spaCy と BERT を用いた関係詞の抽出とそれに基づくリーダビリティ指標の算出モジュールの開発

相良碧[1]　中野明[1]

概要：本研究の目的は，関係詞に着目した用法判定器の構築と，構造的複雑さを考慮したリーダビリティ指標の提案ならびに算出モジュールの開発である．用法判定には，spaCy と BERT を併用し，関係詞の省略を含む 158 文のテストで正答率 100%であった．この判定器を用いて，単語難易度（CEFR-J 準拠）と文法構造難易度（DLT 理論準拠）を統合した指標を提案している．提案手法により算出した値と教科書データの学年進行との相関は 0.769 となり，従来の単語のみの指標を上回る精度であった．

キーワード：言語解析，テキスト処理，語学学習

Development of a Readability Assessment Module Based on Relative Clause Extraction Using spaCy and BERT

AOI SAGARA[†1]　AKIRA NAKANO[†1]

Abstract: This research proposes a new readability assessment model that accounts for the complexity of relative clauses by utilizing spaCy, BERT, and Dependency Locality Theory (DLT). Compared with conventional metrics based solely on word difficulty, the proposed approach demonstrates a higher correlation with the grade levels of Japanese textbooks.

Keywords: Natural language analysis, Text processing, Computer-assisted language learning,

1. 緒言

近年，グローバル化が急速に進展する中で，特定の専門分野に限らず幅広い場面において英語能力が求められている．しかしながら，現在の日本の学校教育では，小学校・中学校・高等学校という校種間の接続に依然として課題が残されている[1]．この接続を円滑にし，一貫した教育を行うことは，学習指導要領が掲げる「外国語を用いて積極的にコミュニケーションを図ろうとする態度」や「情報を的確に理解し伝達する能力」を育成するために極めて重要である．

この校種間の接続を強化するためには，学習者の習熟度に合わせて段階的に難易度を調整した教材提供が不可欠である．そのためには，単語レベルの難易度だけでなく，文構造の読みやすさを客観的に評価するリーダビリティ判定器の開発が求められる．しかし，正確なリーダビリティを算出するためには，表面的な単語数だけでなく，複雑な文構造を正確に検出する判定器が必要となる．

そこで本研究では，日本人学習者が特に習得困難な文法要素の１つである関係詞に着目し，以下の２点を目的としてモジュールの開発を行った．

1. 関係詞の用法を自動的かつ正確に抽出・分類する判定器の開発．
2. 関係詞の構造的複雑さを考慮した，リーダビリティ算出モジュールの開発．

特に第２の目的に関しては，既存のリーダビリティ指標として投野らが作成した CEFR-J[2]，および小篠らが開発した教科書分析指標である OFYL（Ozasa-Fukui Year Level)[3]を参考に開発を行った.近年の自然言語処理ではリーダビリティ算出に深層学習を用いる手法も存在するが，判定の根拠が不透明になりやすいという課題がある．教育現場での利用を考慮すると，判定結果に対する説明可能性が重要である．そこで本研究では，関係詞の用法検出の一部に BERT を用いる一方で，リーダビリティ値の算出そのものには機械学習モデルを用いず，抽出された言語特徴量に基づく明示的な計算式を採用した．これにより，読みやすさの判定根拠が明確で，透明性の高いリーダビリティ算出手法を提案する．

1 久留米工業高等専門学校
National Institute of Technology, Kurume College, Hukuoka 830-8555, Japan

2. spaCy と BERT を用いた関係詞判定

2.1 判定器概要

本判定器は，関係詞の多角的な解析を実現するために，基本的な構文解析と判定を担当する spaCy（モデル名：en_core_web_trf）と，関係代名詞・関係副詞の省略形の推論を担当する BERT（モデル名：bert-large-uncased），および関係代名詞 what の識別を行うファインチューニング済みの BERT の 3 つの機能を統合して構築した．これらを組み合わせた判定器の処理の流れを図 1 に示す．

