# 日本語のかな漢字テキストとローマ字テキストのタイピングにおける変換係数について

河原達也
（京都大学）

インテルステノでは毎年春（概ね３月～４月）にインターネットでタイピングのコンテスト（Internet Contest: IC）を実施している。一つの言語でも参加可能であるが、多言語（今年は 18 の言語）で実施されており、多くの参加者が挑戦している。日本語も 2015 年から採用されているが、海外の多くの参加者は日本語のかな漢字の読み書きができないので、ローマ字（transliterated）のテキストも併用されており、参加者はいずれかを選択する。かな漢字のテキストとローマ字のテキストでは、文字数や入力の難易度が異なるので、両者を公平に比較できるように入力文字数の変換係数（=かな漢字の１文字入力がローマ字の何文字入力に相当するか）を定める必要がある。この係数の値として導入当初から 2.2607 が採用されてきたが、これはローマ字テキストに有利なことが経験的に知られ、日本人の参加者の多くもかな漢字のテキストでなくローマ字テキストを選択する事態になっていた。このたび、この係数の推定に取り組み、結果的に 2024 年のコンテストから 2.95 という値に改訂することができたので、報告する。

## 1.　日本語入力とキーストロークの考察

日本語のテキストは主にかな漢字から構成される。かなの１文字は概ね、ローマ字２文字に対応する。漢字の１文字は概ね、ローマ字２～４文字程度に対応する。漢字を入力する際には、まずローマ字またはかな文字を入力してから、漢字に変換する。その際に変換キーを１回以上押すとともに、変換が正しいか視認した上で確定キーを押す必要がある。例えば、「兼子」と入力する場合は、ローマ字で“kaneko”と入力してから、変換キーを１～２回押した上で確定キーを押す。すなわち、アルファベットのキーを６回に加えて２～３回のキーストロークが必要となる。変換キーと確定キーを押す手間は他のキーに比べて迅速にできるが、所望の漢字であるかの視認には時間を要する。変換がうまくいかない場合に入力を区切り直したりすることもあるが、ここでは考慮しない。一方で、ローマ字の場合は単語の間にスペースを入力する必要がある。かな漢字変換を文節単位で行うとすると、変換／確定キーはローマ字のスペースに対応するとみなすことができる。なお、句読点とピリオド・コンマは概ね対応すると考えられる。

以下にローマ字テキストとかな漢字テキストの例を示す。

Nihongo no nyuuryoku wa muzukashii desu. (40 文字+1 大文字=41 ストローク)

日本語の入力はむずかしいです。(15 文字)

単純に文字数の比をとると 41/15=2.73 となるが、入力の手間を考慮する必要がある。このかな漢字テキストの入力には、典型的には以下のキーストロークが必要となる。

nihongono<SP>nyuuryokuha<SP>muzukasiidesu。<ENT>

(34 文字 ＋2 回の<SP>＋1 回の<ENT>＝37 キーストローク)

ここで <SP> は変換キー、 <ENT> は確定キーである。この例では、変換／確定は、自立語 1 つと付属語からなる文節単位で行われている。自立語と付属語の後にスペースが入る反面、変換キーを複数回押す場合もあるので、単純な比較は容易でないが、スペースキーや変換／確定キーの入力自体は他のキーストロークに比べて手間は小さいと考えられる。これに対して入力のボトルネックになるのは、変換の際に所望の漢字になっているか視認する手間と考えられる。したがって、漢字とカタカナからなる単語の出現する回数に応じた補正を行う必要があると考え、理論的な変換係数として以下の式を考えた。

理論的な変換係数 ＝（ローマ字文字数＋漢字・カタカナの単語数）／かな漢字文字数

（式 1）

ここで、漢字・カタカナの単語数は変換キーを押す回数に概ね相当する。上記の例では、変換係数の推定値は、(41+2)/15=2.86 となる。

## 2. テキスト分析による推定

インターネットコンテストの 2021 年から 2025 年に用いられた日本語のかな漢字テキストとローマ字テキストを対象として分析を行った。ローマ字テキストはかな漢字テキストから自動的に生成されたもので、同一内容である。統計量を表 1 に示す。

ローマ字テキストとかな漢字テキストの単純な文字数の比においてもすべて 2.65 以上となっており、導入当初の 2.2607 という値が小さすぎることは明らかである。さらにかな漢字変換の手間を考慮すると「適正な」変換係数は 2.90-3.10 あたりであることが予想される。

