現代日本語書き言葉均衡コーパスに対する難易度付与

佐藤 理史　（名古屋大学大学院工学研究科）

Assigning Readability Score to Every Text in BCCWJ

Satoshi Sato　(Graduate School of Engineering, Nagoya University)

1　はじめに

我々は、これまで、『現代日本語書き言葉均衡コーパス (BCCWJ)』[1] の各サンプルテキストに難易度を付与すべく、準備を行なってきた。我々が開発した難易度測定システム『帯 2(obi2)』[1] は、あらかじめ用意した、難易度の規準となるコーパス (規準コーパス) に基づき、与えられたテキストの難易度を決定するシステムである。このシステムでは、まず、規準コーパスに含まれる、それぞれの難易度に対応するサブコーパスに対し、言語モデル (文字 bigram モデル) を作成する。次に、未知のテキストに対して、それぞれの言語モデルに対する尤度を計算し、最大の尤度をとる言語モデルに対応する難易度を出力する[2]。このように、本システムでは、規準コーパスが難易度スケールを規定する。

『帯 2』が最初に提供した難易度スケールは、13 段階の学年区分 (小学 1 年から高校 3 年、および、大学) に対応する obi2/T13 である。このスケールの規準コーパスには、各学年の教科書から抽出したテキストで構成する「教科書コーパス」[2] を用いている。これとは別に、均衡コーパスから規準コーパスを作成する方法を考案し、BCCWJ のモニター公開データ (2009 年度版) に含まれる書籍データ 10,423 サンプルから、相対的難易度を表す難易度スケール obi2/B9 を作成した [3]。その後、31 名の被験者実験の結果との比較により、obi2/B9 の難易判定能力が、平均的な人間と同程度であることを明らかにした [4]。

今回、2012 年 1 月に 2 枚組 DVD として正式にリリースされた『現代日本語書き言葉均衡コーパス』(以下、BCCWJ リリース版) に含まれる、書籍の可変長サンプル 18,000 件を用いて、難易度スケール obi2/B9 を再構成した。さらに、再構成した obi2/B9 を用いて、BCCWJ リリース版に含まれる全サンプルに難易度を付与した。本稿では、これらの内容について報告する。

2　難易度スケール obi2/B9 の再構成

本節では、難易度スケール obi2/B9 の再構成について、前回 [3] との差分を中心に述べる。

2.1　使用したサンプル

今回の obi2/B9 の再構成には、BCCWJ リリース版に含まれる、出版サブコーパスの書籍 (PB) および図書館サブコーパスの書籍 (LB) の可変長サンプル、全 20,688 サンプルのうち、有効 bigram 数が多い 18,000 サンプルを利用した。有効 bigram とは、『帯 2』が難易度計算のために使用する可能性がある文字 bigram (連続する 2 文字) のことで、2 つの文字は、いずれも、ひらがな、カタカナ、JIS 漢字第 1 水準のいずれかである。使用するサンプル数は、前回の反省から、1000 で割り切れる数とした。最も有効 bigram 数が少ないサンプルは、1,074 個の有効 bigram を含む。

2.2　難易度スケール B9 の構成手順

難易度スケール obi2/B9 の構成手順は、おおむね、前回の構成手順 [3] を踏襲した。具体的手順を以下に示す。

[1] http://kotoba.nuee.nagoya-u.ac.jp
[2] 実際の計算は、スムージングを適用するため、もう少し複雑である。

表 1: Stanine

stanine	1	2	3	4	5	6	7	8	9
割合	4%	7%	12%	17%	20%	17%	12%	7%	4%
範囲	0–4%	4–11%	11–23%	23–40%	40–60%	60–77%	77–89%	89–96%	96–100%

(1) 難易度付きコーパス B_0 を用意する。
(2) $i = 1$
(3) 難易度付きコーパス B_{i-1} を用いて、与えられた 2 つのテキストの難易を判定する比較器を構成する。
(4) 構成した比較器を用いて B_{i-1} に含まれるサンプルを難易順にソートする。
(5) ソート結果に基づき、B_{i-1} の各サンプル (テキスト) に 9 段階の難易度を付与する。難易度の決定には、stanine (後述) を用いる。
(6) こうして得られた 9 段階の難易度コーパスから、その一部 (各難易度に対して同数のサンプル) を取り出す。これを規準コーパス C_i として、obi2/B9$_i$ を構成する。
(7) obi2/B9$_i$ を用いて、コーパス B_{i-1} の各テキストに難易度を付与する。この結果として得られる難易度付きコーパスを B_i とする。
(8) i=$i+1$ として、(3) へ。

実際には、この手続きを、$i = 7$ まで実行した。上記の手続きに出てくる stanine は、ソートされたリストの各要素に、表 1 に示す割合に従って、1 から 9 の整数を割り当てる方法である。整数は、難易度が高いものほど値が大きくなる方向で割り当てる。

今回の obi/b9 の再構成においては、一貫して、前述の 18,000 サンプルを用いた。すなわち、上記の B_i は、すべて同一の 18,000 サンプルから構成されている。添字の違いは、各サンプルに付与されている難易度の値の違いのみである。