また，本判定器では，spaCy の係り受け解析結果に基づき，関係代名詞（that および wh-語の主格・目的格・所有格）と関係副詞を判定対象とした．また，リーダビリティに影響を与える要素として，修飾関係（名詞・代名詞・文修飾），非制限用法，および前置詞の位置に基づく文体（Formal/Casual）の分類も行う．さらに，類似構文との誤判定を防ぐため，同格の that や間接疑問文の識別機能も実装した ．

2.2 spaCy による判定ロジック

本判定器では spaCy の係り受け解析結果(依存関係ラベル,品詞タグ,構文木)を利用して判定を行った．表 1 に，各関係詞の判定条件を示す．なお，表中の主要部とは構文木上の親のことを指す．

表 1　判定と依存関係ラベル(関係詞 what を除く)

文法	関係詞の依存関係ラベル	関係詞の主要部の依存関係ラベル	関係詞の品詞タグ
関係代名詞(主格)	nsubj, nsubjpass	relcl, advcl, acl	WP, WDT
関係代名詞(目的格)	dobj, pobj, dative, mark	relcl, advcl, acl	WP, WDT
関係代名詞(所有格)	poss	relcl, advcl, acl	WP$
関係代名詞 that(主格)	nsubj, nsubjpass, attr	relcl, acl	
関係代名詞 that(目的格)	dobj, pobj, dative	relcl, acl	
関係代名詞 that(目的格)	mark	relcl	
関係副詞		relcl, advcl, acl	WRB

表 1 にて判定していない関係詞の省略形及び関係代名詞 what に関しては，表層的な構文構造のみに依存するルールベースの手法（spaCy）では識別が困難である．これらの課題に対し，BERT を用いたアプローチで解決を図った．詳細は 2.3，2.4 に示す．

2.3 関係詞の省略形判定

関係詞が省略された文において，その省略語が「関係代名詞」か「関係副詞」かを特定するため，BERT（モデル名：bert-large-uncased)の推論能力を活用した.具体的には，spaCy で関係節構造（relcl）を持つが関係詞が存在しない箇所を特定し，先行詞と節の間に［MASK］トークンを挿入して BERT に入力する．BERT は文脈確率から単語を予測するが，文脈に関わらず汎用的な that が最上位（Top-1）に出力される傾向が強いため,本システムではあえて that を除外し，次点以降（Top-2）の候補語（when, where, which 等）を採用することで正確な用法を特定するロジックを実装した．

2.4 関係代名詞 what の判定

関係代名詞 what（例：*This is what I wanted.*）と間接疑問文の what（例：*I asked what he wanted.*）は，構文木上での構造的差異が極めて乏しく，ルールベースでの判別は不可能である．そこで本研究では，BERT（モデル名：bert-base-uncased)を特定のタスク用にファインチューニングすることで，文脈に基づく 2 値分類器を構築した．学習データには，Wikipedia から抽出した文に対し手動でラベル付けを行った計 667 文（関係代名詞：449 文，間接疑問文：218 文）を使用した．本判定器は，spaCy が判定不能とした what を含む文に対し，この分類器を適用することで最終的な用法を決定する．

2.5 付随する判定事項

本判定器は，修飾関係（名詞・代名詞・文修飾），非制限用法，および前置詞の位置に基づく文体（Formal/Casual）の分類を行う．修飾関係判定では先行詞の品詞タグ（NOUN, PRON 等）を参照して分類を行い，非制限用法判定では関係詞直前のトークンが句読点(コンマ)であるかを検査する．文体の判定においては，関係詞と前置詞の係り受け構造を確認し，前置詞が関係詞節内に残留しているか（Preposition Stranding），関係詞の前に移動しているか（Pied-piping）によって判定を行う．