表 1　2021～2025 年のインターネットコンテスト(IC)で用いられたテキストの分析

| | IC2021 | IC2022 | IC2023 | IC2024 | IC2025 |
|---|---|---|---|---|---|
| (1) かな漢字テキストの文字数 | 4002 | 3413 | 3906 | 4572 | 3912 |
| (2) 漢字の文字数 | 1341 | 1090 | 1769 | 1683 | 1576 |
| (3)カタカナの文字数 | 298 | 546 | 200 | 418 | 42 |
| (4)漢字とカタカナの単語数 | 879 | 845 | 1104 | 1116 | 1038 |
| (5) ローマ字テキストの文字数 | 11120 | 9055 | 11135 | 12591 | 11055 |
| (6)スペースの数 | 2051 | 1536 | 1803 | 2018 | 1776 |
| (5)/(1) ローマ字文字数／かな漢字文字数 | 2.78 | 2.65 | 2.85 | 2.75 | 2.83 |
| ((5)+(4))/(1)（ローマ字文字数＋単語数）／かな漢字文字数 | 3.00 | 2.90 | 3.13 | 3.00 | 3.09 |

3.　タイピング実験による算出

次に 2021 年(IC2021)と 2022 年(IC2022)のテキストを用いて、インターネットコンテストに参加した経験のある方にタイピングを行ってもらった。同一内容のかな漢字テキストとローマ字テキストの両方を入力してもらい、そのスコア（所定時間内に正しく入力した文字数）を比較した。かな漢字テキストとローマ字テキストは同一内容であるため、テキストの内容を記憶する影響を考慮し、公正な比較を行うために、参加者を 2 つのグループに分けて、以下の順番で 4 種類のテキストの入力を行ってもらった。

グループ 1:　IC2021 ローマ字 → IC2022 かな漢字 → IC2021 かな漢字 → IC2022 ローマ字

グループ 2:　IC2021 かな漢字 → IC2022 ローマ字 → IC2021 ローマ字 → IC2022 かな漢字

タイピングはすべて 5 分間で行い、Practical Type(http://www.second-way.com/)を用いてスコアの算出を行った。その結果を表 2 に示す。

表 2　2021 年・2022 年のテキストによるタイピングスコアの比較

| タイピスト ID | グループ | IC2021 | | | IC2022 | | |
|---|---|---|---|---|---|---|---|
| | | ローマ字のスコア | かな漢字のスコア | スコア比 | ローマ字のスコア | かな漢字のスコア | スコア比 |
| A | 2 | 2673 | 997 | 2.68 | 2696 | 897 | 3.01 |
| C | 1 | 1880 | 843 | 2.23 | 1819 | 780 | 2.33 |
| D | 2 | 2316 | 785 | 2.95 | 1980 | 732 | 2.70 |
| E | 2 | 1316 | 609 | 2.16 | 1265 | 418 | 3.03 |
| F | 1 | 2825 | 930 | 3.04 | 2827 | 1018 | 2.78 |
| G | 2 | 2156 | 765 | 2.82 | 2087 | 765 | 2.73 |
| H | 1 | 2879 | 874 | 3.29 | 2921 | 793 | 3.68 |
| I | 1 | 1795 | 594 | 3.02 | 2065 | 644 | 3.21 |
| J | 2 | 2164 | 579 | 3.74 | 2370 | 586 | 4.04 |
| 平均 | 1 | 2345 | 810 | 2.90 | 2408 | 809 | 3.00 |
| | 2 | 2125 | 747 | 2.87 | 2080 | 680 | 3.10 |
| | 1&2 | 2223 | 775 | 2.88 | 2226 | 737 | 3.06 |
| 標準偏差 | 1 | 508 | 129 | 0.40 | 475 | 134 | 0.50 |
| | 2 | 446 | 149 | 0.51 | 477 | 164 | 0.49 |
| | 1&2 | 487 | 144 | 0.46 | 503 | 165 | 0.50 |

