英文リーダビリティ指標 FKGL はほぼ平均文中音節数である

江原 遥 [1]
[1] 東京学芸大学
ehara@u-gakugei.ac.jp

概要

Flesch-Kincaid 式は，英語の可読性指標として古典的だが代表的であり，近年の大規模言語モデルの出力評価においても用いられている．これらの式は単語あたりの平均文長と音節あたりの平均単語長の線形和であり，この線形和には数十年の長期の使用に耐える，人間の認知に基づく何らかの理論的要因の存在が示唆される．本研究では，これらの式の理論的解析を行いこの理論的要因を明らかにした．先行研究とは異なり，これらの式が「1 文あたりの平均音節数」として解釈できることを示した．学年が上がるにつれて語彙の範囲は拡大する可能性があるが，音節の範囲は学年や年齢に関係なく一定に保たれる．これが，長期に使われ続ける要因であろう．評価実験では，BNC を用いて本理論枠組みの妥当性を確認し，他言語版 FKGL の提案も行う．

1 はじめに

Flesch-Kincaid 式，特に Flesch-Kincaid Grade Level (FKGL) [1] および Flesch Reading Ease (FRE) [2] は，大規模言語モデルが生成するテキストを含め，英語テキストの可読性を評価するために広く使用されている [3, 4, 5]．これらの指標が広く普及している理由は，FKGL スコアが解釈しやすいこと，そして単語リストに依存しないことである．単語リストは維持が困難であり，新たな用語（例：「スマートフォン」など）が登場するたびに対応しなければならない．Flesch-Kincaid 式が長期間にわたり受け入れられている理由の一つは，その堅牢性である [1]．単語リストに依存する手法とは異なり，新しい単語が出現しても影響を受けにくいためである．では，これらの式が一貫して信頼性を維持できる理由は何か．我々は，その理由が人間の認知特性に基づいていると仮定し，本研究を進めた．後述するように，これらの式は認知的特性に基づいていることを示す．

本研究では FKGL に焦点を当てる．FRE にも同様の論理が適用できるため，表記を標準化し，値が高いほど難易度が高いことを示すように定義する．

$$\begin{aligned} \text{FKGL} &= 0.39\left(\frac{\text{総単語数}}{\text{総文数}}\right) \\ &+ 11.8\left(\frac{\text{総音節数}}{\text{総単語数}}\right) - 15.59 \end{aligned} \tag{1}$$

FKGL の根拠は以下の通りである．1 つ目の項である「文あたりの単語数」は，文の複雑さの指標として機能する．しかし，単語数だけでは文の難易度を完全には捉えられない．たとえ短い文であっても，教育水準に対して難解な単語が含まれていれば，その文は学生にとって理解が困難である．したがって，文に含まれる語彙の難易度も考慮する必要があり，2 つ目の項がこの補正を行う．

しかしながら，この補正は過度にヒューリスティックであり，スコアが過剰に補正される理論的保証が欠けている可能性がある．この洞察は，より大規模なアノテーション付きデータセットを用い，言語的特徴を考慮することで，補正手法を改善する方向性を示唆する．しかし，その代償として，時間的な堅牢性が損なわれるリスクがある．自動可読性評価に関する重要な先行研究として [6] が挙げられ，他の研究は [7] で調査されている．

本研究では，式 1 に関連する以下のリサーチクエスチョン（**RQ**）に取り組む．

RQ1　平均単語数と平均音節数の線形結合が有効に機能するのはなぜか？この種の線形結合は，適切に係数が選択されれば狭い範囲の M を持つ「1 文あたりの平均音節数」の積として表すことができる．我々は，FKGL が「1 文あたりの平均音節数」を用いて難易度を決定していることを示した．item[RQ2] **過剰補正の可能性はあるか？すなわち，1 単語あたりの平均音節数が過度に高くなる可能性はあるか？**上記のように，最大の FKGL 値は M の最大値を決定することで得られる．この手法により，上限が確立

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

された．

RQ3　FKGL の認知的根拠は何か？人は成長するにつれて語彙が増加する．そのため，少数の単語から成る文であっても難解な単語が含まれていれば，文は複雑になる．しかし，音韻レパートリーは年齢と共に増加しない．つまり，認識可能な音節の種類は時間が経過しても一定に保たれる．したがって，音節数の平均が多い文は，音節数が少ない文よりも確実に複雑である．

