E-009

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
A Study on Readability Index Adapted to both English and Japanese Sentences

赤木　信也† 納富　一宏†
Shinya Akagi Kazuhiro Notomi

1. はじめに

文章の読みやすさを計算機に推定させる研究（リーダビリティ研究）は，英語に対しては 1920 年頃から，日本語に対しては 1940 年頃から行われてきた．英語リーダビリティ指標としては, Flesch Reading Ease(FRE), Flesch Kincaid Grade Level(FKG),Automated Readability Index(ARI), Coleman Liau Index(CLI), SMOG などが提案されている．FRE は保険証書の読みやすさ保証を目的として州保険法に数値基準が規定されていたり，Microsoft Word の文章校正機能に組み込まれていたり，一般的指標として確立している．日本語リーダビリティ指標としては，①線形式に基づくもの，②語彙リストに基づくもの，③言語モデルに基づくものがそれぞれ提案されているが，何れも一般的指標として普及してはいない．

本研究は，より一般的な日本語リーダビリティ指標の開発を目的とし，英語リーダビリティ指標の「変数置き換え」によって，英文と日本語文の両文（英日両文）に適応可能な汎用性のあるリーダビリティ指標の検討を行うものである．これにより，英日両文の読みやすさを統一基準で判定・比較することができるようになる．

本稿では，「NHK ニュースで英会話」[1]の記事をプロの翻訳者による推敲済み文章とし，英文と日本語翻訳文のリーダビリティ指標の比較実験を行った結果について述べる．また，比較実験で得られた評価値について，英文と日本語翻訳文の評価値の差を考察し，リーダビリティ指標の結果も述べる．

2. 変数置き換え方法

英語リーダビリティ指標の変数置き換えに関しては，先行研究として，酒井[2], 赤木ら[3]の研究が存在する．

酒井はSMOGの変数置き換え方法として，多音節語の代わりに4文字以上の漢字数を用いる方法を提案しており，4文字以上の漢字数による読みやすさ判定の可能性を示している．

赤木らはFRE, FKG, ARI, CLI の変数置き換え方法として形態素分割と品詞結合，および，文字種の種類数に応じたシャノン情報量による重み付けを用いる方法を提案している．置き換えた指標をそれぞれjFRE, jFKG, jARI, jCLI と定義し，他の日本語リーダビリティ指標である「帯2」[4] と比較を行い，中程度の相関のある判定結果が取得できることを示している．

本稿では，文字種ごとの分割（字種分割）と再分割を用いて，英語リーダビリティ指標（FRE, FKG, ARI, CLI）を日本語に置き換えた指標（jFRE, jFKG, jARI, jCLI，以降は「提案指標」とする）を考える．提案指標は，英文に対しては従来指標と同じ評価値を得ることができ，評価式は式1～式4のように示される．

†神奈川工科大学　Kanagawa Institute of Technology

$$jFRE = 206.835-(1.015\times ASL)-(84.6\times ASW) \quad \text{.....(式1)}$$
$$jFKG = (0.39\times ASL)+(11.8\times ASW)-15.59 \quad \text{.....(式2)}$$
$$jARI = (4.71\times ACW)+(0.5\times ASL)-21.43 \quad \text{.....(式3)}$$
$$jCLI = (5.88\times ACW)-(29.6/ASL)-15.8 \quad \text{.....(式4)}$$

※ASL = Average Sentence Length
（字種分割数／センテンス数）
※ASW = Average number of Syllables per Word
（字種分割数と再分割数／字種分割数）
※ACW = Average number of Characters per Word
（シャノン情報量に基づく重み／字種分割数）

3. 英文と日本語翻訳文の比較実験

2014 年 4 月付けの「NHK ニュースで英会話」の記事（計 22 件）において，英文と日本語翻訳文のそれぞれに対して提案指標を用いた評価値の取得を行う．提案指標の取得は，字種分割と漢字 1 字単位の再分割を基本とし，ひらがな・カタカナについて，①再分割なし，②6 字単位の再分割，③5 字単位の再分割，④4 字単位の再分割，⑤3 字単位の再分割を行う，計 5 種類の方法を用いる．