以下に、上記の手続きの各ステップの詳細を示す。

難易度付きコーパス B_0 の作成 18,000 サンプルに対して、前回作成した (すでに存在する) obi2/B9 を用いて、9 段階の難易度を付与した。これを、難易度付きコーパス B_0 として使用した。

比較器の構成 前回と同一の手順で、比較器を SVM として構成した。ただし、頻度パラメータ f_c (SVM の属性として使用する、bigram の最低頻度) は、$f_c = 100$ を用いた。

コーパスのソートと難易度付与 コーパスのソートは、18,000 サンプルを 18 個のパーティションに分け、パーティションごとにソートする方法を採用した[3]。stanine に基づく難易度付与も、パーティションごとに行なう。この結果、9 段階の難易度が付与された 1,000 サンプルが、18 パーティション分、得られる。

規準コーパスの作成と難易度スケールの構成 1 つのパーティションから、1 つの難易度に対して、40 サンプルを選ぶ[4]。この結果、各難易度に対して 720 サンプル、全部で 6,480 サンプルが選ばれる。これを規準コーパス C_i として obi2/B9$_i$ を構成する。このとき、難易度の測定に使用する bigram の最低頻度 f を $f = 100$ とした。

[3] 今回使用した計算機では、このソートに約 1 週間かかる。
[4] 難易度 1 および 9 の割合は 4%なので、1 つのパーティションに 40 サンプルずつしかない。

表 2: 難易度付きコーパスの変化

	R	RMSE	s_0	s_1
B_0–B_1	0.974	0.504	0.763	0.997
B_1–B_2	0.985	0.350	0.888	0.998
B_2–B_3	0.987	0.326	0.899	0.999
B_3–B_4	0.986	0.334	0.898	0.998
B_4–B_5	0.987	0.315	0.911	0.998
B_5–B_6	0.988	0.304	0.914	0.998
B_6–B_7	0.989	0.295	0.919	0.998
B_7–$B_{7(f=75)}$	0.997	0.165	0.975	0.999
B_0–$B_{7(f=75)}$	0.958	0.631	0.666	0.987

コーパスに対する難易度の再付与 こうして得られた obi2/B9$_i$ を用いて、18,000 サンプルに難易度を付与し直す。この結果、新たな難易度付きコーパス B_i が得られる。

2.3 難易度付きコーパスの変化と 2 分割交差検定の変化

ここでは、上記の手続きの繰り返しの過程で、難易度付きコーパス B_i の各難易度がどのように変化したかに着目する。この変化は、2 つのコーパスの各難易度の値を、同じサンプルに対する添字が同一になるように並べたのち、2 つの列 $A = \{a_j\}$ と $B = \{b_j\}$ $(j = 1, 2, \ldots, n)$ を比較することによって得られる。この比較には、次の 4 つの指標を利用する。

$$\text{相関係数 } R(A,B) = \frac{\sum_{j=1}^{n}(a_j - \bar{a})(b_j - \bar{b})}{\sqrt{\sum_{j=1}^{n}(a_j - \bar{a})^2}\sqrt{\sum_{j=1}^{n}(b_j - \bar{b})^2}} \tag{1}$$

$$\text{root mean square error } RMSE(A,B) = \sqrt{\frac{(a_j - b_j)^2}{n}} \tag{2}$$

$$\text{一致率 } s_0(A,B) = \frac{1}{n}\sum_{j=1}^{n} d(a_j, b_j, 0) \tag{3}$$

$$\text{差 1 を許容した一致率 } s_1(A,B) = \frac{1}{n}\sum_{j=1}^{n} d(a_j, b_j, 1) \tag{4}$$

ここで、関数 $d(a_j, b_j, v)$ は、a_j と b_j の差が v 以下かどうか調べる関数で、次のように定義する。

$$d(a,b,v) = \begin{cases} 1 & \text{when } |a-b| \le v \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

上記の 4 つの指標のうち、RMSE を除く 3 つの指標は、列 A と B が似ているほど大きな値をとる。RMSE は、逆に、列 A と B が似ているほど小さな値をとる。

表 2 に、コーパス B_{i-1} と B_i の変化を上記の 4 つの指標で計測した結果を示す。この表より、コーパス B_{i-1} と B_i の差は、かなり小さいことがわかる。B_0 と B_1 の差が最も大きいが、それでも相関係数 $R = 0.974$、一致率 $s_0 = 76.3\%$、差 1 を許容した一致率 $s_1 = 99.7\%$ である。最後の 2 つのコーパス B_6 と B_7 の一致率 s_0 は、91.9%である。これは、最後の 1 回のループにおいて難易度の値が変化したサンプルは、全体の 8.1%にすぎないことを意味する。

表 3 に、C_i を規準コーパスとする obi2/B9$_i$ の 2 分割交差検定の結果を示す。この表より、2 分割交差検定の評価値の変化は小さいことがわかる。ただ、$i = 1$ と $i = 7$ の評価値を比較すると、いず

表 3: obi2/B9$_i$ の 2 分割交差検定

obi2/B9$_i$	R	RMSE	s_0	s_1
1	0.976	0.568	0.700	0.994
2	0.978	0.544	0.725	0.994
3	0.976	0.564	0.711	0.992
4	0.978	0.542	0.718	0.996
5	0.977	0.551	0.721	0.994
6	0.977	0.554	0.714	0.995
7	0.978	0.542	0.724	0.996
7($f = 75$)	0.978	0.538	0.726	0.996