2　FKGL の分析

式 1 には，文あたりの単語数と単語あたりの音節数が現れる．ここでは，各文の単語数に注目する．計算言語学では，文を単語の連続として捉えることが一般的であり，文末には常に明示されないが存在する文末記号（EOS）があると仮定する．その結果，EOS の数は総文数と一致する．したがって，EOS 出現の確率は以下の式で表される．簡単のため，この確率を p_{sw} と定義する．ここで，s は文を，w は単語を表す．つまり，文あたりの単語数は文境界を示す単語の出現確率の逆数と解釈できる．

$$p_{\mathrm{sw}} \equiv \frac{\text{総文数}}{\text{総単語数}} \tag{2}$$

同様に，単語あたりの音節数は音節の連続と見なせる．文との混同を避けるため音節には記号 “l” を用い，この確率を p_{wl} と表す．

$$p_{\mathrm{wl}} \equiv \frac{\text{総単語数}}{\text{総音節数}} \tag{3}$$

さらに，定数 $a = 0.39, b = 11.8, c = -15.59$ を設定すると，FKGL は以下のように書き換えられる．

$$\begin{aligned} \mathrm{FKGL} &= \frac{a}{p_{\mathrm{sw}}} + \frac{b}{p_{\mathrm{wl}}} + c \\ &= \frac{1}{p_{\mathrm{sw}} p_{\mathrm{wl}}} (a p_{\mathrm{wl}} + b p_{\mathrm{sw}}) + c \end{aligned} \tag{4}$$

ここで，1 文あたりの音節数 p_{sl} を導入する．

$$\begin{aligned} p_{\mathrm{sl}} &\equiv \frac{\text{総文数}}{\text{総音節数}} \\ &= \frac{\text{総文数}}{\text{総単語数}} \frac{\text{総単語数}}{\text{総音節数}} \\ &= p_{\mathrm{sw}} p_{\mathrm{wl}} \end{aligned} \tag{5}$$

その結果，式 4 は以下のように書き換えられる．

$$\mathrm{FKGL} - c = \frac{1}{p_{\mathrm{sl}}} (a p_{\mathrm{wl}} + b p_{\mathrm{sw}}) \tag{6}$$

式 4 の右辺は，第 1 項 $1/p_{\mathrm{sl}}$ と第 2 項 $a p_{\mathrm{wl}} + b p_{\mathrm{sw}}$ に分解できる．式 6 までの過程では単純な式変形のみを行っており，近似は一切行っていない．次の節では，式 6 に基づく RQ について議論し，後続の節でこれらの質問を検証する．

2.1　RQ への回答

第 1 の RQ は「**なぜ平均単語数と平均音節数の線形結合は有効に機能するのか？**」である．この点は，式 6 によって部分的に説明できる．式 6 において，FKGL は本質的に $\frac{1}{p_{\mathrm{sl}}}$（1 文あたりの平均音節数）と M の積として表される．ここで，M は以下のように定義される．

$$M = (a p_{\mathrm{wl}} + b p_{\mathrm{sw}}) \tag{7}$$

実験では，一般的なコーパスを用いて FKGL における M が大きく変動しないことを示す．

第 2 の RQ は「**過剰補正の可能性はあるか？すなわち，1 単語あたりの平均音節数が過度に大きくなることはあるか？**」である．ここで，式 7 は p_{wl} と p_{sw} が確率値であるため，有界であることが容易にわかる．したがって，$0 \le M \le a + b$ が成り立つ．これを式 6 と組み合わせることで，式 1 に対する以下の上限が導出される．

$$c \le \mathrm{FKGL} \le \frac{1}{p_{\mathrm{sl}}} (a + b) + c \tag{8}$$

式 8 において，c は負の値であり，FKGL の場合，$c = -15.59$ である．一方，a と b は正の値である．したがって，FKGL は 1 文あたりの音節数によって上限が定まる．つまり，1 単語あたりの平均音節数が過度に大きくなったとしても，FKGL は 1 文あたりの平均音節数によって制限される．我々の知る限り，この理論的な上限について言及した先行研究は存在しない．したがって，これは新しい結果であり，本研究における貢献の一つである．