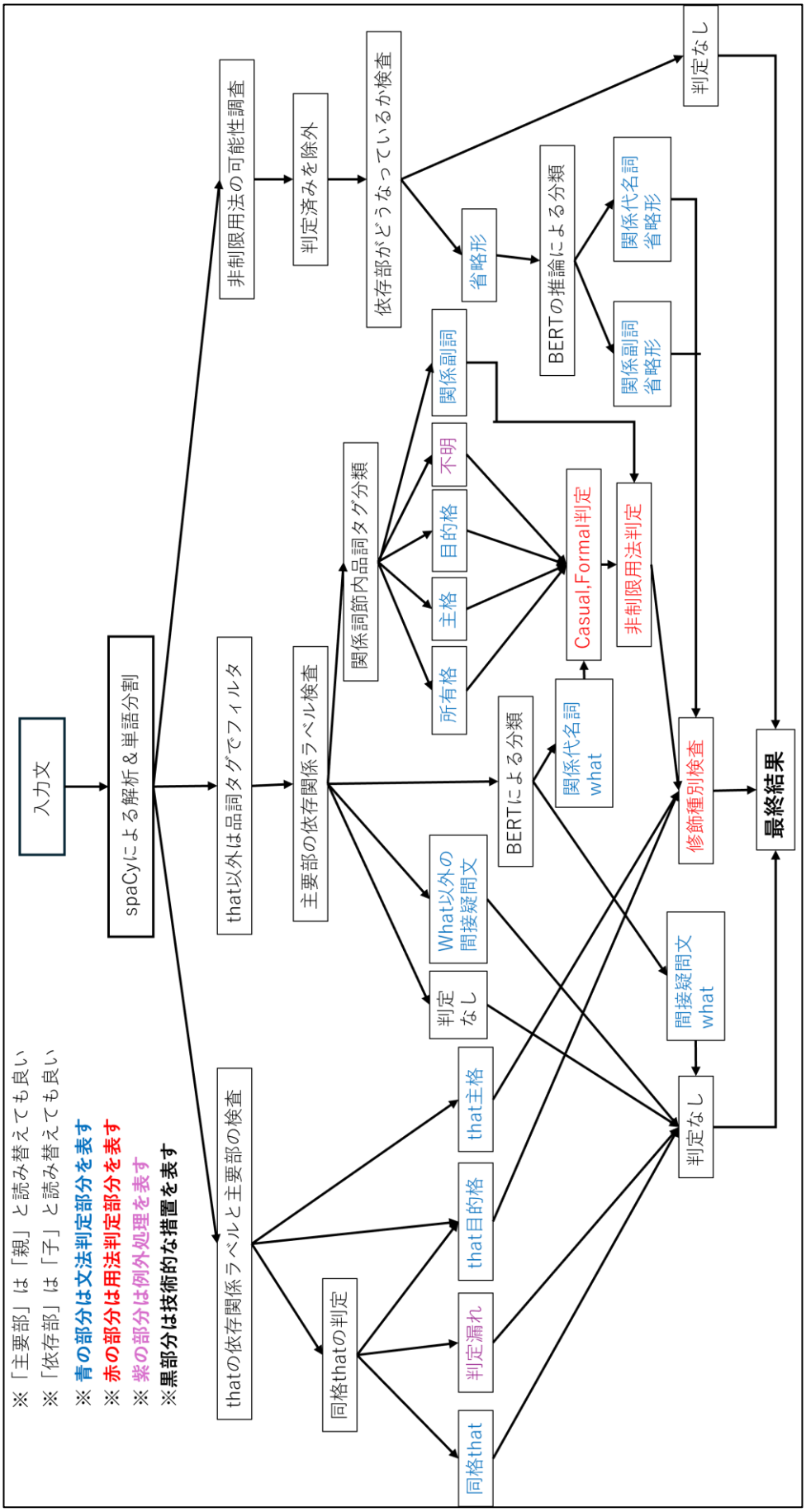

図 1　本判定器の処理の流れ

3. 関係詞判定器の評価

本判定器の性能評価は，主要な文法事項および識別が困難な類似構文を含む計 158 文を入力し，その出力結果を人手によって精査することで実施した．評価用データセットには，データの客観性と多様性を確保するため，特定の教科書に偏らないよう，複数の市販の英語参考書および英語学習用 Web サイトからランダムに抽出した例文を使用した．評価対象とした文法項目およびその内訳を表 2 に示す．なお，評価対象の文法以外にも同時に判定される項目があるが，今回はその部分の正誤も確認している．