れの評価値の向上している。このことは、前述の手順による繰返しにより、より一貫した難易度が付与される方向に動いていることを、間接的に示している。

2.4 最終的に採用した難易度スケール

最終的に採用した難易度スケールは、規準コーパス C_7 から作成した。ただし、難易度の測定に使用する bigram の最低頻度 f の値は、$f = 100$ ではなく、$f = 75$ を採用した。表 3 の最終行に、$f = 75$ の場合の 2 分割交差検定の評価値を示す。この表に示すように、$f = 75$ の評価値は、$f = 100$ と比べ、RMSE と s_0 において、若干良い。

なお、表 2 に示すように、18,000 サンプルに対して、$f = 75$ の obi2/B9$_7$ で難易度を付与した場合 ($B_7(f = 75)$) と、$f = 100$ の obi2/B9$_7$ で難易度を付与した場合 (B_7) の差は、非常に小さい。事実、97.5%のサンプルで、難易度の値は同一である。

最終的に採用した、新しい obi2/B9 は、33,360 種類の有効 bigram に基づいて難易度を決定する。なお、以前の obi2/B9 が利用する有効 bigram は、29,017 種類であった。

表 2 の最終行に、18,000 サンプルに、以前の obi2/B9 によって難易度を付与した場合 (B_0) と、最終的に採用した新たな obi2/B9 によって難易度を付与した場合 ($B_7(f = 75)$) の差を示す。この表の他の値と比較すると、これら 2 つの難易度付きコーパスの差は大きいように見えるが、実際には、66.6%のサンプルで難易度の値が一致しており、値に差があるものも、そのほとんど (98.7%) が ± 1 の範囲にある。つまり、BCCWJ のリリース版を用いて作り直した obi2/B9 は、以前の obi2/B9 とそれほど大きな差はない。

2.5 被験者実験との比較

以前行なった被験者実験との比較 [4] と全く同じことを、新たに構成した obi2/B9 を用いて行なった。この文献 [4] の表 2 の末尾に、新たな obi2/B9 のデータ (nB9) を追加したものを表 4 に示す。3 つの指標 d_0, d_c, d_i の値は、以前の B9 (oB9) より減少している。これらの値は、いずれも人間の多数派の回答とのハミング距離 (差の数) を表しており、その値の減少は、人間の多数派の回答に、より近づいていることを意味する。

3 BCCWJ に対する難易度付与

新たに構成した obi2/B9 を用いて、BCCWJ リリース版に含まれる全サンプルに対して、難易度を付与した。

3.1 固定長サンプルに対する難易度付与

表 5 に、固定長サンプルに対する難易度付与の結果を示す。図 1 は、白書 (OW) を除く各レジスタの難易度分布を、棒グラフ化したものである。このグラフの縦軸は、全体に対する割合を示してお

表 4: 被験者実験との比較 (120 ビット 1 対比較コード)

