Vol.2019-CE-148 No.12
2019/2/17

文の節構造に着目したレポート文における推敲支援方法の検討

大野博之[†1]　稲積宏誠[†2]

概要：学生のレポートには，わかりづらい文も散見されるが，それを書いている筆者自身は意図を理解しているためその問題に気づきにくい．もし語彙や語種の使用状態の全体像や，文の節構造が複雑な長文などを客観的に概観できれば，学生自身の推敲作業の支援が可能となるだけでなく，指導者としても指導のポイントを見つけやすく，添削のコストを低減できる．そこで ICT を活用し，全体像が把握できる環境を用意すると共に，個々の文に対しても，節構造等に着目して，どのような修正が考えられるかを検討する．

Study on elaboration support method focused on structure of clauses

HIROYUKI OONO[†1]　HIROSHIGE INAZUMI[†2]

1. はじめに

客観的・論理的な文章理解・作成能力の育成は，専門領域を問わず，近年の大学における必須の課題であり，基礎教養教育の一部としてアカデミックライティングの授業が展開されてきている[1][2]．しかし，学習者の語彙力や文章構成力が均一ではないことや教員による添削を中心とした指導が必要とされることから，適切な教材の選定や効率的な作文指導方法の改善が必要である．

学生のレポート文を添削・指導するうえでは，文章全体の構成や流れを指導する面と，文の作成そのものを指導する面がある．前者については，テーマの発見から章・節の構成，パラグラフライティングなど多くの書籍で解説されている内容を講義形式で指導することができる．一方，後者においては校正の面と推敲の面があり，一定の形式的な指導はできるものの，実際に書かれた個々の文に対して問題がないかどうかをチェックするのは，指導者にとって負担が大きい．

こうした問題に対して我々は，機械的な校正チェックが可能な支援ツールを構築してきた[3]．しかし，文を校正しただけでは，読みづらい・理解しづらい文の問題をすべて解消することはできない．何が原因となっていて，どのように修正すべきかまでは，このツールでは指摘できていない．もし，学生の書く文の特徴を機械的にとらえることができれば，校正面の形式チェックとは異なる「推敲」の指導支援につなげることができるようになる．

推敲の関連研究として，係り受けの複雑さに着目した取り組みがある[4]．これは，「係り受けの複雑さ」を人間の短期記憶を模した係り受け解析過程モデルによって定義し，修正すべき文の指摘と修正候補の生成を行うものである．修正候補として，修飾節の入れ替え案と長い文の分割案を提示している．個々の文を対象としているため，具体的な推敲に結び付けられる試みであるが，文章全体の概観までは行えない．

また，日本語を対象とした読みやすさ・リーダビリティに関する研究としては，次の 3 つが挙げられる．柴崎・原の取り組み[5]では，文の平仮名の割合，平均述語数，平均文字数，平均文節数を用いて，9 学年もしくは 12 学年のいずれに該当する文かを予測するリーダビリティ公式を作成している．また，佐藤の取り組み[6]では，独自に作成した教科書コーパスを基に bigram による言語モデルを作成し，13 段階の学年区分を決定している．李の取り組み[7]では，平均文長，漢語率，和語率，動詞率，助詞率を用いて，重回帰分析による公式を定め，難易度として 6 段階（初級前半・初級後半・中級前半・中級後半・上級前半・上級後半）を結果としている．これらの読みやすさ関連の研究に共通するのは，対象文全体のリーダビリティを検討している点であり，個々の文章を評価するものではない点である．そのため，文章全体のおおまかな傾向と修正を促すことは可能であるが，個々の文に対する指摘には結びつかない．

