# テキストの難易度に対する人間の判断と機械の判断

佐藤理史　（名古屋大学 大学院工学研究科）†
柏野和佳子　（国立国語研究所 言語資源研究系）‡

# Which Text is Easier?
# —Judgement by Human and Machine—

Satoshi Sato　(Graduate School of Engineering, Nagoya University)
Wakako Kashino　(Dept. Corpus Studies, NINJAL)

## 1　はじめに

いま、被験者に、1000 字の日本語テキストを印刷した 2 枚の紙を渡し、どちらのテキストがやさしいかを回答してもらうことを考えよう。多くの被験者に、このような課題を与えたとき、彼らの回答は一致するのであろうか。

我々は、母国語である日本語のテキストを読んで、そのテキストの難易を判定する能力を有している。たとえば、上記の課題で、小学校の教科書から抜き出したテキストと大学の教科書から抜き出したテキストを比較するのであれば、その判断は容易であろう。そして、その回答は、日本語を母国語とする成人であれば、まず間違いなく一致するだろう。その一方で、我々は、実生活上の経験から、「テキストの難易の判断には個人差がある」ことを知っている。つまり、任意の 2 つのテキストに対して上記のような課題への回答を被験者に求めた場合、その回答は、被験者間で必ずしも一致しないと予想される。

テキストの難易度に関する研究は、リーダビリティ研究として長い歴史を持ち、特に英語に対して、これまで多くの研究がある [1, 2]。その中核は、テキストの難易度と強い相関を持った特徴量 (使用語彙や文の長さ) の発見と、それらを用いた難易度推定公式の提案である。その一方で、人間にテキストの難易の判定を直接求める被験者実験は、ほとんど行なわれていない。すくなくとも、我々が知る限り、日本語テキストに対して、そのような被験者実験を行なった結果は報告されていない。

本論文では、この未調査領域に着目して実施した、テキストの難易の判定を求める被験者実験とその結果について述べる。今回の実験の主要な調査項目は、次の 2 点である。

**Q1** 人間の判断の一貫性：人間の判断は、被験者間で一致するのか。
**Q2** 機械の判断の妥当性：人間の判断と機械の判定結果は、どの程度一致するのか。

## 2　テキストの理解とテキストの難易

テキストの難易という概念は、かならずしも明確な概念ではない。我々人間がテキストを理解する過程において、理解の促進・阻害には、多くの要因が関係している。これらは、おおまかに、次のように整理される。

1. 読むもの (テキスト) 側の要因
   (a) テキストの表示の見やすさ・見にくさ (legibility)：表示媒体、フォーマット、印刷状態など
   (b) テキスト表現のやさしさ・難しさ (readability)：文章表現、語彙、文体など
   (c) 書かれている内容の複雑さ
2. 読み手側の要因

† ssato@nuee.nagoya-u.ac.jp　‡ waka@ninjal.ac.jp

(a) その言語の運用能力
(b) 書かれている内容に対する背景知識
(c) その時の身体状況

これらの要因のうち、我々は、まず、読み手側の要因を、平均的な読み手を仮定することによって捨象する。次に、読むもの側の要因のうち、1(a) の表示に関わる要因を、同一条件となるように制御して排除する。こうして残る要因は、1(b) と 1(c) となる。

一般に、伝える内容が複雑になれば、伝える文章も難しくなるのが普通である。このため、1(b) と 1(c) を完全に分離することは難しい。たとえば、「ゆで卵の作り方」と「ブフ・ブルギニョン (牛肉の赤ワイン煮) の作り方」では、伝達すべき内容 (料理の作成手順の詳細) は、後者の方が複雑である。それを反映する形で、後者のテキストでは、より多くの語彙が使われ、文や文章の構成が複雑化する。

しかし、その一方で、しかし、ほぼ同じ内容を伝える文章において、平易な文章と難しい文章があるのも事実である。たとえば、プロの料理人向けの「ブフ・ブルギニョンの作り方」にはフランス語由来の専門用語が使われるが、これを日常的な一般語に置き換えれば、すくなくとも語彙という観点では平易になる。また、テキストライティングの本には、伝達内容をあまり落さずに、文や文章を平易にする方法が示されている。

