 | 0.96 | 1.00 |

どちらが難しいかを返す判定器を意味する。このような比較器は、訓練例（どちらが難しいか判明しているテキストの組）から学習によって構成することができる。ただし、学習によって構成される比較器は 100% 信頼できるものとはならないので、上記の (2) と (3) を何回か繰り返し、安定したところで終了する。

上記の方法を、均衡コーパス（「現代日本語書き言葉均衡コーパス」モニター公開データ (2009 年度版 ) に含まれる書籍 10,102 サンプル）に対して適用し、最終的に、すべてのサンプルに、stanine に従う９段階の難易度を付与した。詳細な手順は、文献 [3] を参照されたい。

得られた難易度付き均衡コーパスから、3600 サンプルを選び、これを規準コーパスとして obi2 の難易度モデルを構成したものを、obi2/B9 と呼ぶ。これが、我々が作成した、均衡コーパスを規範とする難易度測定システムである。

## 4 新書コーパスに対する難易度測定

ここでは，難易度測定システム obi2/B9 を用いて、実際にテキストの難易度を測定した結果について述べる。測定対象としては、表３に示す新書コーパスを用いた。このコーパスは、新書 12 冊と一般書籍２冊の計 14 冊の本から抽出した、総計 148 のテキストサンプルから構成されている。各サンプルは、それぞれの本においてページ番号が 20 の倍数（S2007pro は 10 の倍数）となるページの見開きから、段落単位で抽出した。サンプルのサイズは、おおよそ 1000 字である。S2007pro と S2007cre は新書ではないが、それぞれ、新書より確実にやさしい本、難しい本という位置付けで、このコーパスに含めてある。

実際の難易度測定では、すべてのサンプルを、難易度測定システム obi2/B9 で測定した後、それらの結果をそれぞれの本毎に集計する。この集計では、測定で得られた難易度の平均値を求めるが、最小値と最大値をはずれ値として除外して計算する。得られた値を表２の最右欄に示す。この表では、14 冊の本を平均値の小さい順に並べてある。

新書 12 冊の各サンプルの難易値は、ほとんどが５から７である。つまり、これらのテキストは、平均的な難易度、あるいは、それより少し難しい難易度を持つと判定された。これは、我々の直感と一致する。また、S2007pro は最もやさしいと、S2007cre は最も難しいと判定された。

表２　新書コーパスに対する難易度測定結果

| ID | 書名 | 著者 | 出版社 | 難易度 |
|---|---|---|---|---|
| S2007pro | 世界一やさしい問題解決の授業 | 渡辺健介 | ダイヤモンド社 | 4.6 |
| S2003bak | バカの壁 | 養老孟司 | 新潮新書 | 5.1 |
| S2004com | コミュニケーション力 | 齋藤孝 | 岩波新書 | 5.2 |
| S2005sao | さおだけ屋はなぜ潰れないのか？ | 山田真哉 | 光文社新書 | 5.6 |
| S1999ren | 日本語練習帳 | 大野晋 | 岩波新書 | 5.8 |
| S2005kok | 国家の品格 | 藤原正彦 | 新潮新書 | 5.9 |
| S1981sak | 理科系の作文技術 | 木下是雄 | 中公新書 | 6.3 |
| S1993sei | 「超」整理法 | 野口悠紀雄 | 中公新書 | 6.3 |
| S2005kar | 下流社会 | 三浦展 | 光文社新書 | 6.3 |
| S2006goo | Google 既存のビジネスを破壊する | 佐々木俊尚 | 文春新書 | 6.4 |
| S2006tan | 他人を見下す若者たち | 速水敏彦 | 講談社現代新書 | 6.5 |
| S2006web | ウェブ進化論 | 梅田望夫 | ちくま新書 | 6.8 |
| S1995net | インターネット | 村井純 | 岩波新書 | 7.0 |
| S2007cre | 創造的想像力 [ 増補版 ] | マイケル・ポラニー | ハーベスト社 | 7.2 |

## 5 BCCWJコアの難易度測定

「現代日本語書き言葉均衡コーパス (BCCWJ)」には、コアと呼ばれる部分集合が存在する。このコアに含まれる 571 の固定長サンプルの難易度を測定した結果を、表 3 に示す。

571 のサンプルは、書籍 83 サンプル、新聞 340 サンプル、雑誌 86 サンプル、白書 62 サンプルから構成されている。書籍の難易度は、難易度 1 から難易度 9 に渡って、正規分布に近い形で分布する。これは設計どおりである。これに対して、新聞の難易度は、大半が難易度 5 から 7 に集中する。つまり、新聞の難易度は、さきほどの新書の難易度と同じようなレベルにあるということである。雑誌の難易度は、比較的広く分布し、新聞より少しやさしい方にかたよる。最も難易度がかたよるのは白書の難易度で、すべてのサンプルが難易度 8 または 9 と判定された。

## 6 おわりに

本稿では、均衡コーパスを規範とした難易度測定法について述べた。この方法は、均衡コーパスに対して、難易度の値が正規分布に従うように設定したスケールで、難易度を測定する新しい方法である。先に述べたように、「現代日本語書き言葉均衡コーパス」は、2011 年 10 月に正式公開の予定である。この公開後、我々は、正式公開版を用いて再度、難易度モデルを作り直し、obi2 システムのダウンロード版に含める予定である。

本方法の残された課題は、このような方法で測定される難易度が、人間が感じる難易度を十分に反映したものとなっているか、という懸念である。この疑問に回答を与えるために、現在、国立国語研究所と共同で、人間が感じる難易度の調査を進めており、予備的な段階では、肯定的な結果を得ている。この調査結果については、稿を改めて報告する予定である。

### 参考文献


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

[2] 小島健輔 , 佐藤理史 , 藤田篤 (2009). 文字 bigram モデルを用いた日本語テキストの難易度推定 . 言語処理学会第 15 回年次大会発表論文集 , pp.897-900.

[3] 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 , Vol. 52, No. 4, pp.1777-1789.

[4] 柏野和佳子 , 丸山岳彦 , 稲益佐知子 , 田中弥生 , 秋元祐哉 , 佐野大樹 , 大矢内夢子 , 山 崎 誠 (2009).『現代日本語書き言葉均衡コーパス』における収録テキストの抽出手順と事例 , 国立国語研究所内部報告書 LR-CCG-08-01, 国立国語研究所 .

[5] Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada (2010). Sorting by Readability. Computational Linguistics, Vol. 36, No. 2, pp.203-227.


表 3　BCCWJ コアの難易度測定結果

| 難易度 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 計 |
|---|---|---|---|---|---|---|---|---|---|---|
| 書籍 | **6** | **5** | **13** | **13** | **20** | **12** | **8** | **5** | **1** | 83 |
| 新聞 | | | 4 | 28 | 92 | 76 | 89 | 49 | 2 | 340 |
| 雑誌 | 2 | 2 | 8 | 24 | 15 | 15 | 16 | 4 | | 86 |
| 白書 | | | | | | | | 33 | 29 | 62 |
| 計 | 8 | 7 | 25 | 65 | 127 | 103 | 113 | 91 | 32 | 571 |