そこで本研究では，レポート文などの全体を概観するだけでなく，個々の文を対象にすることも踏まえ，自然言語処理技術によって文の特徴量を算出し，わかりやすい文への修正に向けた支援に使用できるかどうかを検討したい．ただし，同じ文でも，それを読む人の語彙量や背景知識，理解力などには個人差があるため，「わかりやすい文」の絶対的な定義は断定できない．そのため文章のタイプとして，初等・中等教育で使用される教科書，白書，新聞の 3 つを取り上げ，学生のレポート文がいずれのタイプの傾向を持っているか，そしてどのような特徴があるかを調査する．なお，10 文節未満の短い文や 25 文節を超えるような長い文は，ここでは議論の対象外とする．

以上のことから，次の流れで検討を行う．まず文の特徴量として，文字種・語種に関するものと文の構造（係り受

†1 東京医療保健大学
Tokyo Healthcare University
†2 青山学院大学
Aoyama Gakuin University

ⓒ 2019 Information Processing Society of Japan

け構造，節構造など）に関するものを定める．そして，これらの特徴を概観するツールを試作する．また，これらの特徴量を使い，学生のレポート文と既存の教科書・白書・新聞の文と比較し，どのような推敲支援が可能か検討する．

2. 使用する文の特徴量の選定

文を特徴づけるものが何かは定かではないため，まずは文を構成する要素や係り受け関係による構造をもとに，文の特徴量として考えられるものをいくつか選定した．文字種・語種に関する特徴量，係り受け構造に関する特徴量，節構造に関する特徴量，チャンクに関する特徴量，文の並びに依存しない特徴量の 5 つに分けてそれぞれ説明する．なお，各特徴量を算出するにあたり，「教科書コーパス語彙表および BCCWJ 主要コーパス語彙表[8]」と「新阪本教育基本語彙[9]」の 3 つを基準データとして用いた．また，自然言語処理ツールとして，ipadic を辞書とする MeCab [10] および CaboCha [11]を使用する．実際の算出には，これらを組み合わせて各数値を出力する専用のツールを作成し用いた．

2.1 文字種・語種に関する特徴量

文字種・語種に関する特徴量としては，表 1 に示すように 12 項目とした．なお，文字数や形態素数のように他の要素で代替できるものや，受動態表現・使役表現など出現頻度が低いものは除外している．

表 1　文字種・語種に関する特徴量

項目	説明
文節数	CaboCha で判定された文節に基づく値
平仮名率 片仮名率 漢字率	文を構成する文字種の割合
形容詞率 形容動詞率 副詞率 助詞率	MeCab で判定された品詞に基づいた，文を構成する形態素の各品詞の割合
和語率 漢語率 外語率	「教科書コーパス語彙表」・「BCCWJ 主要コーパス語彙表」・「新阪本教育基本語彙」に登録されている語種に基づき，判定した各語種の割合
動詞・サ変文節率	CaboCha で判定された文節に基づき，動詞やサ変名詞によって動詞のように扱われている文節の割合

2.2 係り受け構造に関する特徴量

係り受けに構造に関する特徴量は，表 2 に示す通り 4 項目とする．「係り受け距離」とは，係り受け先が何文節先かを表した数値を意味し，その文の文節数により正規化した値である．また「修飾語の順番評価値」は，次のように重みと変量を定め，図 1 で示すように加重平均により算出したものである．なお，2 つ以上の係り元がある文節を対象とし，係り元が 1 文節の場合はそれを除外して考える．また，複数個所ある場合は，その平均としている．

- 最終文節に係る修飾節数を n
- 重み w_i は最終文節に係る修飾節の文節数
- 変量 x_i は最初の修飾節を 1，最後の修飾節を 0.01 とし，その間は図 1 の式で算出

この値は，最大値 1，最小値 0.01 となり，長い修飾語が文の前半に来るほど 0.505 より値が大きくなり，後半に来るほど 0.505 より小さくなる．例文でこれらの評価値例を図 2 および表 3 にて示す(小数点以下第 3 位で四捨五入)．

