Vol.2022-NL-253 No.11
2022/9/30

変数置き換えモデルを用いた医薬品情報の可読性分析と検索件数を用いた複合名詞の文章平易化の検討

Readability analysis of medical document using variable replacement model and text simplification of compound norm using the number of searches

赤木　信也*
Shinya Akagi

*NTTデータ先端技術株式会社, Email: akagis@intellilink.co.jp

概要：英文と日本語文の両文に適用可能な可読性指標として, 変数置き換えモデルによる可読性指標を提案した.言語モデルである帯２との比較により, 日本語文を大まかに分類できること, 形態素分割より字種分割を用いる方法が最適であることが示された.また, 英文と日本語翻訳文の比較により, 日英両文に適用可能であること, 対応付けとして字種分割（ひらがな・片仮名の再分割なし）を用いる方法が最適であることが示された.更には, 医薬品添付文書とくすりのしおりの比較により, 古典的な手法よりも正確に判定できること, および英語圏の質保証基準を援用できることが示された. 最後に, 検索件数を用いた助詞『の』の自動補完による複合名詞の文章平易化を検討した結果, jFREの値が45未満の文章をjFREの値が45以上になるように平易化できることが示された. 形態素解析と複合名詞抽出を調整したところ, 形態素解析において辞書を用いず, 複合名詞抽出時に『一般』『固有名詞』『サ変接続』『接尾』に着目し, 先頭または末尾の形態素が1文字となる複合名詞を除外し, 4つ以上の形態素からなる用語を複合名詞に含め, 『-』や『/』が含まれる複合名詞を除外し, 検索データベースとしてBingを用いることで, 不必要な自動補完が少ない, 複合名詞の文章平易化を実現できることが示された.
キーワード：文章解析, 可読性, リーダビリティ, 変数置き換えモデル, 文章校正, 検索件数, 助詞補完

はじめに

現在, 日本語学習者層の拡大と多様化が進んでおり[1], この多様な方々に対して日本語のニュースや公的文書などを分かりやすく伝えることの重要性が高まっている. 特に健康保険関連の文章が分かりやすいことは重要であると考えられ, 米国では, 文章の読みやすさを定式化した『可読性指標』に基づいて文章の読みやすさを保証することが州の保険法に定められている.日本でも同様に, 可読性指標に基づいて文章の読みやすさを保証することができれば, 多様な方々にとって分かりやすい情報提供が実現されるようになると考える.
そこで, 本研究では, 簡便かつ汎用性のある可読性指標の開発を進めるとともに, 医薬品関連文書の可読性分析を通じて, 計算機を用いた文章の読みやすさ保証の有効性および評価基準に関する検証を実施する.更に, 文章校正に応用できないか検討する.

可読性

可読性とは, 文章の読みやすさのことである.文章の読みやすさに影響する項目としては, (1)文長, (2)語彙, (3)構造, (4)文体, (5)配置などが挙げられ, 文章の読みやすさを定式化して定量的に判定しようとする研究が存在する.定式化の研究では, 文章の意味的な内容は考慮せず, 文章の表層的な内容を考慮することで, 読みやすさの基準を作成している.
本研究では, 可読性を『人による視覚的な黙読しやすさ』と定義し, 文章の表層的な内容を考慮した可読性分析を実施する.

可読性指標

可読性指標とは, 文章の読みやすさを定式化して定量的に判定できるようにしたものである.可読性指標としては, 線形モデル, 語彙リストモデル, 言語モデル, 変数置き換えモデルが存在する.
英語の線形モデルの可読性指標としては, Flesh Reading Ease Score（FRE）[2], Flesh Kincaid Grade Level（FKG）[3], Automated Reading Index（ARI）[4], Coleman Liau Index（CLI）[5], SMOG[6]などが存在しており, FREは文章の読みやすさを保証することを目的として, 州の保険法に数値基準が定められている[7].FREの評価式は下記の通りである.

FRE = 206.835 - 1.015 wps - 84.6 spw
* wps = センテンスあたりの単語数
* spw = 単語あたりの音節数

FRE score	学年水準	評価基準
100-90	5年次	とても易しい
90-80	6年次	易しい

80-70	7年次	やや易しい
70-60	8年次・9年次	ふつう
60-50	10-12年次	やや難しい
50-30	大学	難しい
30-	大学卒業	とても難しい

日本語の可読性指標としては, 線形モデルの指標として建石ら[8], 李ら[9], 語彙リストモデルの指標として川村ら[10], 言語モデルの指標として佐藤ら[11]の指標が存在しているが, いずれも標準的指標としては確立していない.

日本語の特徴と可読性指標

日本語は英語と異なり, 分かち書きしないこと, ひらがな, カタカナのような表音文字と漢字のような表意文字が混在することといった特徴を持つ.分かち書きをしないため, 文章の区切りについて, 形態素単位の分割と文字種による分割の２通りを考えることができる.表音文字と表意文字が混在し, 英語の可読性指標よりも説明変数が増えるため, 指標自体および指標を用いた可読性分析の信頼性や妥当性を担保しづらくなっている.日本語の可読性の研究では, 主にひらがな・カタカナ・漢字に着目した指標が０から作成されているが, どの指標も標準化には至っていない.この原因の一つに, 日本語の複雑さに伴う信頼性や妥当性の担保しづらさが影響しているものと考える.
本研究では, 日本語の可読性指標を０から作成する方法では標準化には至らないのではないかと考え, 既に標準的指標が確立している英語の可読性指標を拡張する方法を考えた.

変数置き換えモデル

変数置き換えモデルとは, 既に標準的指標が確立している英語の可読性指標を拡張して, 英語と日本語の対応表を作成し, 値の対応付けを実施することにより, 英語の可読性指標の基準などをそのまま援用する方法である.
英語と日本語の両文に適用可能であり, かつ, 評価基準（質保証基準）を０から作成しなくて良いという特徴を持つ.
変数置き換えモデルとしては, SMOGの多音節数を4字以上の漢字の数で置き換える酒井の指標[12]と, 本研究で提案する, 単語数を文字種による分割語数で, 音節数を漢字の数で, 文字数をシャノン情報量に基づく重みで置き換える赤木の指標[13]が存在する.