第 3 の RQ は「**FKGL の認知的根拠は何か？**」である．$\frac{1}{p_{\mathrm{sl}}}$ は 1 文あたりの平均音節数である．1 文あたりの平均音節数は，1 文あたりの平均単語数とは大きく異なる．これは，文中に許容される単語の平均数が学年に応じて変化するためである．直感的に理解できるように，学年が上がるにつれて許容され

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

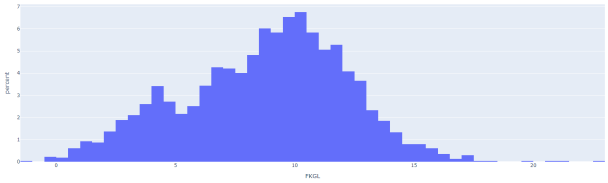

図 1　BNC における FKGL のヒストグラム．

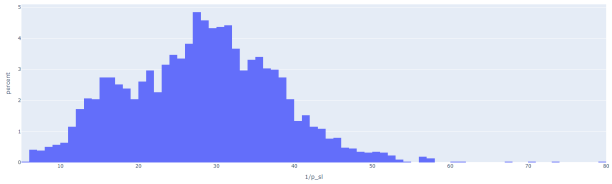

図 2　BNC における $1/p_{\mathrm{sl}}$ のヒストグラム．

る語彙が増加する．教材は各学年レベルでの語彙を増やすように作成されている．これは，1 文あたりの平均単語数だけではテキストの複雑さを測定できないことを示している．学習者の学年に応じた許容語彙を予測し，それを計画に組み込む必要がある．1 文あたりの平均単語数と受容語彙の変化を考慮したテキスト複雑性の計算は，元の式 1 のようなシンプルな式では実現できないように思える．

しかし，導出された式式 6 は全く異なる視点を提供する．これは，FKGL が 1 文あたりの平均音節数として解釈できることを示している．学習者の語彙数（語彙在庫）は学年が進むにつれて増加する．しかし，認識できる音韻の種類数（音韻数）は変化しない．音韻数は言語固有のものであり，一度母語を習得すると，その言語の母語話者における音韻数は固定される．受容可能な音韻数の大きさは一定であるため，1 文あたりの平均音節数の増加は，文の複雑さの増加を確実に表す．さらに，語彙とは異なり，音韻数は歴史的な時間の流れに堅牢である．「スマートフォン」のような単語はここ数十年で一般的になったが，この期間に音韻数が急激に増加または減少した言語は事実上存在しない．

FKGL に基づく 1 年あたりの音節数の増加　式 6 の式を用いることで，1 年あたりの音節数の増加も FKGL からモデル化できる．式 6 において，特定の学年に対する FKGL に注目する．次に，その 1 年上の学年（FKGL+1）に対する FKGL を考える．FKGL+1 について，M は一定であると仮定する．

FKGL+1 では，M は一定のままで，p_{sl} から p'_{sl} に変化するとする．このとき，以下の式が成り立つ．

$$\mathrm{FKGL{+}1} - c = \frac{1}{p'\mathrm{sl}} M \quad \mathrm{FKGL} - c = \frac{1}{p\mathrm{sl}} M \tag{9}$$

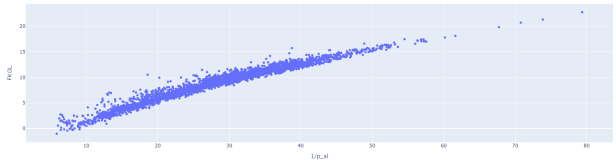

図 3　BNC における FKGL と $1/p_{\mathrm{sl}}$ の関係．

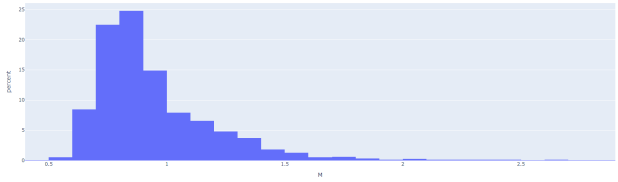

図 4　BNC における M のヒストグラム．

ここから，$\left(\frac{1}{p_{\mathrm{sl}}} - \frac{1}{p_{\mathrm{sl}}}\right) = \frac{1}{M}$ が得られる．$\frac{1}{p_{\mathrm{sl}}}$ は特定の学年における 1 文あたりの平均音節数を示し，$\frac{1}{p'_{\mathrm{sl}}}$ はその 1 年後の平均音節数を示す．これは，1 年あたりの 1 文あたりの平均音節数の増加を示している．さらに，$\frac{1}{M}$ は 1 年あたりの 1 文あたりの平均音節数の増加を表すことがわかる．