ID B のタイピストは実施する順番を間違えたので除外した

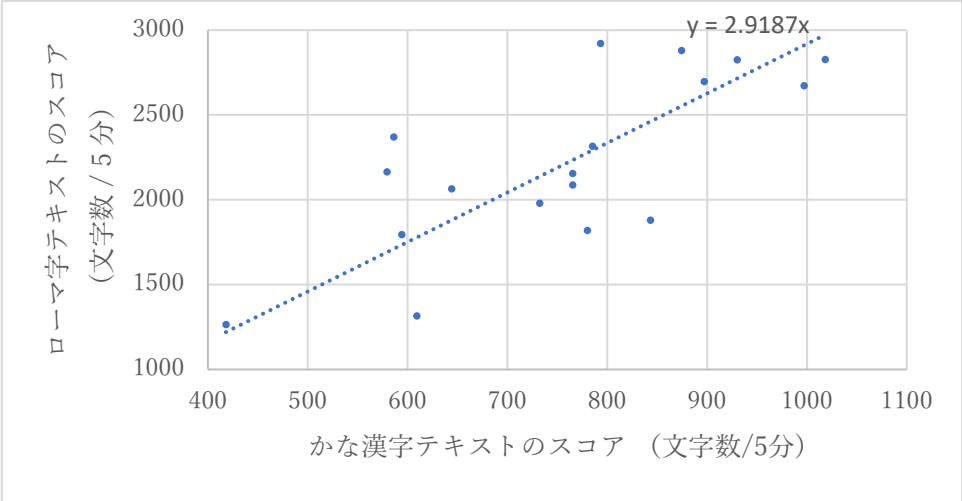


図１　かな漢字テキストとローマ字テキストのスコアの相関

ローマ字テキストとかな漢字テキストのスコアの平均の比は、２つのグループ間でほぼ一致しており、2.87 から 3.10 の間の値となっている。全体の平均値は 2.97 である。この値は、表 1 の最後の行に示した理論的な変換係数とほぼ同じである。タイピスト毎の２つのスコアの関係をプロットしたものを図 1 に示す。２つのスコアに概ね相関があることが確認でき、回帰直線の傾き（図 1 の直線の傾き）は 2.92 である。

## 4.　結論

日本語のかな漢字テキストとローマ字テキストを分析して、両者のタイピングを比較・正規化する係数として 2.90-3.00（IC2021 と IC2022 のみの場合）の値を推定した。両方のテキストを用いたタイピングによる実験において変換係数として 2.88-3.06 の値が得られた。これらの結果を総合して、変換係数として **2.95** を提案した。

## 5.　その後の経緯と限界

この提案は 2022 年 8 月のインテルステノ大会中に開催された科学委員会で審議・承認され、2023 年の理事会・評議会で審議・承認の末、2024 年のインターネットコンテストから採用された。その間、理事会から一度照会（差戻し）があり、議論の補強を行った。

本考察の最大の問題点は、根拠とするデータ、特に表 2 と図 1 のサンプル数が少ないことである。この点が理事会においても問題視された。しかしながら、この実験は、同一内容のかな漢字とローマ字の複数のテキストを熟練した方にタイピングしてもらう必要があり、協力者の確保を含めて容易に実施できるものでない。しかしながら、過去のコンテストにおいて、日本人でもローマ字を選択した競技者の方がかな漢字を選択した競技者より有意に（平均で 35%）高いスコアとなっていたことが改定の根拠となった。

一方、表 1 に示した日本語テキストの分析は、いくらでもデータを増やすことができる。

ただし厳密には、式 1 に示す理論的な変換係数の値は用いるテキストによって変わる。IC2022 は「ラーメン」に関する文章でカタカナが多い一方、IC2023 は漢字が多いテキストのため、両者の間で変換係数の推定値が大きく異なった。したがって毎年のコンテスト毎に、理論的な変換係数の推定値を計算して設定することも考えられる。ただし、IC2021 と IC2022 で、表 1 に示す理論的な推定値と表 2 に示す実際のスコア比には、0.15 程度の乖離と両者の逆転現象がみられ、その程度の不確実性がある。しかしながら、2.95 という値は平均的にみて概ね妥当と考えられる。

## 6. タイピングテキスト課題の作成法

当方はタイピングコンテストに関わった経験はなかったが、インテルステノの教育委員を務めている関係で、2021 年から課題テキストの作成を依頼された。著作権の問題のない文書から、入力しやすい文を選別してテキストを作成している。具体的には、人名等の固有名詞や数字を含まないよう指示があり、さらに難読漢字や変換が難しそうな単語は極力避けている。これは国内におけるコンテストでも同様であろう。

しかし、ローマ字テキストを作成するのはかなりの労力を要する。これには、漢字に読みを付与し、単語単位に分割する必要がある。日本語において単語の単位の明確な定義はないので、形態素解析を行う。形態素解析には Juman、ひらがなからローマ字への変換には kakasi といういずれも京都大学の先生によって開発されたソフトを使用した。ただし、単語分割や読み付与には一定の誤りが不可避であり、手作業で確認・修正を行う必要がある。さらに、文頭を大文字にする、句読点をピリオド・コンマに置換する、などの処理を行う。この問題について知見のある方は是非連絡頂きたい。

## 7. 謝辞

本件を含めて、日頃からインテルステノに関してご助言を頂きます兼子次生様に深く感謝します。本件を含めて、インターネットコンテストの日本人参加者の取り纏めをして頂いている中山貴之様にも厚く御礼申し上げます。中山様には本稿の閲読もして頂きました。本考察については京都大学の森信介教授と米国・カーネギーメロン大学のグラム・ニュービッグ准教授にコメント・確認を頂きました。また、タイピング実験に協力頂いた 10 名の方に感謝します。