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}	d_0	d_c	d_i
m	111000	101001	110000	111011	111111	001111	000011	101001	000110	000100	111011	000100	100001	000001	000000	111111	000100	010100	101001	000001			
ci	−+?+−+	−?+−+	++??++	−+−++−	+−−+−+	−+?+−+	−−+?++	−?+−++	−−+?+	??+?++	+++−??	−+−+?+	?−+−++	+−++??	+−−+−+	?++?+?	++−?++	++−+−	−++−+−	?+++??	120	96	62
p_{01}	.0...	.....	....1.	.....	.....	.....	.....	.....	.....	1..0..	.....	.....	.....	1....0	.....	0.....	...0..	.....	....1.	....1.	11	2	2
p_{02}	.0...	.1....	.....	.....	.....	...0..	....0.	.1....	.....	.....	....0.	.....	.....	....1.	1.....	0	.....	.....	.....	....1.0	11	4	4
p_{03}	.....	.1...0	.1....	.....	.....	...0..	...1..	.0....	.1....	.....	....0.	.....	0	.....	.....	00....	...0..	.....	.....	.....	12	6	6
p_{04}	.0....	.10..0	...1..	.....	0....0	.....	.....	.1....	.....	.....	00.0..	.....	.....	....1.	.....	0.....	.....	.....	.....	11.1.0	17	8	7
p_{05}	.....	.....	..11..	.....	.....	..00..	...1.0	.1....	.....	1..0..	.....	.....	.....	.....	.....	.....	...0.1	.....	.....	1.....	12	3	3
p_{06}	1	.1...0	..11..	.....	.....	.1....	...1..	.1....	.....	.1....	....0.	.....	.....	1.1.1.	.....	1.....1	0....0	.....	.1....	....1.	18	7	7
p_{07}	.....	.....	..11..	.....	.....	0.1....	...1..	.....	.....	.....	0..1..	.1..1.	.....	....0.	.....	.....	...0..	1.....	.....	1.1.1.	15	7	7
p_{08}	.....	..0...	.0...1	.....	.....	.0..0	1.1.0.	..0...	....0.	11....	000...	...0..	0.....	...1.0	11....	000...	.1..1.	.....	0.0...	....1.	30	22	18
p_{09}	.00...	.1...0	.0....	.0..0.	.....	.1....	....0.	..0...	.....	.1....	....0.	.1....	0.....	..1.1.	...1..	.....	.....	.....	....1.	.....	18	12	12
p_{10}	.....	..0...	..1..1	.....	.....	.0..0	.11..	.....	....0.	11.0..	...1..	...0..	0.....	.111.	1.....	...00.	.....	.....	0.0...	11.1.0	26	13	12
p_{11}	.0....	....1.	..1111	0..0.	.....	..0...	...1..	.1....	.....	11....	.....	.0.1.0	0.....	0	...1..	...00.	.....	.....	.0....	...110	25	11	11
p_{12}	.01..	0...1.	.....	.....	.....	.....	.....	.....	.....	.....	...1..	...0..	.....	..1.1	.....	...000	.....	.....	.....	1.1...	17	7	6
p_{13}	...1..	.1....	.....	.....	.....	.....	00	.1...0	.1..0.	.....	....0.	.....	0	.....	.....	.....	.0....	.....	.....	1....0	13	7	7
p_{14}	.....	..0...	...1..	.0....	.....	.0.....	111...	.1....	....0.	.....	....0.	.1....	.....	....1.	.....	.....	.....	.....	.....	0	14	8	6
p_{15}	.0....	.1...0	.....	...1..	0	11....	...100	.0..0	.....	.....	...1.0	....1.	0.....	.....	...1..	.0.0..	...0..	.....	.0....	1.1...	23	14	13
p_{16}	.....	.....	00...1	.....	.....	.0..0	1...00	.1....	.....	1.0..	00	...0..	0.....	.....	11....	...0..	.....	.....	.....	.1.1.	20	12	10
p_{17}	.0....	.1....	.0....	.....	.....	.....	.....	....1.	.1....	.....	....0.	....1.	.....	.....	....1.	.....	.....	.....	....1.	.....	9	5	5
p_{18}	.....	.1...0	..1...	.....	.....	....0.	.....	.1....	.....	.....	.....	....1.	.....	....1.	....1.	.....	....1.	.0.0..	....1.	1.1.1.	16	8	7
p_{19}	.....	.1....	0.1...	...1..	0.....	.....	...1..	1...0	.....	1..0..	...1..	.0.1.0	0.....	.1.1.	1.....	.00000	...0..	.....	1....	0	25	12	12
p_{20}	.....	.....	0..11.	....0.	0.....	.....	.0000	111.00	..0...	.11...	11....	0001..	.1..1.	.....	...1.1	1...1	0....0	...0..	.....	...1.0	33	22	17
p_{21}	.....	.....	00.1..	...1..	.....	.....	.....	1	.....	...0..	...1..	....1.	.....	.1.1.	.....	.....	.....	.0....	.....	1....0	14	7	7
p_{22}	.....	..0..0	..11.	....0.	.....	110000	...1..	..0...	.....	11....	.....	....1.	.....	0	.1.1..	...00.	...0..	.....	.....	.....	24	13	10
p_{23}	.....	.1....	..1..1	.....	.....	...0..	...1..	.1....	....0.	.1....	.....	.....	.....	.....	.....	0	.....	.....	....1.	0	11	3	3
p_{24}	.00...	....1.	..11..	.....	0...0	.....	..11..	1.1.0	....0.	...0..	.0..0.	.1..1.	.....	0	.1....	0.....	.....	1....	.....	..111.	24	12	11
p_{25}	..11.	..0...	..1...	.....	000.00	11...0	...1..	.....	....0.	...0..	...1..	...0.1	0.....	0	1.1.11	000...	.....	.0....	.....	1.1.1.	31	21	14
p_{26}	.....	..0...	.0....	.....	0	..0..0	.....	.1....	.....	...0..	.....	.....	0.....	1.....	1....1	0....0	.1....	.....	.....	.....	14	8	8
p_{27}	.00...	.1...0	00...1	.....	0.....	.0....	...11.	.1...0	1.....	11....	...1..	.....	.....	.....	.....	000..0	...0..	.....	.....	...11.	26	12	12
p_{28}	.....	.....	..1...	.....	.....	.0..0	.....	.0....	.....	...0..	.....	....1.	0.....	.....	.....	...000	...0..	.0....	.....	1.....	14	5	5
p_{29}	.00...	.....	..11..	.....	.....	.....	.....	.....	....0.	11....	...1.0	.....	.....	..1.1.	.....	...0..	.....	.0....	.....	11...0	16	5	5
p_{30}	.....	.1....	..11..	.....	0.....	..00..	...1..	.0....	....0.	.....	0..0..	.1....	0...0	1...0	.....	0...0	...0..	1.....	.....	1....0	22	8	8
p_{31}	...1..	.1...0	.0.....	....0.	.....	.....	.....	.....	.....	.....	...1.0	...0.1	0.....	111..0	1....1	...000	...011	1.....	.....	.....	23	16	15
av.																					18.8	9.7	8.7
oB9	.....	.?....	.0...?	....0?	0.....	.?....	....?.	.....	....?.	?..0..	....0.	.1....	0....0	.....	.....	....?.	.....	.?..?.	.....	....?.	14.5	8.5	8.0
nB9	.?...1	.?....	.0...?	....?.	0.....	.?....	..??..	.?....	.....	....?.	....0.	.1....	0....?	....?.	.....	.....	.....	.....	....?.	....?.	12.5	7.5	7.5