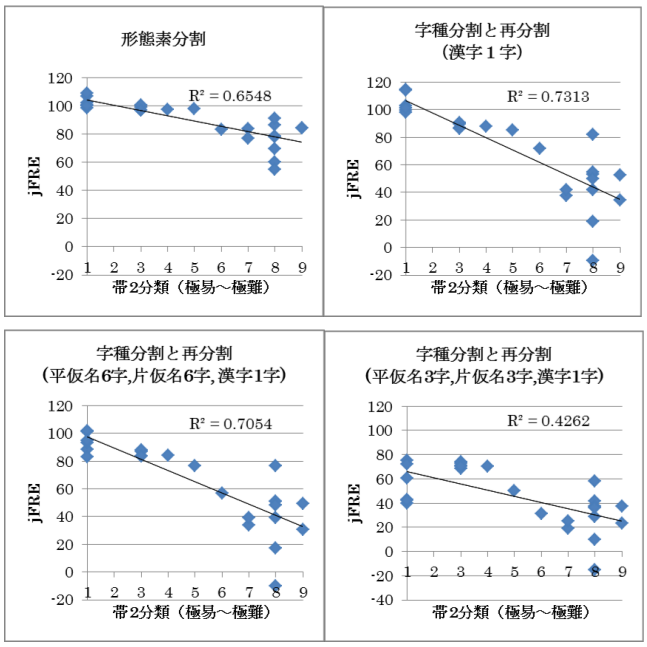
赤木の変数置き換え指標

本研究では, 英語の単語数を文字種による分割語数で, 英語の音節数を漢字の数, 文字数をシャノン情報量に基づく重みで対応づける変数置き換えモデルの指標（変数置き換え指標）を提案する.
FREを赤木の変数置き換えモデルで対応付けたものをjapanese FRE (jFRE)と呼ぶ.評価式は下記のとおりである.評価基準表はFREと同じものを使用する.

jFRE = 206.835 - 1.015 wps - 84.6 spw
* wps = センテンスあたりの単語数（センテンスあたりの字種分割語数）
* spw = 単語あたりの音節数（字種分割語あたりの漢字1字単位での再分割語数）

対応付けの妥当性に関する研究（1）

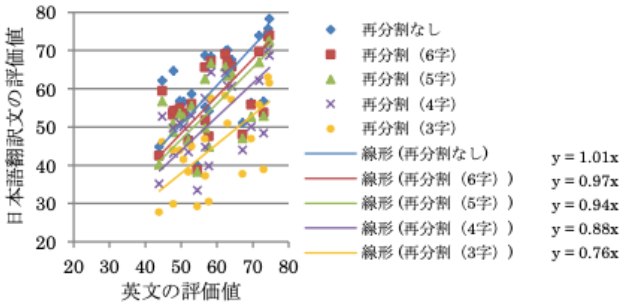
単語あたりの音節数の対応付け方法について, 形態素分割語数, 字種分割と漢字・ひらがな・片仮名の上限値による再分割を用いる方法を検討し, 計25件の日本語文(内訳:読み聞かせ文2件, 青空文庫の絵本2件, 青空文庫の小説8件, 法律3 件, 白書4件, コンピュータに関するWeb文章4件, コンピュータに関する学術論文2件)に対する可読性評価の結果を佐藤らの指標（帯２）との寄与率を用いて比較分析した[13].



その結果, 字種分割と漢字1字単位での再分割を用いる方法が最も寄与率が高い値（0.7313）を示し, 赤木の変数置き換え指標でも日本語文を大まかに分類できること, および, 形態素分割を用いる方法より字種分割を用いる方法の方が単語数および単語あたりの音節数の対応付けとして適切であることが示された.

対応付けの妥当性に関する研究（2）

「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ついて, 字種分割とひらがな・片仮名の上限値による再分割を用いる方法の比較実験を実施した[14].英文と日本語翻訳文は同じ内容を説明する文章であるため, 英語と日本語翻訳文の可読性指標の値が同じになると仮定し, 英語の可読性評価値と日本語翻訳文の可読性評価値を用いたグラフの傾きが1に近いほど対応付けが適切であるとする実験方法である.



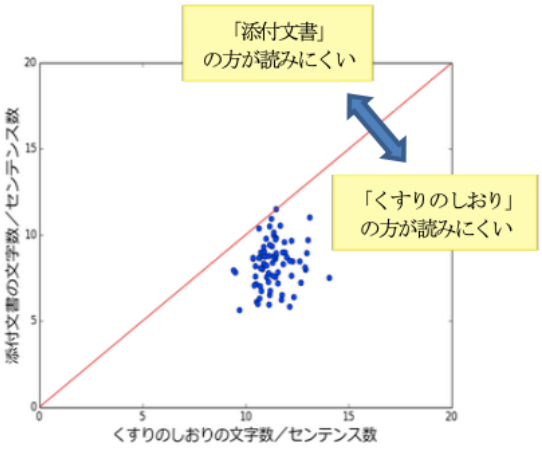
その結果, 英文の可読性評価値（FRE）と日本語翻訳文の可読性評価値（jFRE）について, 字種分割と漢字1字単位での再分割（ひらがな・片仮名の再分割なし）を用いる方法が最も傾きが1に近い値（1.01）を示し, 英文と日本語文の両文に適用可能な指標であること, および, ひらがな・片仮名の再分割を用いない方法が最適であることが示された.