以上のような考えに基づき、我々は、1(b) と 1(c) は概念的には独立な要因であるという立場を採用する。もちろん、1(c) の要因の影響を受けることは避けられない。しかし、我々が「テキストの難易」という用語で指し示すものは、テキストの理解過程における 1(b) の要因とする。これを日本語テキストを対象とする場合において、より平易に言い直せば、「そのテキストがどのような日本語で書かれているか、その日本語は理解しやすいか」ということである。

## 3 課題設計

テキストの難易の判定を求める被験者実験において、どのような課題を用いるかは自明ではない。我々の知る限り、日本語テキストに対して、このような被験者実験を行なった例はなく、新たに課題を設計する必要がある。

### 3.1 テキストサイズ

第一に考えるべき問題は、難易の判定を求めるそれぞれのテキストの長さをどの程度にすべきかという問題である。これは、「テキストの難易度というものは、どの程度の長さのテキストに対して定義するのが適切か」という問題でもある。

被験者にテキストを読むことを求めるのであるから、あまり長いテキストを用いるのは非現実的である。他方、文の難易ではなく、テキスト (文章) の難易を計りたいのであるから、それなりの分量があってしかるべきである。

我々は、最終的に、約 1000 字というサイズを定めた。これは、以下に示す理由を勘案した総合的判断に基づく。ただし、難易度を定義するサイズとして、1000 字が最適なサイズであるということを主張するものではない。

1. 我々の経験では、書籍を開いて、その見開き 2 ページを読めば、その書籍がどの程度難しいか、おおよそ見当がつく。書籍の 1 ページ当たりの文字数は千差万別であるが、新書では、約 1000 字程度である。
2. A4 サイズの紙に印刷すると、ほぼ 1 ページの分量となる。紙に印刷して見比べるには都合がよい。

3. テキストを選ぶ母集団に予定している「現代日本語書き言葉均衡コーパス (BCCWJ)[3]」の固定長サンプルのサイズが 1000 字である。
4. テキストサイズが 1000 字あれば、文字や語彙に関する統計量がそれなりに安定する。

### 3.2 課題形式

次に考えるべき問題は、被験者に課す具体的な課題形式である。我々は、次の 2 つの形式を検討した。

1. 1 対比較課題：2 つのテキストを与えて、どちらがやさしいかを回答させる。
2. 並べ替え課題：$n$ 個のテキストを与えて、それらをやさしい順に並べさせる。

テキストの難易の基本となるのは、テキストの 1 対比較である。1 対比較課題は、これをそのまま課題とするものである。この場合、2 つのテキストに対して回答が 1 つ得られるので、テキスト 1 件当たりの 1 対比較結果は 1/2 個となる。

これに対して、並べ替え課題の場合、読むテキストの数に対して、より多くの 1 対比較結果が得られる。たとえば、$n = 4$ の場合を考えよう。4 つのテキストを $A$, $B$, $C$, $D$ と表した場合、被験者の回答は、たとえば、“$B < A < D < C$” のような形となる。この被験者の回答は、6 個の 1 対比較結果、すなわち、“$B < A$, $B < D$, $B < C$, $A < D$, $A < C$, $D < C$” に同意しているとみなすことができる。このように解釈した場合、4 個のテキストを読めば、6 つの 1 対比較結果が得られることになるので、テキスト 1 件当たりの 1 対比較結果は 3/2 個となる。

一般に、$n$ を大きくすれば、テキスト 1 件当たりに得られる 1 対比較は多くなる。その一方で、課題の遂行が繁雑になり、難しくなる。我々は、まず、$n = 5$ の実行可能性を探ったが、5 枚の紙を見比べるのは、認知負荷的にも作業スペース的にもきつかったため、これを断念し、$n = 4$ を採用した。