比較実験の結果を図 1，図 2 に示す．

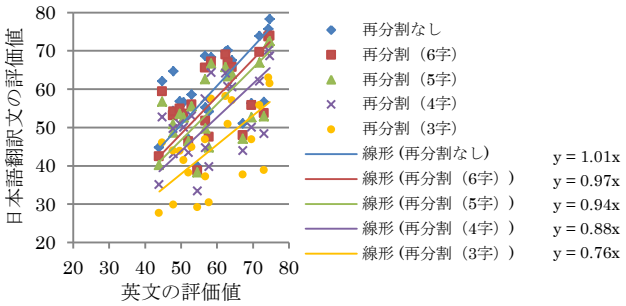

図 1　英文と日本語翻訳文の評価値の関係（jFRE）

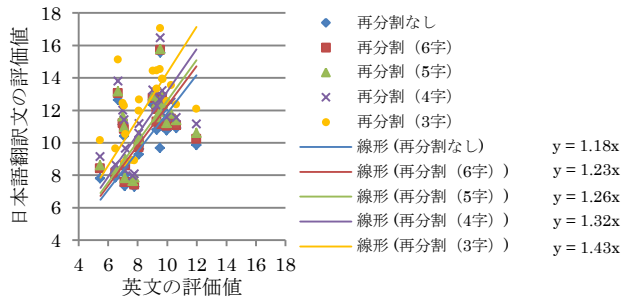

図 2　英文と日本語翻訳文の評価値の関係（jFKG）

本研究では，jFRE の結果（図 1）と jFKG の結果（図 2）において，1 に近い傾きの場合を「英日両文の適応において，提案指標の取得方法が最適である」とする．図 1 では，1 より大きい傾きの場合，日本語翻訳文の方が読みやすく，

1 より小さい傾きの場合，日本語翻訳文の方が読みにくく判定されていることを示す．図 2 では，1 より大きい傾きの場合，日本語翻訳文の方が読みにくく，1 より小さい傾きの場合，日本語翻訳文の方が読みやすく判定されていることを示す．

図 1，図 2 より，再分割なしの方法が 1 に最も近い傾きになっている．すなわち，英日両文の適応において，字種分割と漢字 1 字単位の再分割を用い，かつ，ひらがな・カタカナの再分割を行わない変数置き換え方法が最適であることが示唆されている．ただし，jFRE は 1 に近い傾きが得られているが，jFKG は最も 1 に近い傾きの値が 1.183 であり，1 との差が大きくなっている．

4. 考察

4.1 評価値の平均

比較実験において取得した「NHK ニュースで英会話」の評価値の平均を表 1 に示す．English は英文，それ以外は日本語翻訳文に対して，評価値を取得した結果である．

表 1　「NHK ニュースで英会話」の評価値

方法	ASW	ASL	jFRE	jFKG
English	1.581	14.000	58.851	8.529
再分割なし	1.468	21.848	60.452	10.255
再分割（6 字）	1.502	21.848	57.597	10.653
再分割（5 字）	1.525	21.848	55.610	10.931
再分割（4 字）	1.567	21.848	52.103	11.420
再分割（3 字）	1.653	21.848	44.848	12.432

一文あたりの平均単語数（ASL）の平均値は，英文より日本語翻訳文の方が 7.848 高い値を示している．

jFKG の平均値は，再分割なしの方法が英文に最も近い値を示しているが，それでも評価値の差が 1.726 ある．jFKG は式の関係上，ASL が 3 増えると値が 0.39×3 = 1.17 増えるような，ASL の影響を強く受ける指標である．そのため，ASL が高いことが jFKG の評価値の差の直接的な原因になっているものと考える．