古典的な可読性分析との比較および可読性評価基準の援用に関する過去研究

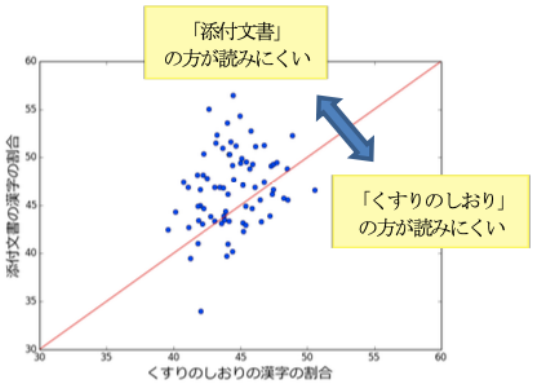
2014年4月~2015年3月の期間中, 「QLife お薬検索」[15]の月別薬品検索ランキング上位50件に1回以上現れた計74件の医薬品名を対象として, 医薬品添付文書[16]とくすりのしおり®[17]の可読性分析を実施した[18].くすりのしおり®は, 患者向けに読みやすくした文章であるため, 医薬品添付文書よりもくすりのしおり®の方が読みやすい文章であると判定されると仮定できる.

古典的な可読性分析手法であるセンテンスあたりの文字数, 漢字の割合を用いた可読性評価, およびjFREを用いた可読性評価の結果を以下に示す.

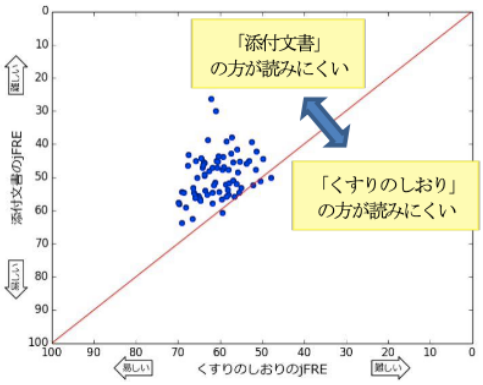
	文字数	文字数/センテンス数	漢字の割合	jFRE
医薬品添付文書	7327.0	20.8	46.8	49.7
くすりのしおり®	1407.4	27.4	44.4	60.8



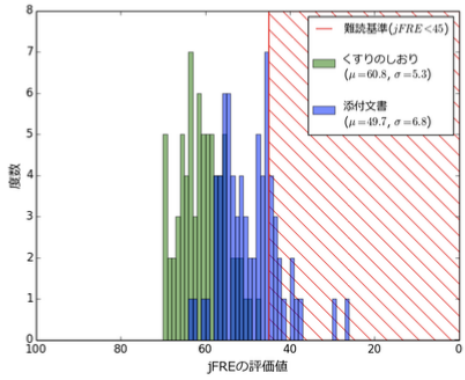
センテンスあたりの文字数については, くすりのしおり®の方が平均値が高く, 値が高いほど読みにくい文章とするため, くすりのしおり®の方が読みにくい文章であると判定されている.また, 個々の文章を比較しても, 74件中１件において, 医薬品添付文書の方が読みにくい文章であると判定されている.この結果より, センテンスあたりの文字数では適切な可読性評価ができない場合があることが示された.



漢字の割合については, 医薬品添付文書の方が平均値が高く, 値が高いほど読みにくい文章とするため, 医薬品添付文書の方が読みにくい文章であると判定されている.しかし, 個々の文章を比較すると, 74件中52件において, 医薬品添付文書の方が読みにくい文章であると判定されている.この結果より, 漢字の割合だけでは適切な可読性評価ができない場合があることが示された.



jFREについては, 医薬品添付文書の方が平均値が低く, 値が低いほど読みにくい文章とするため, 医薬品添付文書の方が読みにくい文章であると判定されている.個々の文章を比較しても, 74件中70件において, 医薬品添付文書の方が読みにくい文章であると判定されている.この結果より, jFREは古典的な分析手法より, 精度の高い可読性評価を実施できていることが示された.



医薬品添付文書とくすりのしおり®のjFREの値について, ヒストグラムを作成し, フロリダ州の州保健法におけるFREの質保証基準である45以上を援用し, 45未満の値をjFREの難読基準として図に示した.医薬品添付文書では, 74件中16件の文書が難読基準に該当し, くすりのしおり®では74件中0件の文書が難読基準に該当した.この結果より, jFREを用いることで, くすりのしおり®のような文章平易化を実施したものを定量的に読みやすい文章として判定できること, および英語圏の評価基準を問題なく援用できることが示唆された.

医薬品添付文書

『医薬品添付文書』とは, 医薬品に関する用量・用法・注意事項などが記載された文書のことであり, 薬事法において, 添付文書に記載すべき項目が指定されている.医薬品添付文書の項目は下記のとおりである.

・警告
・禁忌
・組成・性状
・効能・効果
・用法・用量
・使用上の注意
・（特定の背景を有する患者に関する注意）
・相互作用
・副作用
・（高齢者への投与）
・（妊婦, 産婦, 授乳婦等への投与）
・（小児等への投与）
・（過量投与）
・薬物動態
・臨床成績
・薬効薬理
・有効成分に関する理化学的知見
・承認条件
・包装
・主要文献

くすりのしおり®

『くすりのしおり®』とは, 医薬品添付文書を患者向けに分かりやすく要約した文章であり, 下記の項目から構成されている.

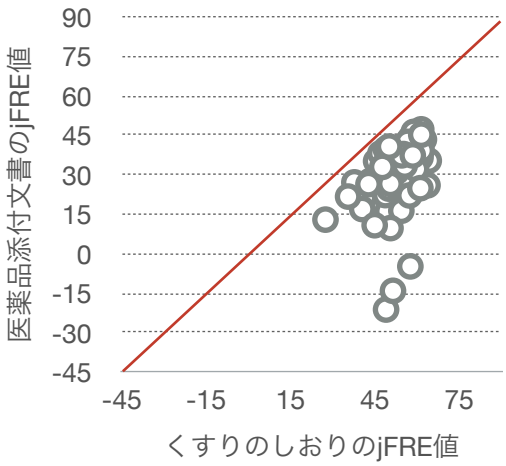
・商品名
・この薬の作用と効果について
・次のような方は使う前に必ず担当の医師と薬剤師に伝えてください
・用法・用量（この薬の使い方）
・生活上の注意
・この薬を使ったあと気をつけていただくこと（副作用）
・保管方法その他