## 4 課題セットの編纂

我々は、今回、上記で述べた並べ替え課題 ($n = 4$) を 20 課題作成した。

### 4.1 方針

実際の課題の編纂において、次の方針を立てた。

**方針 1** 課題に使用するテキストとして、「現代日本語書き言葉均衡コーパス (BCCWJ)」の固定長サンプル (1000 字) を用いる。

**方針 2** 1 つの課題に含まれる 4 つのテキストのジャンルを揃える。具体的には、日本十進分類 (NDC) の 3 桁が一致するテキストを選択する。

**方針 3** 1 つの課題に含まれる 4 つのテキストに、難易度が異なると思われるものを含める。

方針 1 は、多くの研究者が、使用したテキストにアクセスできるようにするために定めた。被験者実験に使用したテキストが入手可能であれば、実験結果を利用した各種の調査 (たとえば、語彙や文長などの調査) が可能となる。

方針 2 は、「まったく異なるジャンルのテキストを比較することは難しいだろう」という予想に基づき定めた。3 桁の NDC が一致したとしても、かならずしも同じような内容のテキストとは限らない。しかしながら、NDC を揃えないよりは揃えた方が、書かれている内容に対する依存度を軽減することができると考え、このような方針を採用した。

方針 3 は、「似たようなレベルの難易を判定することはかなり難しい」という、これまでの経験に基づき定めた。4 つのテキスト群の中に、相対的にやさしいテキストや相対的に難しいテキストが存

在していれば、並べ替えは比較的容易になると考えられる。テキスト群には、被験者が判断に迷うものも含まれていてよいが、そのようなものばかりだと、被験者に過度の負担を強いることになる。我々は、難易度の値の分布は正規分布に従うと考えており、その仮定が正しいとすると、ランダムサンプリングでは、平均的な難易度のテキストが多数、選ばれることになる。そのため、ランダムサンプリングは採用せず、難易度の値が異なると思われるものを、意図的に選ぶ。

### 4.2 編纂の実際

課題セットの編纂には、「現代日本語書き言葉均衡コーパス」(2009 年モニター版) のうち、書籍 (BK) の固定長サンプル (9,428 サンプル) を用い、次の手順で 20 課題からなる課題セットを編纂した。

1. 被験者への提示にふさわしいテキストサンプルを選択する。具体的には、そのサンプルにおいて、
   (a) article (同一著者による、同一テーマのひとまとまりの文書要素) は 1 つ、
   (b) title (特定範囲の文書要素の内容を代表する記述) は 2 つ以下、
   (c) caption (図表についてのタイトルや説明)、quotation (当該 article 要素とは異なる著作物からの引用や、発話・心内発話の引用・描写・書き起こし)、rejectedBlock (サンプル範囲内において、削除対象となったブロック要素の存在)、verse (詩、和歌、俳句、歌謡などの韻文) はそれぞれ 1 つ以下、

   という条件を満たすものを選択した。4,032 サンプルが選択された。
2. サンプル集合を、日本十進分類 (NDC コード 3 桁) でグルーピングする。なお、以下では、それぞれのグループを NDC グループと呼ぶ。
3. 含まれるサンプル数および分野のバランスを考慮して、20 個の NDC グループを選ぶ。
4. 選ばれたそれぞれの NDC グループに対して、以下を実行する。
   (a) すべてのサンプルに、obi2/B9[4] で難易度を付与する。この難易度は、1(とてもやさしい) から 9(とても難しい) までの 9 段階の値をとる。
   (b) obi2/B9 難易度で、サンプル集合をソートする。
   (c) ソートした列を 5 グループに等分し、中央のグループを除く 4 グループから、それぞれランダムに 1 つづつサンプルを抜き出す。
   (d) 抜き出した 4 つのサンプルを BCCWJ の ID 順にソートし、A から D の記号を付与する。

作成した 20 個の課題 ($t_1, t_2, \cdots, t_{20}$) に使用したテキストサンプルの一覧を表 1 に示す。この表において、各サンプルは、BCCWJ の ID で表示し、その直後の数字は、そのサンプルの obi2/B9 難易度を表す。

## 5 被験者実験

編纂した課題セットを用いて、31 名の被験者に対して実験を行なった。被験者は、すべて日本語を母国語とする成人であり、男女比は 6 名対 25 名、年齢層は 20 代 9 名、30 代 8 名、40 代 12 名、50 代 2 名である。