4.2 評価値の差が大きい文章の調整

評価値の差が大きい文章として，2014 年 4 月 29 日の記事の分析を行った．「意図しないセンテンス数の計算」，「翻訳時の文章結合」が見受けられ，英文と日本語翻訳文でセンテンス数が異なっていた．

「意図しないセンテンス数の計算」では，略語である“U.S.”という語の“.”をセンテンスの区切りとして計算していた．そこで，“U.S.”を区切りとして認識しないように調整して，評価値の取得を行った．調整前の結果を「English」，調整後の結果を「English*」として，表 2 に示す．

「翻訳時の文章結合」では，言及対象が同じである文を結合して表現したり，話者情報（人名，組織名）と会話内容を別々に表現している文章されている文を結合して一文で表現したりしていた．そこで，英文のセンテンス数と同じになるように翻訳文を分割して，評価値の取得を行った．調整前の結果を「再分割なし」，調整後の結果を「再分割なし*」として，表 2 に示す．

表 2　評価値の差が大きい文章の調整結果

評価値	English	English*	再分割なし	再分割なし*
Num. of Letters	776	776	487	487
Num. of Sent.	12	11	7	11
Letters per Sent.	64.7	70.5	69.6	44.3
ASW	1.406	1.406	1.424	1.429
ASL	14.583	15.909	29.286	18.455
jFRE	73.109	71.764	56.607	67.246
jFKG	6.685	7.202	12.639	8.464

表 2 より，英文と日本語翻訳文のセンテンス数が同じになるよう調整した結果，ASL，jFRE，jFKG の誤差が大きく縮まり，英文と日本語翻訳文において同程度の評価値を取得することができている．

この結果より，評価値の差の原因として，提案指標の分割手法や重み付け手法の影響が小さいと言える．また，1 に最も近い傾きになっている再分割なしの方法について，センテンスの結合を文章が読みにくくなったと検出し，文章の調整によって同程度の評価値を取得できることが示されている．

4.3 リーダビリティ指標で評価できないこと

「翻訳時の文章結合」では，言及対象が同じであるセンテンスの結合を行っている場合があった．これは，重複語や代名詞を削減するという処理であり，意味の流れを整える処理，すなわち，広義では読みやすさを改善する処理として考えることができる．しかし，リーダビリティ指標の評価対象は「表層情報に基づく読みやすさ」であるため，意味の流れのような「意味に基づく読みやすさ」を評価することはできない．

4.4 今後の課題

表 2 では，センテンス数を調整しても ASL の差が 2.546 あり，日本語翻訳文の方が読みにくいと判定されている．この差はリーダビリティ指標自体が影響していると考えられるため，今後，ASL が低くなるように，提案指標の分割手法や重み付け手法を調整する必要がある．

5. まとめ

「NHK ニュースで英会話」の記事（計 22 件）の英文と日本語翻訳文に対し，字種分割を用いた変数置き換え方法による評価値の取得，および，評価値の比較実験を行った．その結果，英日両文の適応において，字種分割と漢字 1 字単位の再分割を用い，ひらがな・カタカナの再分割を行わない変数置き換え方法が最適であることが示唆された．

評価値の差を生む ASL の高さについて，リーダビリティ指標自体が影響していると考えられるため，今後の課題として，調整方法の検討と調整結果の検証が求められる．

参考文献

[1] NHK, “ニュースで英会話”, https://cgi2.nhk.or.jp/e-news/.
[2] 酒井由紀子, “患者向け説明文書の可読性判定”, 三田図書館・情報学会研究大会発表論文集, pp.45-48, (2006).
[3] 赤木信也，納富一宏, “英文の読みやすさ指標の拡張による日本語文の読みやすさ推定”, 電子情報通信学会総合大会発表論文集, (2015).
[4] 佐藤理史, “帯 2：日本語テキストの難易度推定”, http://kotoba.nuee.nagoya-u.ac.jp/sc/obi2/obi.html.