可読性分析用データセット

Answer Newsに掲載された2018年度 国内医薬品売上高ランキング[19]のランキング上位50件の医薬品名について, 医薬品添付文書とくすりのしおり®の文章を抽出した.医薬品添付文書からは『警告』, 『禁忌』, 『効能・効果』, 『用法・用量』, 『使用上の注意』 『副作用』の文章を抽出し, くすりのしおり®からは『この薬の作用と効果について』, 『次のような方は使う前に必ず担当の医師と薬剤師に伝えてください』, 『生活上の注意』, 『この薬を使ったあと気をつけていただくこと（副作用）』の文章を抽出している.組成・性状, 相互作用など, 文章として抽出が難しかったり, 比較分析に使用しづらかったりした項目は一部抽出していない.また, 医薬品添付文書の表に記載されていた文章を抽出した際, 改行や「, 」で区切って体裁を整えており, 元の記載内容と構成が少し変わっている.しかしながら, 可読性評価および比較分析を実施する上で大きな影響はなく, 可読性分析だけでなく文章要約分析にも活用できる有用なデータセットになっていると考える.サイラムザ, リュープリン, アリムタの説明文が存在しなかったり, ランキング上で重複していた医薬品名を１つにまとめたりしたため, 医薬品添付文書の医薬品名は49件, くすりのしおり®の医薬品名は計46件となっている.
本データセットは Github上に公開している[20].

データセットの可読性分析結果

上記のデータセットを用いて医薬品添付文書とくすりのしおり®の可読性分析を実施した.
医薬品名計46件の可読性評価を実施した結果, 医薬品添付文書のjFRE平均値が27.1, くすりのしおりのjFRE平均値が49.0となっており, jFREは値が低いほど読みにくい文章であることを表すため, 医薬品添付文書の方が読みにくい文章であると判定された.



	jFRE平均値	jFRE45未満
医薬品添付文書	27.1	43/46 (93.5%)
くすりのしおり	49.0	8/46 (17.4%)

各文章のjFREの値を比較すると, いずれの文章でも医薬品添付文書の方が低い値（難しいこと）を示しており, くすりのしおり®が患者向けに文章を平易化できていることが定量的に示されている.
過去研究の文章では, くすりのしおり®のjFREの値が全て45以上だったが, 本研究の文章ではjFRE値が45未満となる文章が8件存在した.この結果は, フロリダ州の質保証基準を援用してjFREの値が45以上という数値基準を設けると, くすりのしおりの文章でも8件は不適切と判定されてしまうという結果であり, （1）変数置き換えモデルの対応付けに問題があること, （2）数値基準の値が不適切であること（3）くすりのしおりによる文章平易化の取組に改善の余地があること, いずれか, または複数を示唆している.過去研究において, 変数置き換えモデル自体の有効性と評価基準の援用の有効性が示唆されていたことを踏まえると, 今回の可読性評価の結果は, （3）くすりのしおりによる文章平易化の取組に改善の余地があることが示唆されているものと考えられる.

難しい文章の特徴

くすりのしおり®において, jFREの値が45未満となった文章は８件存在しており, jFREの最低値は27.4でオプジーボに関する文章であった.オプジーボに関する文章では, 副作用の項目において, 『全身倦怠感、むくみ、発汗、体重減少[甲状腺機能障害]』や『筋力低下、眼瞼下垂、呼吸困難、嚥下障害、筋肉の痛み、動悸、胸痛[重症筋無力症、心筋炎、筋炎、横紋筋融解症]』という記述が存在しており, 単語あたりの漢字の数が多いため, jFREの値がそれぞれ–119.7, –61.2を示していた.症例と疑われる病症名を列挙するだけでは, 使用される単語自体の難しさに起因した文章の難しさを改善しきれず, 文章全体が読みにくいままであると判定されてしまったものと考えられる.

文章平易化プロセスと改善案

文章平易化は, 重要情報抽出→言い換え→要約文作成の各段階で実施される.くすりのしおり®のオプジーボに関する文章では, 副作用の項目において, 症例と病症名が列挙されているだけとなっている.これは, 重要情報の抽出のみを実施され, 言い換えや要約文作成が未実施と

なっていることを示している.
文章平易化の改善案としては, 重要情報の抽出, 言い換え, 要約文作成それぞれの段階で改善を考えることができるが, くすりのしおり®の文章の場合は, 重要情報をこれ以上削減することはできないと考えられるため, 重要情報抽出時における改善は難しい.そこで, くすりのしおり®の文章平易化では, 言い換え時, 要約文作成時のいずれかで対処すべきだと考える.

言い換え時の文章平易化案

言い換え時の文章平易化としては, 『全身倦怠感』という表現を『全身の気だるさ』と言い換えることができると考えられる.このような言い換えを実施することにより, jFREの値は-119.7から-2.2となり, より良い文章平易化が実現される.

要約文作成時の文章平易化案

要約文作成時の文章平易化としては, 助詞などを補い, 『全身倦怠感、むくみ、発汗、体重減少の症状が現れた場合、甲状腺機能障害が疑われます.』という要約文を作成することができると考えられる.このような要約文作成を実施することにより, jFREの値は-119.7から11.3となり, 文章量は増えてしまうが, より良い文章平易化が実現される.

文章平易化の実現方法

文章平易化の実現方法については, 言い換え時の文章平易化として, 難易度辞書（パラレルコーパス）を用いる手法[21], 検索件数を用いて助詞の補完を行う手法[22], 要約文作成時の文章平易化として, 文章要約技術を応用する手法が考えられる.
ここで, 医薬品添付文書とくすりのしおり®の比較実験における難しい文章の特徴について考えると, 『全身倦怠感』や『甲状腺機能障害』など, 連続する漢字が用いられている.これらは, 『全身の倦怠感』や『甲状腺の機能障害』といったように, 助詞『の』を補完するだけでも文章がより平易になる.そこで本研究では, 文章平易化の恩恵が大きい複合名詞を対象として, 汎用性の高い助詞『の』の補完に絞って, 言い換え時の文章平易化を実現できないか検討した.