表 2　評価する文法事項とその文の数

No.	文法項目	用意した文の数 [個]
1	関係代名詞(主格)	10
2	関係代名詞(目的格)	10
3	関係代名詞(所有格)	10
4	関係代名詞 名詞修飾	10
5	関係代名詞 代名詞修飾	2
6	関係代名詞 文修飾	3
7	Formal 関係代名詞	10
8	Casual 関係代名詞	10
9	関係代名詞 非制限用法	10
10	関係代名詞(省略)	8
11	関係副詞	19
12	関係副詞 非制限用法	2
13	関係副詞(省略)	10
14	関係代名詞 what	10
15	ダミー(同格 that)	6
16	ダミー(間接疑問文)	23
17	ダミー(強調構文)	5

表 2 のすべての例文を判定器に入力し，人力で精査したところ，全 158 文において誤判定は確認されず，誤判定率は 0%となった．この結果から，本判定器は中学・高校の検定教科書や一般的な学習参考書で扱われる標準的な英文に対して，極めて高い判定精度を有することが示された．

以上の評価により，本判定器はリーダビリティ算出の基盤として十分な信頼性を有していることが確認された．

4. リーダビリティの算出

関係詞判定器を利用した，構造的複雑さを考慮したリーダビリティ算出手法について示す．

本研究においては，OFYL を参考にした指標[3]と，DLT 理論[4]および 98%ロジック[5]を統合し，本研究のリーダビリティ指標を策定した．本指標は，大きく以下の 2 つの要素を合わせたものとして算出される．

1. 単語難易度指標（Modified OFYL）：単語レベルや文章長に基づく基礎的な難易度．
2. 文構造難易度指標 ： DLT 理論および 98%ロジックに基づく構造的な難易度．

なお，今後の指標は全て，文章(複数の文を含んだもの)に対するものである．

4.1 単語難易度指標の算出($Diff_{ofyl}$)

単語難易度の算出にあたっては，Ozasa et al.[3]による OFYL 指標を修正したモデルを提案する．原著の算出式には「熟語難易度（IdiomDiff)」が含まれているが，本システムにおいては熟語の定量的評価が困難であるため，その代替として，文章全体の語数が読解負荷に与える影響を考慮し，総単語数の対数値を新たな項として追加した．また，原著の算出式内の単語難易度については，CEFR-J[2]の単語難易度を使用した．$\log_{10} Words$等の変更の統計的な妥当性については今後のさらなる検証課題とするが，本研究においては暫定的な指標としてこの定義を採用する．改変後の算出式（式 1）を以下に示す．

$$Diff_{ofyl} = 0.0995 \times \frac{Words}{S} + 0.4302 \times \frac{Syllab}{W} + 0.9799 \times \frac{WordDiff}{W} + 0.15 \times \log_{10} Words + 0.2815 \quad (1)$$

ここで，各変数の定義は以下の通りである．

・Words/S：1 文あたりの平均単語数．
・Syllab/W：1 単語あたりの平均音節数．
・WordDiff/W：1 文あたりの平均単語難易度．
・$log10$(Words)：入力した文章(複数文を含む)の総単語数の常用対数．

ここで，$log10$(Words)の項にかかる係数 0.15 は，既存の公開ツールである『CEFR-J Level Calculator』の難易度判定との整合性を図るために導入した補正値である．予備実験において，本算出式の出力値と既存ツールの判定結果を比較検証し，両者の難易度レベルが近似するように経験則的に決定した．CEFR-J との違いは OFYL の式を用いることで，難易度の理由を明示できる点にある．

4.2 文構造難易度指標の算出($Diff_{grammar}$)

文構造難易度指標は，学習者の語彙理解度を保証する「98%ロジック」に基づく基準レベルと，文構造の複雑さに起因する「DLT コスト」，および関係詞の「学年別難易度」を統合して算出する．