被験者には、次のような指示を与えた。

1. それぞれのセクションにおいて、$A$ から $D$ までの 4 つのテキストが示されています。それぞれのテキストの長さは、約 1000 字です。(テキストは、文章の途中から始まっていることもあります。)

表 1: 課題セットに使用したテキストサンプルの一覧

| 課題 ID | NDC | A | | B | | C | | D | |
|---|---|---|---|---|---|---|---|---|---|
| $t_1$ | 049 | LBh0_00001 | 6 | LBi0_00008 | 3 | PB30_00080 | 4 | PB40_00029 | 5 |
| $t_2$ | 159 | LBm1_00036 | 6 | PB21_00081 | 3 | PB41_00019 | 6 | PB51_00022 | 5 |
| $t_3$ | 188 | LBo1_00028 | 5 | LBo1_00031 | 4 | PB21_00057 | 6 | PB51_00071 | 6 |
| $t_4$ | 210 | LBg2_00067 | 7 | LBp2_00050 | 5 | PB12_00061 | 6 | PB42_00048 | 6 |
| $t_5$ | 289 | LBc2_00043 | 6 | LBc2_00044 | 8 | LBn2_00008 | 4 | PB22_00135 | 3 |
| $t_6$ | 291 | LBm2_00069 | 5 | PB22_00295 | 7 | PB52_00028 | 5 | PB52_00097 | 3 |
| $t_7$ | 302 | LBe3_00057 | 4 | LBh3_00090 | 7 | LBn3_00122 | 7 | PB23_00011 | 5 |
| $t_8$ | 312 | LBa3_00038 | 7 | LBg3_00076 | 6 | LBm3_00066 | 8 | PB53_00488 | 6 |
| $t_9$ | 335 | LBe3_00049 | 5 | LBh3_00101 | 8 | LBo3_00061 | 6 | PB53_00191 | 8 |
| $t_{10}$ | 361 | LBi3_00090 | 6 | LBr3_00135 | 6 | PB23_00122 | 7 | PB33_01003 | 8 |
| $t_{11}$ | 367 | LBl3_00039 | 8 | LBt3_00058 | 3 | PB23_00258 | 7 | PB43_00904 | 5 |
| $t_{12}$ | 369 | LBj3_00094 | 5 | PB33_00465 | 8 | PB43_00470 | 4 | PB53_00341 | 9 |
| $t_{13}$ | 493 | LBa4_00018 | 3 | PB14_00183 | 4 | PB14_00232 | 6 | PB24_00023 | 7 |
| $t_{14}$ | 498 | LBc4_00017 | 3 | LBo4_00052 | 5 | LBs4_00017 | 7 | PB34_00399 | 6 |
| $t_{15}$ | 673 | PB36_00116 | 3 | PB36_00165 | 4 | PB56_00020 | 5 | PB56_00064 | 8 |
| $t_{16}$ | 783 | LBm7_00049 | 7 | LBq7_00031 | 5 | PB17_00054 | 5 | PB27_00075 | 3 |
| $t_{17}$ | 913 | LBm9_00025 | 2 | LBo9_00127 | 5 | PB39_00681 | 3 | PB59_00309 | 5 |
| $t_{18}$ | 914 | LBf9_00067 | 4 | LBk9_00080 | 5 | LBp9_00155 | 2 | PB49_00275 | 7 |
| $t_{19}$ | 916 | LBl9_00262 | 5 | LBm9_00267 | 3 | PB29_00053 | 5 | PB49_00126 | 3 |
| $t_{20}$ | 933 | LBg9_00165 | 2 | LBj9_00197 | 3 | LBq9_00073 | 4 | PB59_00371 | 3 |

2. それぞれのテキストに目を通し、あたなが感じた素朴な印象 に基づいて、やさしい順に並べて下さい。
3. 順位が付けるのが難しい場合は、何度も読み比べてもかまいません。
4. どうしても難易度の順位が付けられない場合は、回答シートの同じ箱に、複数の記号を記述して下さい。(ただし、可能である限りは、順位を付けて下さい。)
5. コメント欄には、難易度の順位付けが容易だったか難しかったかを記述して下さい。
   例： $A$ が最もやさしく、$C$ が最も難しいという判断は容易だったが、$B$ と $D$ は判断に迷った。最終的には $D < B$ としたが、難易度にはほとんど差がないように思う。
6. その他、気付いたこと、感想等があれば、コメント欄に記述して下さい。