検索件数を用いた助詞の自動補完

まず, Mecab[23]とmecab-ipadic-neologd[24]を用いて文章の形態素解析を実施し, 複合名詞を抽出する.複合名詞は, 名詞（固有名詞, 一般, サ変接続, 数）の連続を対象とした.次に, 複合名詞の形態素の間に助詞『の』を追加した検索クエリ（補完用語）を作成して, Bing Web Search API[25]を用いて, Bing検索エンジンに対して完全一致検索を実施し, 検索件数を取得する.ここで, 検索件数が0件となった補完用語は, 助詞『の』による自動補完が誤りであるとして, 複合名詞の置換自体を行わない.そして, 検索件数が最も大きかった補完用語で複合名詞を置換する.

難読文として判定されたくすりのしおり®の文章における助詞の自動補完

jFREの値が45未満と判定された計8件のくすりのしおり®の文章について, 検索件数を用いた助詞の自動補完を実施した.各文章におけるjFREの値, 複合名詞の数, 助詞『の』が自動補完された数, 補完後のjFREの値を以下に示す.

くすりのしおり®	jFRE（Bing）	複合名詞の数	助詞自動補完の数	jFRE (補完後)
オプジーボ	27.4	50	41	46.4
イーケプラ	35.5	38	36	55.1
シムビコート	37.7	48	32	49.7
キイトルーダ	40.5	47	36	52.0
アバスチン	41.1	39	34	52.7
グラクティブ	42.7	39	34	53.1
セレコックス	43.4	29	22	51.2
プログラフ	44.9	32	26	52.7

検索件数を用いた助詞『の』の自動補完を実施することで, 難読文として判定されていたくすりのしおり®の文章を質保証基準を満たすjFRE45以上の値になるように平易化することができた.
助詞の自動補完については, 『全身倦怠感』が『全身の倦怠感』に,『医療担当者記入』が『医療担当者の記入』に, 『３週間間隔』が『３週間の間隔』に, 『1日1回服用』が『1日1回の服用』に置換され, 文章平易化につながる助詞の自動補完が行われていた.一方で, 補完用語の検索件数が0件のために, 『グルカゴン濃度低下作用』が『グルカゴン濃度の低下作用』に, 『口腔咽頭不快感』

が『口腔咽頭の不快感』に置換されなかったり, 補完用語の検索件数が0より大きかったために, 『声がれ』が『声のがれ』に, 『体重50kg』が『体重の50kg』に, 『L2』が『Lの2』に置換されたりすることがあり, 自動補完の方法に再調整が必要な複合名詞が存在した.

助詞の自動補完の改善案

今回は, mecab-ipadic-neologdを使用した形態素解析を実施していたが, 『悪性黒色腫』が形態素解析時に固有名詞として分類され, 自動補完の分析対象から外れていた.そのため, 分析対象を多く扱いたい場合は, 形態素解析時に, mecab-ipadic-neologdを使用しないことを検討した方が良いと考える.
今回は, 複合名詞抽出時の品詞項目に『数』を含めていたが, 数に関する用語は自動補完時に検索件数が0件になりにくく, 不必要な自動補完がされやすかった.そのため, 複合名詞抽出時に, 品詞項目『数』は除外した方が良いと考える.
『声がれ』や『L2』のように, 1文字の形態素を含む用語において不必要な自動補完がされやすかった.そのため, 複合名詞抽出時に, 1文字の形態素を含む用語は除外した方が良いと考える.

検索データベースによる違い

補完用語の検索件数は検索データベースの違いに左右されるため, 使用する検索データベースによって自動補完の精度に大きな差が生じる可能性が考えられる.そこで, Bing, Google, Wikipediaについて, 自動補完にどのような差が生じるか比較実験を行なった.Googleでは, Google Custom Search Engine API[26]を用いて, Google検索エンジンに対する完全一致検索を実行し, Wikipediaでは, MediaWiki API[27]を用いて, Wikipediaの記事に対する完全一致検索を実行した.
くすりのしおり®のオプジーボに関する文章ついて, 各検索データベースを使用した際の自動補完の数（置換数）, 自動補完後のjFREの値を以下に示す.

オプジーボ	複合名詞数	置換数	jFRE（補完後）
Bing	50	41	46.4
Google	50	45	46.7
wikipedia	50	27	41.5

自動補完の件数は, Bingが50件中41件, Googleが50件中45件, Wikipediaが50件中27件となっており, Googleの自動補完の件数が多く, Wikipediaの自動補完の件数が少なくなった.Googleのみが自動補完した用語としては, 『PD-1受容（体）』, 『小細胞肺癌』, 『4回静脈（内）』があり, Wikipediaのみが自動補完しなかった用語としては, 『腫瘍効果』, 『遠隔転移』, 『悪性胸膜中皮腫』, 『医療担当者記入』, 『（使った）あと気（をつけて）』, 『右季肋』, 『甲状腺機能障害』, 『下垂体機能障害』, 『腎障害』, 『副腎障害』, 『皮膚障害』, 『左上腹部』があった.
BingやGoogleでは, 『腎障害』を『腎の障害』に, 『右季肋』を『右の季肋』に置換するなど, 用語を非常に細かく分けるような自動補完が実施されていた.
Googleでは『４回静脈』を『４回の静脈』に, 『PD-1受容』を『PD-1の受容』に置換するなど, 不必要な自動補完が実施されていた.一方で, Bingで自動補完できなかった『小細胞肺癌』の自動補完が実施されていた.
Wikipediaでは, 用語を非常に細かく分けるような自動補完は減っているが, 『悪性胸膜中皮腫』を『悪性の胸膜中皮腫』に, 『医療担当者記入』を『医療担当者の記入』に置換できていなかった.補完後のjFREも45未満であることなどから, 文章平易化を実現するためにはデータ量が不十分だと考える.

形態素解析の調整

形態素解析において, mecab-ipadic-neologdを使用していると, 『悪性黒色種』や『自己免疫疾患』が固有名詞となり, 自動補完の対象外となっており, Wikipediaを使用した場合にjFREの値が改善しきれないことがあったため, 自動補完の対象となる複合名詞を増加させる目的で, mecab-ipadic-neologdを使用しない場合のデータを分析した.