(1) 98%語彙レベルの特定(L_{98})

まず，対象となる文章(複数の文を含む)に含まれる全単語の難易度（CEFR-J レベル）を昇順に整列させる．その上で,文章全体の 98%をカバーするために必要な最小の CEFR-J レベルを特定し，これを基礎となる語彙レベル L_{98}と定義する．この値は，Hu and Nation[5]が提唱した「辞書なしで理解可能な水準」となる学年を意味する．なお，各単語のレベル数値化は以下の表 3 に準ずる．

表 3　CEFR-J レベルの数値化

CEFR-J レベル	本システムでの数値	備考
A1	1～2	小学校～中学校 1 年相当
A2	2～3	中学 2 年～高校 1 年相当
B1	3～4	高校 2 年～大学受験相当
B2	4～5	大学受験～大学教養相当
C1 及び C2	5 以上	ネイティブ相当

(2) 関係詞の学年別難易度($L_{relative}$)

次に，文章中に含まれる関係詞の種類に応じた難易度重みを設定する.中学校学習指導要領[1]および高等学校学習指導要領[6]に基づき,各関係詞用法が導入される学年レベル($L_{relative}$)を定義した.中学生レベルの関係詞が含まれる文には 2.5，高校生レベルの関係詞が含まれる文には 3.5 の重み係数を付与する．各用法の分類を表 4 に示す.

表 4　関係詞の学年分類

中学生レベルの関係詞(2.5)	高校生レベルの関係詞(3.5)
関係代名詞(主格)	関係代名詞(所有格)
関係代名詞(目的格)	Formal 関係代名詞(主格)
Casual 関係代名詞(主格)	Formal 関係代名詞(目的格)
Casual 関係代名詞(目的格)	関係副詞
関係代名詞(省略形)	関係副詞(省略)
	関係代名詞 what
	Formal 関係代名詞 what
	非制限用法を含む文

(3) DLT コストの算出(DLT_{mem}, DLT_{int})

本モジュールにおいて，DLT 理論[4]の 2 つの指標である記憶コストと統合コストは，以下のように定量化した．なお，算出された記憶コストをDLT_{mem}，統合コストをDLT_{int}と定義する．

・記憶コスト（DLT_{mem}）：
関係詞節の入れ子構造の深さと定義する．

・統合コスト（DLT_{int}）:
先行詞と，文構造的に統合される動詞との間の「語数」と定義する．具体的には，先行詞が主節の主語である場合は「主節の動詞」まで，それ以外の場合は「関係節内の動詞」までを計測区間とし，その区間に存在する談話指示対象（名詞・動詞等）の総数をコストとして算出した．

(4) 文構造的難易度指標($Diff_{grammar}$)

最終的に，上記の(1)～(3)の値を基に，文構造難易度（$Diff_{grammar}$）を以下の式で定義する．ここで，文法的な難易度（$L_{relative}$）が語彙レベル（L_{98}）を上回る場合には,その乖離を負荷として加算する補正を行っている．

なお，式中の各重み係数（0.4，0.8，0.08 等）は，算出されるリーダビリティ値を適切な範囲に正規化するために設定した値である．具体的には，難易度が既知である関係詞の英文（A2 相当の平易な文および C レベル以上の難解な文）をサンプルとして入力し，その出力値が本指標の基準範囲（1.0～6.0）に収まるよう，予備的な検証を経て係数の調整を行った．

・$L_{relative} > L_{98}$の場合

$$Diff_{grammar} = L_{98} + 0.4 \times (L_{relative} - L_{98}) + 0.8 \times DLT_{mem} + 0.08 \times DLT_{int} + 0.08 \times (DLT_{mem} + DLT_{int})$$

・その他の場合

$$Diff_{grammar} = L_{98} + 0.8 \times DLT_{mem} + 0.08 \times DLT_{int} + 0.08 \times (DLT_{mem} + DLT_{int})$$