## 6 実験結果の分析

### 6.1 1 対比較コードへの変換

並べ替え課題 ($n = 4$) の回答は、たとえば、“$B < A < D < C$” のような順序列である。これを 6 個の 1 対比較結果とみなし、6 ビットのコードで表すこととする。それぞれのビットは、上位ビットから、$A$ と $B$、$A$ と $C$, $A$ と $D$、$B$ と $C$、$B$ と $D$、$C$ と $D$ の 1 対比較結果を表し、それぞれ、正順ならば 0、逆順ならば 1 と定める。この結果、上記の回答は “10001” と表現されることになる。課題セットは 20 課題から構成されるため、課題セットに対する被験者の回答は、120 ビットのコードで表現されることになる。これを **1 対比較コード**と名付ける。このコードのそれぞれのビットは、ある特定のテキストサンプルの組に対する 1 対比較の結果に対応する。

31 名の被験者 ($p_{01}$–$p_{31}$) の回答を、120 ビットの 1 対比較コードで表現したものを、表 2 に示す。この表では、先頭行 ($m$) に、31 名の被験者の過半数が、正順 (0) と逆順 (1) のどちらを支持しているかを示し (以下、多数派の回答とよぶ)、各被験者の行では、多数派の回答と値が異なるビットのみ、数字で示した。なお、列 $d_0$ は、各被験者の回答と多数派の回答の相違ビット数 (ハミング距離) を示している。また、下から 2 行目 (av.) に、31 名の被験者の平均値を示した。

表 2 で、縦の列に着目すると、31 名の被験者の結果がすべて一致する (すべて ‘.’) 1 対比較もあれば、判定が割れる 1 対比較もあることがわかる。この表から導ける、調査項目 Q1 に対する答えは、

表 2: 実験結果 (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. | ..... | ..... | 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.... | ..... | ..... | ..... | 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..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..0. | ..... | .1.... | ....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.... | ..... | 11.... | ..... | ....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. | ..... | .1.... | .....0 | ..... | ..... | ..... | ..... | ..... | ..... | ....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 | ..... | ..... | ..... | ..... | 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.... | ..... | ..... | ....0. | ....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..... | .1.1. | 1..... | .00000 | ...0.. | ..... | 1.... | ....0 | 25 | 12 | 12 |
| $p_{20}$ | ..... | ..... | 0..11. | ....0. | 0..... | ..... | 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.. | ..... | ..... | ..... | ....0. | ....1. | ..... | ...0.. | ...1.. | ....1. | ..... | ..1.1. | ..... | ..... | ..... | .0.... | ..... | 1....0 | 14 | 7 | 7 |
| $p_{22}$ | ..... | ..0..0 | ..11. | ....0. | ..... | 110000 | ...1.. | ..0... | ....0. | 11.... | ..... | ....1. | ..... | ..0... | ..1.1. | ...00. | ...0.. | ..... | ..... | ..... | 24 | 13 | 10 |
| $p_{23}$ | ..... | .1.... | .1..1 | ..... | ..... | ...0.. | ...1.. | .1.... | ..... | .1.... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ....1. | ....0 | 11 | 3 | 3 |
| $p_{24}$ | .00... | ....1. | .11.. | ..... | 0....0 | ..... | ..11.. | .1.1.0 | ....0. | ...0.. | .0..0. | .1.1.. | ..... | ..... | ....0 | 0..... | ..... | 1..... | ..... | ..111. | 24 | 12 | 11 |
| $p_{25}$ | ..11. | ..0... | .1.... | ..... | 000.00 | 11...0 | ...1.. | ..... | ....0. | ...0.. | ...1.. | ...0.1 | 0..... | ..... | 0.1.11 | 000... | ..... | 0..... | ..... | 1.1.1. | 31 | 21 | 14 |
| $p_{26}$ | ..... | ..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.. | ....1. | ..... | ..... | 0..... | 000..0 | ...0.. | ..... | ..... | ...11. | 26 | 12 | 12 |
| $p_{28}$ | ..... | ..... | .1.... | ..... | ..... | 0..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 |
| B9 | ..... | .?.... | .0...? | ....0? | 0..... | .?..... | ...?.. | ....?. | .....? | ?..0.. | ....0. | .1.... | 0....0 | ..... | ..... | ....?.. | ..... | ....?. | ?..?. | ....?. | 14.5 | 8.5 | 8.0 |