表 5: 固定長サンプルに対する難易度付与結果

サブコーパス-媒体	1	2	3	4	5	6	7	8	9	計	平均	分散
出版-書籍 (PB)	321	567	1299	1433	1976	1284	1298	1205	734	10117	5.34	4.43
出版-雑誌 (PM)	97	61	227	555	622	172	181	75	6	1996	4.60	2.46
出版-新聞 (PN)	5	0	20	218	488	392	260	88	2	1473	5.62	1.41
図書館-書籍 (LB)	536	835	1874	1920	2404	1533	911	315	223	10551	4.51	3.28
特殊目的-白書 (OW)	0	0	0	0	2	8	143	935	412	1500	8.16	0.38
書籍 (PB+LB)	857	1402	3173	3353	4380	2817	2209	1520	957	20668	4.92	4.01
stanine (S9)	4	7	12	17	20	17	12	7	4	100	5.00	3.84

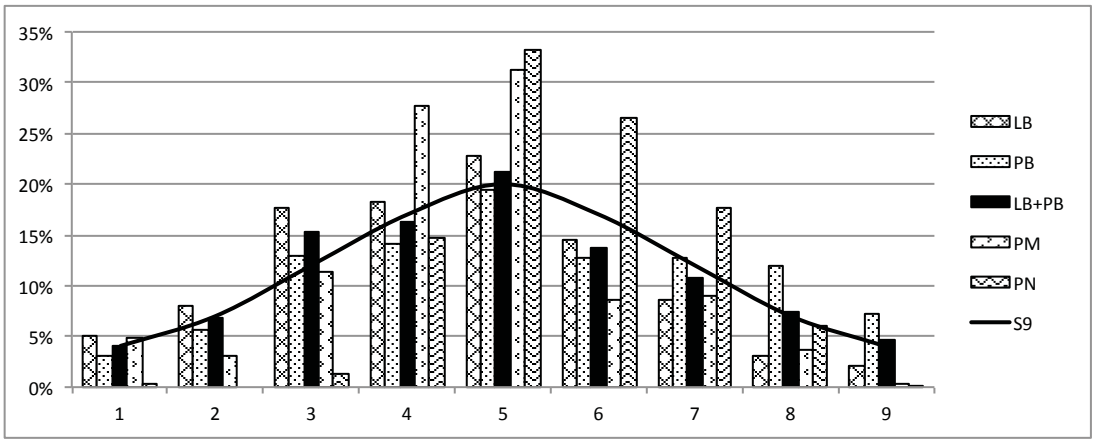

図 1: 固定長サンプル (LB, PB, PB+LB, PM, PN) の難易度分布

り、曲線 (S9) は、stanine の分布を示している。

このグラフより、書籍全体 (PB+LB) の難易度分布は、ほぼ、stanine の分布に従っていることがわかる。これは、難易度分布が、ほぼ設計通りとなっていることを示している。出版-書籍 (PB) と図書館-書籍 (LB) の 2 つのレジスタを比較すると、図書館-書籍 (LB) の方が難易度の平均が低く、分散も小さい。つまり、「図書館サブコーパスの書籍 (LB) は、出版サブコーパスの書籍 (PB) に比べ、難易度が低い方に偏っている」ということである。図書館-書籍 (LB) の母集団は、都内 13 自治体以上の図書館が共通して所蔵している書籍であるのに対し、出版-書籍 (PB) の母集団は、2001 年から 2005 年の間に国内で出版された書籍である [5]。2 つのレジスタの難易度の分布の違いは、このような母集団の違いによるものと考えるのが妥当である。この点については、後で再度、議論する。

雑誌 (PM) および新聞 (PN) は、書籍 (PB や LB) と比較する分散が小さい。つまり、比較的難易度が揃っている。雑誌 (PM) のサンプルの 59%は、難易度が 4 または 5 である。一方、新聞 (PN) は、サンプルの 60%が難易度が 5 または 6 であり、サンプルの 92%が難易度 4 から 7 の範囲である。これらの難易度の集中は、新聞・雑誌の編集において、難易度がコントロールされていることの現れと考えることができる。なお、雑誌 (PM) と新聞 (PN) では、雑誌の方が難易度の平均が低く、分散が大きい。このことは、我々が、日常生活において感じている印象と一致する。

図 1 のグラフには、白書 (OW) を含めなかった。白書のほとんどのサンプルは、難易度 7 から 9 であり、難易度の平均値が非常に高く、かつ、分散は小さい。

3.2 可変長サンプルに対する難易度付与

固定長サンプルと同様に、すべての可変長サンプルに対しても難易度を付与した。ただし、可変長サンプルの長さはさまざまで、1000 字未満のサンプルや、有効 bigram を一つも含まないサンプルも

表 6: 可変長サンプルに対する難易度付与結果 (有効 bigram 数が 300 以上のサンプル)