2 つの式の違いは，$L_{relative} > L_{98}$ の場合に$0.4 \times (L_{relative} - L_{98})$の補正が加わっている点である．

5. リーダビリティの算出結果

本章では，定義した各指標を実際の英文に適用し，その算出結果の妥当性について論じる．

5.1 使用した例文について

リーダビリティ指標の評価用データとして，中学校・高等学校の英語検定教科書から抽出した本文，および令和 7 年度大学入学共通テストの英語（リーディング）問題文を使用した.教科書データには,中学校用『Sunshine English Course 1, 2, 3』（開隆堂出版）[7]，および高等学校用『CROWN English Communication I, II, III』（三省堂）[8]を用い，抽出したデータセットの総数は 234 文章である．なお、教科書であれば 1 レッスンを 1 文章とし,共通テストであれば 1 つの大問を 1 文章としている．

また，データ抽出に際しては，関係詞を含まない文も対象とした．本モジュールは関係詞の有無や構造のみを指標の算出根拠としているため，その他の文法事項については本研究の評価対象外とし，今後の課題とする．

5.2 リーダビリティの分布

前節の各本文は出現順に並べられており，これらの本文の難易度は徐々に上昇するはずである．したがって，本研究で作成した指標の各本文に対する算出値が本文番号と強い正の相関関係を持つならば，その指標は日本の英語教育の実態に即した妥当なものであると評価できる．本節では，各指標と本文番号との相関関係を分析する．

(1) $Diff_{ofyl}$のみを使った算出結果

まず，4.1 節にて述べた単語難易度ベースの指標，$Diff_{ofyl}$を使用して分析を行った．教科書および共通テストの各文の算出結果の分布を図 2 に示す．

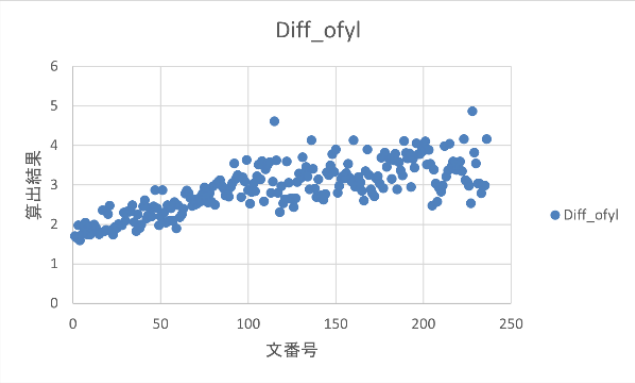

図 2　$Diff_{ofyl}$の分布

グラフ上のプロットデータに基づき相関係数を求めた結果，以下の通りとなった．

相関係数($Diff_{ofyl}$) ： 0.764443

(2) $Diff_{grammar}$のみを使った算出結果

次に，4.2 節にて定義した文構造難易度指標 $Diff_{grammar}$ を使用して分析を行った．同様にグラフ化したものを図 3 に示す．

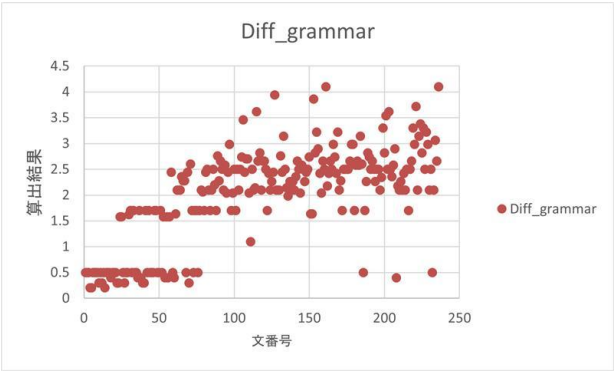

図 3　$Diff_{grammar}$の分布

グラフ上のプロットデータに基づき相関係数を求めた結果，以下の通りとなった．

相関係数($Diff_{grammar}$) ： 0.704855

(3) $Diff_{ofyl}$と$Diff_{grammar}$双方を使った算出結果

$Diff_{ofyl}$ は語彙的な難易度を高精度に反映するが，文構造の複雑さを原理的に考慮できない．一方，$Diff_{grammar}$は文構造の負荷を反映するが，単語の難易度を考慮できない．そこで本研究では，これら双方の特性を相互補完するハイブリッド型の指標$Diff_{select}$を作成した．