「人間の判断が一致するかどうかは、比較するテキスト対に依存する」という予想通りの帰結である。

### 6.2 1 対比較と有意水準

それぞれの 1 対比較に対して、31 名の被験者の回答を集計すると、多数派と少数派に分かれる。多数派の人数を $M$ とするとき、「多数派 (過半数) である」ことが有意水準 1%で統計的に有意であるのは、$M \geq 22$ の場合に限られる[1]。今回の実験結果では、120 個の 1 対比較中、96 個が統計的に有意であった。(表 2 の行 $ci$ で、'+' または '−' の列が、統計的に有意な 1 対比較を表す。)

ここで、我々は、1 対比較 (2 つのテキストの難易) の「正解」をどのように定義すべきかという問題に直面する。テキストの難易判定は人間しかできないのであるから、多くの人々[2]が同意する回答を、正解とするしか方法がない。ここで、「多くの人々」がどれくらいの割合を意味するものとするか—過半数でよいのか、それとも 2/3 以上必要か—には自由度があるが、その最低ラインは過半数である。

以降、本実験の分析では、多数派の人数 $M$ が $M \geq 22$ を満たす 1 対比較 96 個に対して、多数派の回答を難易の正解順序として定義する。表 2 の列 $d_c$ は、正解順序を定義した 96 ビットにおける、正解と被験者のハミング距離を示した。

列 $d_c$ の値からわかるように、被験者と正解との距離は 2 から 22 まで分布し、その平均は 9.7 である。つまり、平均的な被験者は、96 個の 1 対比較において、10 個程度、正解 (多数派) とは異なる回答をするということである。なお、ここでは、詳細は省略するが、特定の 2 名の結果が非常によく似ているという事実はない。つまり、被験者のサブグループに特定の傾向があるということはなく、被験者の回答は、それぞれの個人によってばらつく。

### 6.3 独立な 1 対比較

前述のように「正解」を定義すると、96 個の 1 対比較を、独立な 1 対比較と、そうでない 1 対比較に分割することができる。たとえば、課題 $t_1$ の正解は 11?000 である ('?' は定義されないことを示す)。つまり、定義される正解順は、"$A > B,\ A > C,\ B < C,\ B < D,\ C < D$" となる。このうち、"$B < D\ \ (B < C < D)$" と "$B < A\ \ (B < C < A)$" は、他の 1 対比較の結果と推移律から一意に定まる。これらを除いた 3 つの 1 対比較結果 "$A > C,\ B < C,\ C < D$" は、独立である。それぞれの 1 対比較がどちらの区分に含まれるかを、表 2 の行 $ci$ に、'+'(独立)、'−'(非独立) の記号で示した。独立な 1 対比較は、96 個中 62 個ある。この表の列 $d_i$ は、独立な 1 対比較に限定した場合の、正解と被験者のハミング距離を示している。この表からわかるように、前述の $d_c$ とこの $d_i$ の間には、それほど大きな差はない。すなわち、独立な 1 対比較に限定したとしても、被験者の回答はばらつく。

### 6.4 機械による難易判定

先に述べたように、並べ替え課題 ($n = 4$) の各テキストには、obi2/B9 の難易度 (9 段階) が付与されている。この難易度を、人間の判断と同様の形に変換する。ただし、比較対象とする 2 つのテキストに、同一の難易度が付与されいる場合は、「判定不能」とする。このようにして求めた obi2/B9 による 20 課題に対する回答を、表 2 の最下行 (B9) に示す。ここで、'?' は、判定不能を意味する。