オプジーボ	複合名詞数	置換数（Wiki）	置換数（Google）	置換数（Bing）	jFRE（Wiki）	jFRE（Google）	jFRE（Bing）
num	69	38	65	61	46.8	51.3	51.4
num-d	50	27	45	41	41.5	46.7	46.4

mecab-ipadic-neologdを使用しないことにより, 自動補完の対象となる複合名詞の数が50から69に増加し, 自動補完の数も10以上増加していた. しかし, 『間質性肺疾患』や『医療担当者記入』が複合名詞として抽出

されていなかったり，『PD-1受容』を『PD-1の受用』に，『T細胞』を『Tの細胞』に置換したりしており，自動補完の対象となる複合名詞は増加したが，不自然または不必要な自動補完も増加していた．この原因としては、複合名詞抽出時に品詞項目『接尾』が含まれていないこと，品詞項目『数』が含まれていること，1文字の形態素を利用していることが挙げられる．
Wikipediaを用いた自動補完について，jFREの値が45以上になっており，『 PD-1受容』が検索データベースに存在しないため置換されておらず，GoogleやBingよりは不必要な自動補完が実施されていなかったため，これで良いように思えるが，『間質性肺疾患』や『医療担当者記入』を複合名詞として抽出できず不自然な自動補完が実施されてしまっていたり，『悪性黒色腫』が検索データベースに存在しないため自動補完が実施されていなかったりしており，更なる調整が必要かつデータ量が不十分だと考える．

複合名詞抽出の調整（1）

形態素解析にmecab-ipadic-neologdを使用しないことで，自動補完の対象となる複合名詞を増やすことができたが，不自然または不必要な自動補完が含まれていた．そこで，不自然または不必要な自動補完が減るように品詞項目『接尾』を追加し，品詞項目『数』を除外したデータを分析した．

オプジーボ	複合名詞数	置換数（Wiki）	置換数（Google）	置換数（Bing）	jFRE（Wiki）	jFRE（Google）	jFRE（Bing）
nonum	69	39	65	62	46.8	51.3	51.4
nonum-suffix	100	39	91	77	45.8	58.3	55.0
num	69	38	65	61	46.8	51.3	51.4

品詞項目『数』を除外することによって，『PD-1受用』が『PD-』が複合名詞として抽出され，『PDの-』という自動補完が実施されていた．『-』の品詞特性が名詞，サ変接続なので，複合名詞として抽出されてしまっており，更なる対策が必要だと考える．
品詞項目『接尾』を追加することによって，『間質性肺疾患』や『医療担当者記入』を複合名詞として抽出できており，自然な自動補完が可能になっていた．しかし，『静脈内』を『静脈の内』に，『回あたり』を『回のあたり』に，『生活上』を『生活の上』に置換しており，別途不必要な自動補完が増えてしまっていた．

複合名詞抽出の調整（2）

品詞項目『接尾』を追加することによって，『間質性肺疾患』や『医療担当者記入』を複合名詞として抽出できるようになったが，不必要な自動補完も増えてしまっていた．不必要な自動補完の多くは1文字の形態素を含む複合名詞であった．そこで，先頭または末尾の形態素が1文字となる複合名詞を除外したデータを分析した．

オプジーボ	複合名詞数	置換数（Wiki）	置換数（Google）	置換数（Bing）	jFRE（Wiki）	jFRE（Google）	jFRE（Bing）
nonum-suffix	100	39	91	77	45.8	58.3	55.0
nonum-suffix2	38	20	35	35	38.3	43.7	44.2

複合名詞の数は大幅に削減され，Googleを用いた自動補完の場合でも，jFRE45未満になってしまうような自動補完しか実施できなかった．『PD-』や『静脈内』，『回あたり』などを複合名詞として抽出しないため，不必要な自動補完が大幅に削減された一方で，自動補完して欲しい『全身倦怠感』や『悪性黒色腫』なども複合名詞として抽出されないため，自動補完の対象が少なくなりすぎてしまっていた．

複合名詞抽出の調整（3）

先頭または末尾の形態素が1文字となる複合名詞を除外したことによって，不必要な自動補完が大幅に削減された一方で，『全身倦怠感』などの自動補完が必要な用語が自動補完の対象外となってしまっていた.そこで，『全身倦怠感』などを自動補完の対象とするために，先の条件に加えて，3つ以上の形態素からなる用語を複合名詞に含めるデータを分析した．

オプジーボ	複合名詞数	置換数（Wiki）	置換数（Google）	置換数（Bing）	jFRE（Wiki）	jFRE（Google）	jFRE（Bing）
nonum-suffix3	68	29	61	58	42.9	51.2	51.3

複合名詞の数は68件程度に増加し, Wikipediaを用いた場合ではjFREの値が45未満となっていたが, GoogleやBingの場合ではjFREの値が45以上となるような自動補完が実施されていた. 『悪性黒色腫』が『悪性の黒色腫』に、『頭頸部癌』が『頭頸部の癌』に, 『悪性胸膜中皮腫』が『悪性の胸膜中皮腫』に, 『全身倦怠感』が『全身の倦怠感』に置換されており, 自動補完が必要な用語の多くをカバーすることができている. この結果は不必要な自動補完も大幅に削減されていることから, 形態素解析, 複合名詞抽出において, 品詞特性『一般』『固有名詞』『サ変接続』『接尾』に着目し, 先頭または末尾の形態素が1文字となる複合名詞を除外して, 3つ以上の形態素からなる用語を複合名詞に含め, 検索データベースとしてGoogleまたはBingを用いることで, 文章平易化として十分な量の助詞『の』の自動補完が実施されたと言える. 逆にWikipediaは, 不必要な自動補完は少ないものの, 文章平易化を実現するためにはデータ量が不十分だと言える.