表 2　係り受けに構造に関する特徴量

項目	説明
平均係り受け距離 （正規化）	文中での係り受け距離の平均値
係り受け距離標準偏差 （正規化）	係り受け距離の標準偏差
最終文節への入射集中率	係り受け数を分母とした時の最後の文節に係る入射数の割合
修飾語の順番評価値	長い修飾語が先に，短い修飾語が後にくることの評価値

$$x_i = \frac{1-0.01}{n-1}\cdot(n-i)+0.01$$

$$\text{修飾語の順番評価値} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

図 1　修飾語の順番評価値

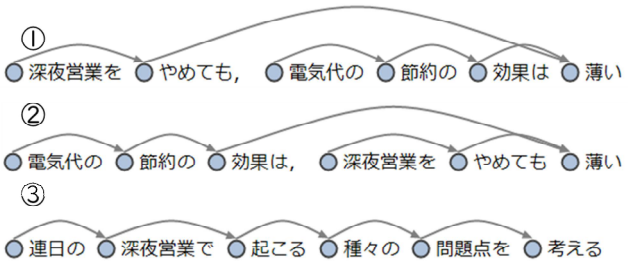

図 2　係り受けの文例

表 3　係り受け構造に関する特徴量例

項目	①	②	③
平均係り受け距離 （正規化）	0.32	0.28	0.2
係り受け距離標準偏差 （正規化）	0.24	0.16	0
最終文節への入射集中率	0.4	0.4	0.2
修飾語の順番評価値	0.41	0.60	0.51

ⓒ 2019 Information Processing Society of Japan

2.3 節構造に関する特徴量

節構造に関する特徴量は，補足節・連体節・副詞節それぞれの節を構成する文節の占有率を表す「補足節占有文節率，連体節占有文節率，副詞節占有文節率」の 3 項目とする．これに伴い，補足節・連体節・副詞節を判定する必要があるため，「基礎日本語文法・改訂版[12]」に基づいて機械的な節構造の判定ルールを算出ツールに実装した．

2.4 チャンクに関する特徴量

特徴量の 1 つとして，チャンク圧縮率を用いる．これは，チャンキングしない場合と比べて，どれだけチャンク数が圧縮されたかを示す値である．つまり，元よりチャンク数が減少すればするほど，圧縮率は大きな値となる．

チャンクとは，心理学者ミラーの提唱した概念であり，人間が情報を知覚する際の「情報のまとまり」をさす．そして，複数のチャンクをグループ化し，より大きな 1 つのチャンクにまとめることをチャンキングという．
人が言語理解過程において，形態素解析や構文解析などを脳でおこなう際に，これらの処理結果を一時的・短期的な記憶装置に保持する[13]．これをワーキングメモリやスタックと呼び，通常 7±2 チャンクの容量を持っているといわれている．

これらの考えに基づき，補足節・連体節・副詞節を構成する文節を除いて，1 文節 1 チャンクを基本に計測する．補足節・連体節は，出現分だけ 1 チャンクを加算し，副詞節は 2 チャンクを加算する．なお，「助詞-連語」「助詞-連体化」「連体詞」で接続される文節も，合わせて 1 チャンクとする．

図 3 の例では，①は 10 分節あるが，「橋の上に」はチャンキングされたものとするため，9 チャンクとなり，圧縮率は 0.1 となる．②では同じ 10 文節であるが，連体節が入れ子状態で 3 つあり，連体節に関係のない文節は 3 文節あるため，6 チャンクとなり，圧縮率は 0.4 となる．