サブコーパス-媒体	1	2	3	4	5	6	7	8	9	計	平均	分散
出版-書籍 (PB)	288	534	1341	1369	1933	1290	1251	1181	717	9904	5.34	4.38
出版-雑誌 (PM)	100	58	209	559	606	166	171	65	4	1938	4.57	2.40
出版-新聞 (PN)	6	5	25	211	333	247	177	80	13	1097	5.54	1.84
図書館-書籍 (LB)	490	727	1993	1844	2396	1519	881	305	215	10370	4.53	3.19
白書 (OW)	0	0	0	0	2	11	111	969	397	1490	8.17	0.35
教科書 (OT)	47	2	7	22	85	69	125	29	0	386	5.46	4.07
広報紙 (OP)	0	0	0	0	27	90	162	73	2	354	6.81	0.75
ベストセラー (OB)	24	144	405	359	280	111	35	6	3	1367	3.91	1.75
Yahoo!知恵袋 (OC)	145	290	2945	2995	1596	309	290	113	111	8794	4.01	1.75
Yahoo!ブログ (OY)	887	1753	3825	3421	2063	561	421	142	44	13117	3.64	2.18
韻文 (OV)	69	16	104	58	5	0	0	0	0	252	2.66	1.35
法律 (OL)	0	0	0	0	0	0	0	0	334	334	9.00	0.00
国会会議録 (OM)	0	0	0	0	53	54	19	12	21	159	6.33	1.83
書籍 (PB+LB)	778	1261	3334	3213	4329	2809	2132	1486	932	20274	4.93	3.94

サブコーパスを明示していないものは、特殊目的サブンコーパスに属する。

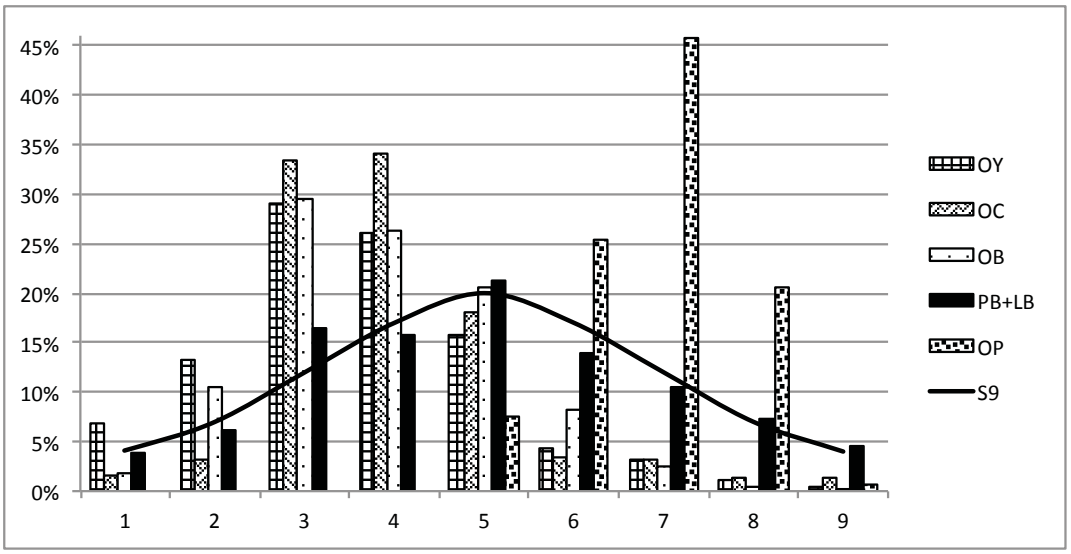

図 2: 可変長サンプル (OY, OC, OB, PB+LB, OP) の難易度分布

存在する[5]。『帯 2』が難易度を安定して決定するためには、少なくとも数百の有効 bigram が必要である。そのため、難易度分布の分析では、難易度を付与したすべてのサンプルを対象とするのではなく、サンプルサイズに対して適当な閾値を設定し、閾値以上のサンプルのみを対象とするのが適切である。

表 6 に、有効 bigram 数が 300 以上のサンプルにおける難易度分布を示す。可変長サンプルのレジスタのうち、固定長サンプルに含まれる出版-書籍 (PB)、雑誌 (PM)、新聞 (PN)、図書館-書籍 (LB)、白書 (OW) については、固定長サンプルと同じ傾向であるので、以下では議論しない。また、韻文 (OV)、法律 (OL)、国会議事録 (OM)、教科書 (OT) の各レジスタは、その特殊性を勘案し、ここでは議論の対象としない。以下では、広報紙 (OP)、ベストセラー (OB)、Yahoo!知恵袋 (OC)、Yahoo!ブログ (OY) について議論する。図 2 に、これらのレジスタの難易度分布の棒グラフを示す。なお、比較のため、書籍 (PB+LB) の難易度分布も合わせて示した。

[5] このような場合、『帯 2』は難易度判定不能として、0 を出力する。

Yahoo!知恵袋 (OC) と Yahoo!ブログ (OY) は、短いサンプルが多く、前者は 9.6% (8794/91445)、後者は 24.9% (13117/52680) のみが、分析対象となっている。分析対象となったサンプルの半数以上が難易度 3 または 4 となる。この 2 つのレジスタの中では、Yahoo!ブログ (OY) の方が難易度の平均は低く、分散は大きい。

ベストセラー (OB) は、書籍全体 (PB+LB) と比較して、かなりやさしい方に偏る。ピークは難易度 3、これに難易度 4 を含めると、全体の 56%に達する。分散は 1.75 で、図書館-書籍 (LB) の分散 3.19 と比べて、かなり小さい。