3　実験

設定　以上に基づき，実験について説明する．この実験では，British National Corpus (BNC)[8] を使用した．文中の平均単語数および平均音節数を決定するために，Python の可読性ライブラリを使用した．

まず，BNC における FKGL のヒストグラムを示す．図 1 にヒストグラムを示す．このヒストグラムは釣鐘型の分布を示している．

次に，本研究の主要な発見を示す．式 6 において，FKGL は再構築可能であり，入力テキストの主要な複雑さは 1 文あたりの平均音節数 $\frac{1}{p_{\mathrm{sl}}}$ によって表されることがわかった．BNC コーパスのテキストから，可読性ライブラリを用いて各テキストに対する $\frac{1}{p_{\mathrm{sl}}}$ を算出し，その結果のヒストグラムを作成した．図 2 のヒストグラムでは，横軸にテキストあたりの平均音節数，縦軸にパーセンテージを示す．図 2 も 図 1 と同様に釣鐘型の分布を示しており，テキストの複雑さが 1 文あたりの平均音節数によって適切に捉えられていることがわかる．

続いて，図 3 は，FKGL と 1 文あたりの平均音節数の関係を示す散布図である．図 3 から，FKGL と 1 文あたりの平均音節数の間には明確な相関があることがわかる．これは，FKGL において 1 文あたりの平均音節数がテキストの複雑さを表す上で重要な要素であることを裏付けている．

式 7 において M がほぼ一定であると仮定してい

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

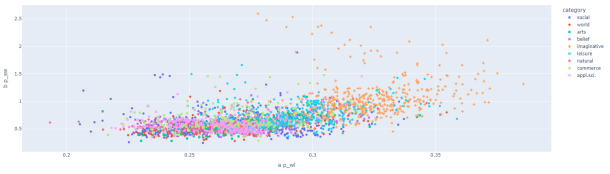

図 5　M とテキスト領域（カテゴリ）．横軸と縦軸の値の合計が M である．

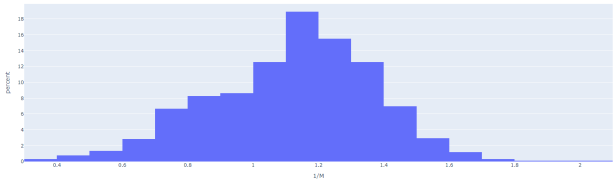

図 6　1 年あたりの 1 文あたりの平均音節数の増加を FKGL で予測した値に対応する $1/M$ のヒストグラム．

るが，これを示すため図 4 に M のヒストグラムを示す．横軸は M の値，縦軸はパーセンテージを示す．M のピークはおおよそ 1 付近に集まっている．式 6 によると，M は 1 文あたりの平均音節数に掛けられる唯一の要素であるため，1 文あたりの平均音節数がほぼ直接的に FKGL で利用されていることになる．実際，およそ 60% の M が 0.7 から 1.0 の範囲に収まっていることが確認された．さらに，裾野が広がっており，高い M の値が稀であることも示されている．

M のドメイン分析　BNC の主要な特性の一つは，その一般的な性質である．これは，コーパスがさまざまなトピックから収集された多様なテキストで構成されていることを意味する．BNC テキストのジャンルは「**ドメイン**」と呼ばれ，これらのテキストのおよそ 4 分の 3 がドメインタグ付けされている[1)]．図 5 は，式 6 における bp_{sw} と ap_{wl} の散布図を示している．横軸と縦軸を統合することで，先に述べたように M の要素が導出される．図 5 では，ドメインがテキストを限定された範囲に閉じ込めることが示されている．したがって，テキストのドメインが固定されている場合，M の値はより一貫性を保ち，大きく変動することはなくなる．その結果，式 6 における $\frac{1}{p_{\mathrm{sl}}}$（1 文あたりの平均音節数）がテキストの複雑さに影響を与える唯一の要素となる．$1/M$ のヒストグラムについて，式 2.1 によると，$1/M$ は 1 年あたりの 1 文あたりの平均音節数の増加として解釈できる．BNC における $1/M$ を導出し，そのヒストグラムを図 6 に示す．興味深いことに，図 6 は BNC における 1 年あたりの 1 文あたりの平均音節数の増加分布を示しており，ピークは 1.2 で，範囲は 0.4 から 2.0 である．我々の知る限り，FKGL を用いた統計的測定によってテキスト複雑性の増加が具体的に示された事例はこれまで存在しない．これは本研究の重要な発見である．