検索データベースの選択と形態素解析および複合名詞抽出の調整

形態素解析と複合名詞抽出を調整し, 大方の不必要な自動補完を除外した. しかしながら, 検索データベースとしてGoogleを使用すると, 『mg/kg』を『mgの/kg』に, 『患者さん』を『患者のさん』に, 『PD-L』を『PDの-L』に, 『髄膜炎』を『髄膜の炎』に置換してしまうなど, 不必要な自動補完が増加する傾向にあるのは変わらなかった. Bingを使用しても, Google同様に『髄膜炎』を『髄膜の炎』に, 『部分発作』を『部分の発作』に, 『配合吸入薬』を『配合の吸入薬』に, 『-グルコシダーゼ阻害剤』を『-のグルコシダーゼ阻害剤』に置換してしまう部分があり, 不必要な自動補完は存在していた.

形態素解析や複合名詞抽出の更なる調整としては, 形態素解析時に『髄膜』を追加したり, 複合名詞抽出時に『-』や『/』を含む用語を除外したり, 複合名詞に含める形態素数を3つ以上から4つ以上に変更したりすることが考えられる, また, 不必要な自動補完が最も少なくなる先頭または末尾の形態素が1文字となる複合名詞を除外する（2）までの段階で調整を止める方法を採用したり, 不必要な自動補完が更に少ないWikipediaを基本的に使用して, Google, Bingを一部併用する手法を検討したりすることも考えられる.

文章平易化のことも考慮すると, 複合名詞に含める形態素数を3つ以上から4つ以上に変更してjFREの値が45以上になるか検討すべきだと考える. また, Wikipediaを基本的に使用して, 形態素数が4つ以上の用語はGoogleまたはBingを用いて自動補完を実施する方法も検討すべきだと考える.

複合名詞抽出の調整（4）

3つ以上の形態素からなる用語を複合名詞に含めた場合, 『悪性黒色種』などが置換された一方で, 『腱鞘炎』などが不必要な自動補完として実施されてしまっていた. そこで, 複合名詞に含める形態素を3つ以上ではなく, 4つ以上にしてjFREの値が45以上になるかデータを分析した.

nonum-suffix4	複合名詞数	置換数（Wiki）	置換数（Google）	置換数（Bing）	jFRE（Wiki）	jFRE（Google）	jFRE（Bing）
オプジーボ	52	20	46	44	38.3	46.9	47.3
イーケプラ	31	12	28	23	42.2	47.9	45.1
シムビコート	33	10	31	24	47.6	59.7	51.8
キイトルーダ	49	19	42	39	50.2	55.6	55.2
アバスチン	44	17	36	33	51.2	56.0	54.8
グラクティブ	35	19	34	31	48.6	54.8	53.1
セレコックス	31	12	25	24	48.7	56.0	54.2
プログラフ	39	14	37	28	52.4	56.6	55.7

GoogleまたはBingを使用した場合に, jFREの値が45以上になることが示された. 『悪性黒色種』などは自動補完の対象外となってしまうが, 『髄膜炎』などの不必要な自動補完が実施されなくなっている. また, Bingを使用することで『患者さん』を『患者のさん』に置換されないようにできる. 以上より, 形態素解析, 複合名詞抽出において, 品詞特性『一般』『固有名詞』『サ変接続』『接尾』に着目し, 先頭または末尾の形態素が1文字となる複合名詞を除外して, 4つ以上の形態素からなる用語を複合名詞に含め, 検索データベースとしてBingを用いる手法が文章平易化を実現する上で最適の調整だと考える.

複合名詞抽出の調整（5）

前項の自動補完を更に調整し，『-グルコシダーゼ阻害剤』を『-のグルコシダーゼ阻害剤』に置換されないように，『-』や『/』を含む用語を複合名詞から除外したデータを分析した.

nonum-suffix5	複合名詞数	置換数（Wiki）	置換数（Google）	置換数（Bing）	jFRE（Wiki）	jFRE（Google）	jFRE（Bing）
オプジーボ	52	20	46	44	38.3	46.9	47.3
イーケプラ	28	12	25	23	42.2	47.9	45.1
シムビコート	33	10	31	24	47.6	59.7	51.8
キイトルーダ	47	19	42	39	50.2	55.6	55.2
アバスチン	44	17	36	33	51.2	56.0	54.8
グラクティブ	34	18	33	30	48.6	54.8	53.1
セレコックス	31	12	25	24	48.7	56.0	54.2
プログラフ	35	14	33	28	52.4	56.5	55.7

複合名詞数がいくらか減少したが, jFREの値については特に変化がなく, GoogleまたはBingを使用した場合は質保障基準であるjFRE45以上を満たす結果となった. 検索データベースとしてGoogleを使用した場合は『患者さん』を『患者のさん』に置換してしまうが, Bingを使用すれば『-グルコシダーゼ阻害剤』を『-のグルコシダーゼ阻害剤』に置換されることがなくなり, 不必要な自動補完の少ない文章平易化を実現できることが示された.

WIKIPEDIAを基本としたシステム

検索データベースとしてWikipediaを用いることで, 文章平易化のためのデータ量は不十分になるが, 不必要な自動補完を少なくすることができる. また, 検索データベースとしてGoogleやBingを用いる場合, 検索のアクセスに上限が設けられており, 大量の処理に向いていないという特徴がある. そこで, 基本的にはWikipediaを参照し, Wikipediaで自動補完できなかった4つ以上の形態素からなる用語を別途GoogleやBingを用いて自動補完を試みるシステムが検討できる. 文章平易化として有効かどうか調整の余地があるが, 試す価値はあると考える.ただし, システム化についてはシステム構成やUIの話となり, 今回実施したソフトウェア上の調整に留まらないため, 今回の実験と分けて今後の課題として実施する.