広報紙 (OP) のピークは、難易度 7 であり、全体の 46%を占める。難易度 6 から 9 のサンプルを合わせると、全体の 91%を占める。ピークが難易度 8 にある白書 (OW) よりはやさしいが、新聞 (PN) と比較すると、平均難易度は高く、分散は小さい。

3.3 レジスタの難易度序列

固定長サンプルおよび可変長サンプルに付与された難易度の分布を総合し、検討の対象としたレジスタ群に難易度の序列を付けると、次のようになろう。

$$\mathrm{OY} < \mathrm{OC} < \mathrm{OB} < \mathrm{LB} < \mathrm{PM} < (\mathrm{PB+LB}) < \mathrm{PB} < \mathrm{PN} < \mathrm{OP} < \mathrm{OW}$$

この結果は、我々が日常的に感じている印象とほとんど矛盾しない。少しだけ意外なのは、広報紙 (OP) の難易度の高さであろうか。この点については、さらに調査を進める必要があるが、広報紙に含まれる、比較的長い漢字連続 (たとえば、「札幌市危機管理対策室」、「緊急輸送道路沿道建築物」、「医療保険年金課高齢者医療係」) が、obi2/B9 難易度を上げている可能性がある。

相対難易度の規準として、我々は、前回より、書籍 (PB+LB) のサンプルを用いている。これは、BCCWJ において、書籍のサンプルが、生産および流通の両側面において適切にサンプリングされており、均衡コーパスとして最もふさわしいという判断に基づく選択であった。今回の BCCWJ の全サンプルに対する難易度付与の結果、書籍 (PB+LB) の難易度の分散は、他のレジスタと比較して、相対的に大きいことが明らかになった。すなわち、難易度の値に多様性があるという観点においても、書籍レジスタを相対的難易度の規準としてを用いるという選択は、妥当である。

3.4 ジャンル別の難易度分布

BCCWJ の書誌情報データには、ジャンル情報が含まれている。ここでは、出版-書籍 (PB)、図書館-書籍 (LB)、ベストセラー (OB) の 3 つのレジスタに対し、ジャンル別に難易度を集計した。日本十進分類法 (NDC) の第 1 次区分に対する難易度の平均値および分散を表 7 に示す。この表の「割合」は、その区分の全体に占める割合を示す。なお、これらは、有効 bigram が 300 以上の可変長サンプルに対する集計結果である。

先に示したように、この 3 つのレジスタにおける難易度の序列は、PB>LB>OB である。表 7 に示すように、平均難易度が高い「3 社会科学」の割合は PB が最も大きく、逆に、平均難易度が低い「9 文学」の割合は OB が最も大きい。このことが、全体の平均難易度が PB>LB>OB となる、主に要因である。

しかし、その一方で、各区分別に平均難易度を比較すると、7 つの区分で、PB>LB>OB という序列が観察される。つまり、ジャンルを固定した場合でも、一般的傾向として、PB>LB>OB という難易度の序列が観察される。

BCCWJ の書誌情報データには、日本十進分類法 (NDC) の他に、「C コード (図書分類コード)」が付与されている。このコードの左から 1 桁目は「販売対象コード」で対象読者を表す。2 桁目は「発行形態コード」で、発行形態を表す。これらのコード別の難易度の平均値と分散を、表 8 および表 9

表 7: NDC 別の難易度分布 (PB, LB, OB)

	出版-書籍 (PB)			図書館-書籍 (LB)			ベストセラー (OB)			
	平均	分散	割合	平均	分散	割合	平均	分散	割合	P>L>O
0 総記	5.91	3.66	(0.032)	5.35	2.83	(0.023)	4.60	2.99	(0.029)	√
1 哲学	5.07	1.18	(0.054)	5.20	1.08	(0.049)	4.55	0.93	(0.102)	
2 歴史	5.10	1.27	(0.086)	5.10	1.05	(0.103)	4.46	0.60	(0.046)	
3 社会科学	6.76	3.13	(0.253)	5.84	2.82	(0.197)	5.29	1.88	(0.127)	√
4 自然科学	6.40	2.76	(0.103)	5.60	1.95	(0.061)	4.96	1.44	(0.020)	√
5 技術・工学	5.92	5.78	(0.091)	5.05	5.07	(0.061)	3.58	3.31	(0.031)	√
6 産業	6.06	3.35	(0.044)	5.46	2.57	(0.033)	4.67	0.89	(0.011)	√
7 芸術・美術	4.71	1.53	(0.064)	4.50	1.45	(0.078)	3.25	1.31	(0.070)	√
8 言語	5.44	1.71	(0.018)	5.02	1.55	(0.018)	4.88	0.36	(0.012)	√
9 文学	3.15	1.06	(0.214)	3.27	1.02	(0.333)	3.43	0.98	(0.530)	
分類なし	4.73	5.22	(0.043)	2.51	4.18	(0.043)	3.41	1.48	(0.021)	–
計	5.34	4.38	(1.000)	4.53	3.19	(1.000)	3.91	1.75	(1.000)	

表 8: 販売対象コード別の難易度分布 (PB, LB, OB)