被験者の場合と同様に、obi2/B9 の結果に対しても、$d_0$、$d_c$、$d_i$ を計算した[3]。このうち、$d_c$ と $d_i$ の値は、順に 8.5、8.0 であり、いずれも被験者の平均値を若干下回る。以上のことから、調査項目 Q2 に対し、「機械 (obi2/b9) の判断は、平均的な人間の判断と同程度である」という帰結が得られる。

---

[1] $z$ 検定を用いた。

[2] どのような母集団を想定するかについては、さらに議論が必要であるが、ここでは、十分な日本語能力を持った成人母語話者を想定する。

[3] ただし、判定不能のビットは、距離 0.5 として計算した。

表 3: 被験者の判断と機械の判断

| | obi2/B9 難易度の差 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 被験者 | +5 | +4 | +3 | +2 | +1 | 0 | −1 | −2 | 計 |
| 31–0 | 1 | 3 | 9 | 9 | 3 | 1 | 0 | 0 | 26 |
| 30–1 | 2 | 2 | 4 | 3 | 1 | 0 | 0 | 0 | 12 |
| 29–2 | 0 | 1 | 1 | 4 | 1 | 2 | 0 | 0 | 9 |
| 28–3 | 1 | 0 | 1 | 4 | 2 | 1 | 0 | 0 | 9 |
| 27–4 | 1 | 1 | 1 | 3 | 2 | 0 | 0 | 0 | 8 |
| 26–5 | 0 | 0 | 1 | 3 | 3 | 1 | 2 | 0 | 10 |
| 25–6 | 0 | 0 | 0 | 2 | 1 | 2 | 1 | 0 | 6 |
| 24–7 | 0 | 0 | 0 | 1 | 4 | 0 | 0 | 1 | 6 |
| 23–8 | 0 | 1 | 1 | 3 | 1 | 0 | 1 | 0 | 7 |
| 22–9 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 3 |
| 小計 | 5 | 8 | 18 | 32 | 21 | 7 | 4 | 1 | 96 |
| 21–10 | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 0 | 5 |
| 20–11 | 0 | 1 | 0 | 3 | 2 | 0 | 0 | 0 | 6 |
| 19–12 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 1 | 7 |
| 18–13 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 17–14 | 0 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 4 |
| 16–15 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 総計 | 5 | 9 | 18 | 37 | 30 | 13 | 6 | 2 | 120 |

表 3 に、被験者による判定結果と機械の判定結果の、より詳細な比較を示す。この表は、被験者の結果が $x$ 対 $y$ だった 1 対比較に対する、obi2/B9 の判定結果 (難易度の差；差が正の場合が多数派の回答と一致) を示している[4]。なお、正解を定義するのは、22 対 9 以上の場合である。

この表からわかるように、obi2/B9 難易度の差が大きいほど、正解とよく一致する。事実、難易度の差が +3 以上あると判定したものは、正解が定義されれない 1 件を除き、すべて正解と一致していた。判定不能を除けば、obi2/B9 が正解と異なる判定を下すのは 5 件だけであり、そのうち 4 件は、難易度の差は −1 である。これらの結果も、機械の判断が人間の判断とよく一致していることを示している。

## 7 まとめ

本研究で得られた帰結は、次の 2 点にまとめられる。

1. テキストの難易に対する人間の判断は、比較対象のテキスト対に依存し、多くの人間の判断が一致するものもあれば、一致しないものもある。テキスト間の難易は、多くの人間の判断が一致するテキスト対に対してのみ、定義されるべきであろう。
2. テキストの難易に対する機械 (obi2/B9) の判断は、平均的な人間と同程度の性能である。

**謝辞** 本研究では、「現代日本語書き言葉均衡コーパス」(2009 年モニター版) の一部を利用した。また、本研究は、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。

## 参考文献

[1] William H. DuBay. *Smart Language: Readers, Readability, and the Grading of Text*. Impact Information, 2007.

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

[3] Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of LREC 2010*, pp. 1483–1486, 2010.

[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.

[4] 対象とする 1 対比較は、全 120 個である。