本指標では，ある文の読解難易度は「語彙」または「文構造」のうち，より難しい要素によって決定づけられると仮定し，以下の選択式を採用する．

$$Diff_{select} = max(Diff_{ofyl}, Diff_{grammar})$$

すなわち，計算された $Diff_{ofyl}$ と $Diff_{grammar}$ のうち，より数値が高い方を最終的なリーダビリティとして採用する．この指標を用いた算出結果の分布を図 4 に示す．

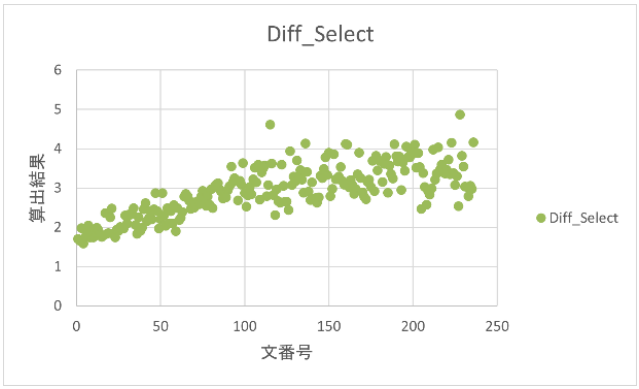

図 4　$Diff_{select}$の分布

グラフ上のプロットデータに基づき相関係数を求めた結果，以下の通りとなった．

相関係数($Diff_{select}$) : 0.768539

(4) 結果のまとめ

各指標の結果を比較すると，ハイブリッド型指標である$Diff_{select}$が最も高い相関係数を示した．このことは，単語の難易度だけでなく，関係詞による文法的な複雑さを考慮に入れることで，より教科書の学年レベル（人間の感覚）に近いリーダビリティ判定が可能になることを示唆している．

6. 結言

本研究では，英語教材の客観的な難易度評価に向け，関係詞の構文的複雑さを考慮した新たなリーダビリティ指標を提案した．spaCy と BERT を統合した判定器は，関係詞の省略や用法を 100%の精度で分類可能であることを確認した．また，提案手法（$Diff_{select}$）を用いた算出結果は，教科書の学年進行に対して相関係数 0.769 を示し，従来の単

語難易度のみの指標（0.764）を上回る精度を達成した．これは，リーダビリティ評価において単語だけでなく文構造の考慮が有効であることを示唆している．しかしながら，暫定的に設定した係数や，$\log_{10} Words$等の変更の妥当性についての検証を行う必要がある．また，リーダビリティ判定の精度向上のため，関係詞以外の文法も考慮しなければならないと考える．

参考文献

[1] 文部科学省. (2017). 『中学校学習指導要領（平成 29 年告示）解説 外国語編』. 文部科学省.
[2] CEFR-J Project. (n.d.). *CEFR-J Wordlist Version 1.5*. Retrieved February 4, 2026, from 〈https://www.cefr-j.org/download.html#cefrj〉
[3] Ozasa, T., Weir, G. R. S., & Fukui, M. (2007). Measuring Readability for Japanese Learners of English.
[4] Gibson, E. (2000). The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity.
[5] Hu, M., & Nation, P. (2000). Unknown Vocabulary Density and Reading Comprehension. *Reading in a Foreign Language*, 13(1), 403-430.
[6] 文部科学省. (2018). 『高等学校学習指導要領（平成 30 年告示）解説 外国語編 英語編』. 文部科学省.
[7] 開隆堂出版.『Sunshine English Course 1, 2, 3』. 開隆堂出版.
[8] 三省堂. 『CROWN English Communication I, II, III』. 三省堂.