1)　ドメインが付与されていないテキストは，すべての実験から除外した．

4　他言語版 FKGL

他言語用の可読性指標を構築する議論においては，新しい可読性指標を提案してしまう研究が大半である．この方法では FKGL と比較した議論を行うなど解釈性が難しい．FKGL と対応させて日本語版 FKGL を求める研究として [9] が挙げられるが，この研究でも FKGL の文中平均単語数と単語中平均音節数に近い値を出す日本語テキストの変数を実験的にのみ示しており理論的背景に乏しい．

$$\text{他言語 FKGL} = d \times \text{文中平均音節数} \times \text{ジャンル定数} + e \tag{10}$$

式 10 において，d, e は言語によって決まる定数であり，ジャンル定数はテキストのジャンルによって定められる定数である．ジャンルが固定されれば，2 テキストあれば $d \times$ ジャンル定数,e の値は求める事ができる．結局，様々な特徴量を用いて精緻な予測モデルを組み立てたほうが性能が良くなることはその通りであるので，ジャンル定数の部分は直接ニューラルネットなどの現代的な手法を用いてテキストから直接値を求めてしまう手法も考えられる．具体的に日本の中学校国語の 2 年生（小学校 6 年分を足して FKGL8 相当）で扱われる「走れメロス」と，1 年生（FKGL7 相当) で扱われる「蜘蛛の糸」の 2 テキストを使って求めた所，$d \times$ ジャンル定数 = −0.0369, e = 8.84 であり，中学 1 年の「坊ちゃん」の推定学年は 7.57 で妥当であった．

5　おわりに

本研究は，Flesch-Kincaid 可読性式，特に FKGL と FRE に関する重要な貢献を行った．従来の自動可読性評価研究とは異なり，これらの式における 1 文あたりの平均音節数がテキストの複雑さを決定する重要な要因であることを示した．さらに日本語版 FKGL の提案も行い定性的に妥当性を確かめた．今後の展望として，他言語 FKGL についての詳細な研究が挙げられる．

謝辞　本研究は JSPS 科研費 22K12287 および JST さきがけ研究費 JPMJPR2363 の助成を受けた．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] J Peter Kincaid, et al. Development and test of a computer readability editing system (cres). final report, june 1978 through december 1979. 1980.

[2] Rudolph Flesch. A new readability yardstick. **Journal of Applied Psychology**, Vol. 32, pp. 221–233, 1948. Place: US Publisher: American Psychological Association.

[3] Teerapaun Tanprasert and David Kauchak. Flesch-kincaid is not a text simplification evaluation metric. In **Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)**, pp. 1–14, 2021.

[4] Joseph Marvin Imperial and Harish Tayyar Madabushi. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz, editors, **Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)**, pp. 205–223, Singapore, December 2023. Association for Computational Linguistics.

[5] Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. BLESS: Benchmarking large language models on sentence simplification. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 13291–13309, Singapore, December 2023. Association for Computational Linguistics.

[6] Joseph Marvin Imperial. Bert embeddings for automatic readability assessment. **arXiv preprint arXiv:2106.07935**, 2021.

[7] Sowmya Vajjala. Trends, limitations and open challenges in automatic readability assessment research. **arXiv preprint arXiv:2105.00973**, 2021.

[8] BNC Consortium. **The British National Corpus**. 2007.

[9] 赤木信也, 納富一宏. 英文と日本語文の両文に適応可能なリーダビリティ指標の検討. 情報科学技術フォーラム講演論文集, 第 14 巻, pp. 215–216, 2015.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).