今後の展望

可読性評価指標としては, 人による判定と異なる結果を示すことがあり, 難読語の数がその要因であると考えられている[28].そこで, 漢字１字単位の再分割に加えて, 漢字検定基準などを用いて漢字ごとに２～４倍程度の重みをつけることで人による判定に近づけることができるか検討する必要があると考える.
可読性評価ツールとしては, 現在いづれのツールも公開していないため, 字種分割ツールとともに可読性評価ツールを公開し, 誰でも気軽に可読性評価が実施できる環境を整備する必要があると考える.
文章校正としては, 助詞『の』以外の自動補完についても検討するとともに, 今回は完全一致検索の検索件数が0件以上であれば自動補完を実施したが, 外れ値を除外するために検索件数が一桁台のものを除外するなどの工夫についても検討する必要があると考える.
文章校正システムとしては, GoogleやBingのデータベースを用いた助詞『の』の自動補完ではアクセス制限が存在し, かつ不必要な自動補完がされやすい傾向にあったため，Wikipediaを基本としたシステム構成を検討したり, 新たに類語辞書と可読性指標を組み合わせたパラレルコーパスを用いた言い換えや, 文章要約技術に可読性指標を組み合わせる方法を検討したりして, より読みやすくなる文章を提案できるようなシステムの構築を検討する必要があると考える.

まとめ

英文と日本語文の両文に適用可能な可読性指標として, 変数置き換えモデルによる可読性指標を提案した.言語モデルである帯２との比較により, 日本語文を大まかに分類できること, 形態素分割より字種分割を用いる方法が最適であることが示された.また, 英文と日本語翻訳文の比較により, 日英両文に適用可能であること, 対応付けとして字種分割（ひらがな・片仮名の再分割なし）を用いる方法が最適であることが示された.更には, 医薬品添付文書とくすりのしおりの比較により, 古典的な手法よりも正確に判定できること, および英語圏の質保証基準を援用できることが示された.
検索件数を用いた助詞『の』の自動補完による複合名

詞の文章平易化を検討した結果, jFREの値が45未満の文章をjFREの値が45以上になるように平易化できることが示された. 形態素解析と複合名詞抽出を調整したところ, 形態素解析において辞書を用いず, 複合名詞抽出時に『一般』『固有名詞』『サ変接続』『接尾』に着目し, 先頭または末尾の形態素が1文字となる複合名詞を除外し, 4つ以上の形態素からなる用語を複合名詞に含め, 『-』や『/』が含まれる複合名詞を除外し, 検索データベースとしてBingを用いることで, 不必要な自動補完が少ない, 複合名詞の文章平易化を実現できることが示された.

参考文献

[1]文化庁, 令和2年度国内の日本語教育の概要: https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/nihongokyoiku_jittai/r02/
[2] Flesch, Rudolph: A new readability yardstick., Journal of applied psychology, Vol.32(3), pp.221-223, (1948).
[3]Kincaid,J.P., Fishburne Jr,R.P., Rogers,R.L., Chissom,B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, No.RBR-8-75, Naval Technical Training Command Millington TN Research Branch , (1975).
[4] SENTER,R.J., SMITH,E.A.: Automated readability index., CINCINNATI UNIV OH, (1967).
[5] Coleman,Meri, Liau,T.L.: A computer readability formula designed for machine scoring., Journal of applied psy-chology, Vol.60(2), pp.283-284, (1975).
[6] McLaughlin, G. Harry.: SMOG grading: A new readability formula, Journal of reading, Vol.12(8), pp.639-646, (1969).
[7]The Florida Senate: Chapter 627 Section 41454 Florida Statutes, https://www.flsenate.gov/Laws/Statutes/2021/627.4145, (2022/7/29アクセス).
[8]建石由香, 小野芳彦, 山田尚彦: 日本文の読みやすさの評価式, 情報処理学会研究報告ヒューマンコンピュータインタラクション(HCI), Vol.25, pp.1-8, (1988).
[9]李在鎬: 日本語文章難易度判別システム (jReadability) alpha版, http://jreadability.net/, 科学研究助成事業:読解教育支援を目的とする文章難易度判別システムの開発, (2013)
[10]川村よし子: 語彙チェッカーを用いた読解テキストの分析, 早稲田大学日本語研究教育センター講座日本語教育, 第 34 分冊, pp.1-22, (1998).
[11]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).
[12]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).
[13]赤木信也, 納富一宏: 字種分割を用いた日本語リーダビリティ判定システムの開発, 電気学会 電子・情報・システム部門大会講演論文集, pp.1192-1197,2015.
[14]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, 情報処理学会 第 14 回情報科学技術フォーラム（FIT2015）講演論 文集, 第 2 分冊, pp.215-216, 2015.
[15]QLife お薬検索, http://www.qlife.jp/meds/.
[16]医薬品医療機器総合機構, 添付文書情報メニュー, https://www.info.pmda.go.jp/psearch/html/menu_tenpu_base.html.
[17]くすりの適正使用協議会:くすりのしおり®, http://www.rad-ar.or.jp/siori/index.html.
[18]赤木信也*, 納富一宏, 斎藤恵一:"変数置き換えモデルを用いた医療関連文書の可読性分析", バイオメディカル・ファジィ・システム学会誌, Vol.19, No.1, pp.19-27, (2017.06).
[19]Answer News, 2018年度 国内医薬品売上高ランキング, https://answers.ten-navi.com/pharmanews/16487/
[20]Github, medical_documents: https://github.com/ShinyaAkagil/medical_documents
[21]梶原智之, 西原大貴, 小平知範, 小町守: 日本語の語彙平易化のための言語資源の整備. 自然言語処理, Vol.27, No.4, pp.801-824, December 2020.
[22]池田和史,柳原正, 服部元, 松本一則, 小野智弘: 口語文書の解析精度向上のための助詞落ち推定および補完手法の提案, 情報処理学会研究報告, 2010年度 (4), 1-8, 2010-12.
[23]MeCab: https://taku910.github.io/mecab/
[24]Github, mecab-ipadic-neologd: https://github.com/neologd/mecab-ipadic-neologd
[25]Bing Web Search API: https://docs.microsoft.com/ja-jp/azure/cognitive-services/bing-web-search/overview.
[26]Google Custom Search API: https://developers.google.com/custom-search/v1/introduction.
[27]MediaWiki API: https://ja.wikipedia.org/w/api.php
[28]赤木信也 納富一宏:"リーダビリティ指標を用いた文章評価システムの開発：計算機と大学生による可読性評価の比較", 情報処理学会 第15回情報科学技術フォーラム(FIT2016)講演論文集, 第4分冊, N-007, pp.293-294, (2016.09)