	出版-書籍 (PB)			図書館-書籍 (LB)			ベストセラー (OB)			
	平均	分散	割合	平均	分散	割合	平均	分散	割合	P>L>O
0 一般	4.42	2.62	(0.544)	4.31	2.32	(0.737)	3.82	1.59	(0.640)	√
1 教養	6.30	2.56	(0.044)	5.74	2.51	(0.068)	5.42	2.66	(0.014)	√
2 実用	6.80	4.95	(0.059)	5.50	4.74	(0.041)	4.88	3.36	(0.006)	√
3 専門	7.38	1.58	(0.197)	6.78	1.93	(0.062)	7.00	1.00	(0.003)	
5 婦人	3.25	5.49	(0.002)	2.48	2.85	(0.004)				
6 学参 I(小中)	3.20	3.36	(0.001)	4.62	5.48	(0.001)				
7 学参 II(高校)	6.50	1.25	(0.000)	5.00	0.00	(0.000)				
8 児童	2.10	2.27	(0.016)	1.90	1.89	(0.037)				
9 (雑誌扱い)	4.71	4.56	(0.034)	4.17	4.89	(0.007)				
コードなし	5.80	3.70	(0.103)	4.64	2.72	(0.043)	3.99	1.79	(0.337)	–

に示す。これらの分類においても、一般的傾向として、PB>LB>OB という難易度の序列が観察される。

以上の調査結果を総合すると、「出版-書籍 (PB)> 図書館-書籍 (LB)> ベストセラー (OB)」という難易度の序列は、内容、対象読者、発行形態を固定しても、かなり一般的に成り立つと判断することができる。これら 3 つのレジスタの母集団は、それぞれ、

出版-書籍 (PB) 2001 年から 2005 年の間に国内で出版された書籍

図書館-書籍 (LB) 都内 13 自治体以上の図書館が共通して所蔵している書籍

ベストセラー (OB) 1976 年から 2005 年までの 30 年間において、『出版年鑑』および『出版指標年報』のいずれかに、各年のベストセラーとして上記 20 位までに挙げられた書籍 951 冊。

である [1, 5]。つまり、「多くの人々に読まれている」という観点では、PB<LB<OB という順序となる。これは、PB>LB>OB という難易度序列とは、ちょうど反対になる。これらのことから、「多くの人々に読まれる書籍は、難易度の低いものに偏る」という帰結を導くことができよう。

4 おわりに

本稿では、BCCWJ リリース版を用いた obi2/B9 の再構成と、再構成した obi2/B9 を用いた BCCWJ リリース版全サンプルへの難易度付与について報告した。今回構成した難易度スケール B9 を含む『帯 2』システム、および、BCCWJ リリース版の全サンプルに対する難易度データは、準備ができ次第、http://kotoba.nuee.nagoya-u.ac.jp において公開する予定である。

表 9: 発行形態コード別の難易度分布 (PB, LB, OB)

	出版-書籍 (PB)			図書館-書籍 (LB)			ベストセラー (OB)			
	平均	分散	割合	平均	分散	割合	平均	分散	割合	P>L>O
0 単行本	5.76	4.09	(0.584)	4.86	3.08	(0.532)	3.93	1.83	(0.489)	√
1 文庫	3.55	1.59	(0.135)	3.44	1.41	(0.208)	2.50	0.25	(0.001)	√
2 新書	4.17	3.10	(0.060)	4.62	3.04	(0.079)	3.75	1.60	(0.124)	
3 全集・双書	5.92	4.30	(0.078)	4.85	4.14	(0.123)	3.83	0.44	(0.039)	√
4 ムック・その他	4.71	4.63	(0.034)	4.20	4.67	(0.008)				
5 事・辞典	6.12	5.03	(0.002)	5.42	3.58	(0.005)	6.50	0.25	(0.001)	
6 図鑑	3.00	4.50	(0.001)	3.53	5.31	(0.002)				
7 絵本	2.71	2.49	(0.001)	2.00	1.71	(0.001)				
9 コミック	4.36	2.05	(0.001)	2.50	0.25	(0.000)	2.50	0.25	(0.009)	
コードなし	5.80	3.70	(0.103)	4.64	2.72	(0.043)	3.99	1.79	(0.337)	–

謝辞　本論文の 3.4 節 (ジャンル別の難易度分布) は、国立国語研究所の丸山岳彦氏の助言に基づくものである。本研究では、『現代日本語書き言葉均衡コーパス』を利用した。本研究の一部は、国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」に基づくものである。本研究は、JSPS 科学研究費基盤研究 (B) 「平易な日本語表現への工学的アプローチ」(課題番号 24300052) の助成を受けている。

参考文献

[1] 国立国語研究所コーパス開発センター. 『現代日本語書き言葉均衡コーパス』利用の手引, 第 1.0 版, 2011.

[2] 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.

[3] 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.

[4] 佐藤理史, 柏野和佳子. テキストの難易度に対する人間の判断と機械の判断. 第 1 回コーパス日本語学ワークショップ予稿集, pp. 195–202, 2012.

[5] 丸山岳彦, 山崎誠, 柏野和佳子, 佐野大樹, 秋元祐哉, 稲益佐知子, 田中弥生, 大矢内夢子. 『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用. 国立国語研究所内部報告書, LR-CCG-10-01, 国立国語研究所, 2011.