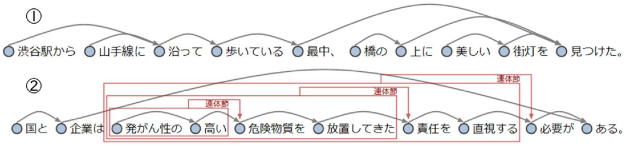

図 3　チャンクの文例

2.5 文の並びに依存しない特徴量

最終文節を木構造のルートとして捉えた時の葉の深さの平均値や標準偏差を特徴量として定めた．これらは係り受けとは異なり，文節の並び順に依存しない評価を行うための尺度である．ただし，係り受け距離を正規化したのと同様に正規化したものを用いる．また，補足節・連体節などでチャンキングされた場合における葉の深さの平均値や標準偏差も同様に特徴量として定めることとした．深さが深いほど，1 つの文節を長く修飾して説明するような文となり，深さが浅いと 1 つの文節を複数の修飾節で説明するような文になる．

また，文内のどれだけの文節が複数個所から修飾されているのかを示す評価値として「分岐率」を定めた．

例えば，「森林は，環境保全に資する多様な機能を持つ」という文であれば，図 4 のように木構造で表され，6 文節中，2 個以上に分岐する文節は 2 個所あるため，分岐率は 2÷6 で求められ表 4 に示したようになる．また，「環境保全に資する多様な機能」は連体節となるため，チャンキングすると「森林は，機能を持つ」という形になるため，表 4 に示したような値になる．

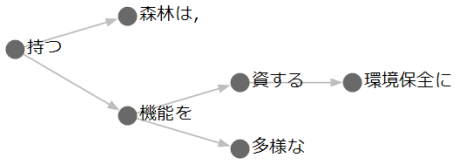

図 4　木構造化した様子

表 4　文の並びに依存しない特徴量例

項目	値
葉の深さの平均（正規化）	0.4
葉の深さの標準偏差（正規化）	0.16
葉(補足節・連体節)の深さの平均（正規化）	0.2
葉(補足節・連体節)の深さの標準偏差（正規化）	0
分岐率	0.33

3. 語彙の難易度による評価対象文の選定

学生のレポート文とコーパスの教科書，白書，新聞データを比較するにあたり，使用語彙の難易度を揃えることとした．しかし，語彙の難易度は，日本語を母語とする者かどうかで基準が変化する．例えば，日本語を学習する留学生であれば，日本語の検定試験や留学生向けの日本語教科書が基準となり，これを日本語を母語とする者の基準とするには対象の語彙量が少ない．そこで，本研究の対象は「日本語を母語とする学習者」であることから，中等教育以下で使用されている語彙とした．以下，語彙の難易度の決定方法および，その難易度に基づいてコーパスからの評価対象文の選定について説明する．

3.1 語彙の難易度

語彙の難易度を定義するために表 5 に示すような「BCCWJ 教科書コーパス語彙表」・「BCCWJ 主要コーパス語彙表」と「新阪本教育基本語彙」の 3 つの基準データを用いる．語彙の難易度としては，小学校低学年をレベル 1，小学校高学年をレベル 2，中学校をレベル 3，高校をレベル

4 とした．また，学校教育で使用されていないが，書籍・雑誌・新聞等では一般的に使用されているものとして「BCCWJ 主要コーパス語彙表」の語彙をレベル 5 と設定した．なお，固有名詞については「教科書コーパス語彙表」の固有名詞における使用度数の上位 100 件については，固有名詞でもよく使用されるものとして，そのままレベルを反映させることとした．それ以外については対象外として難易度を付与しない．これらにより，中等教育で用いられる語彙は，レベル 3 およびレベル 4 相当の語彙ということになる．

実際に語彙のレベルを決定するには，形態素解析の結果から基準データとのマッチングを行う．ただし，これら 3 種の基準データは，形態素解析ツール MeCab を UniDic 辞書と共に用いて短単位で作成されている．一方，本研究では ipadic 辞書で動作する係り受け解析ツール CaboCha を利用する都合上，MeCab も ipadic 辞書を使用する．これにより，形態素解析の基本形表記や品詞表記が異なる場合が起こり得る．そこで，利用する 3 つの基準データに対して，一部人手で修正を行っている．最終的には，これに加え，固有名詞用の基準データと，ユーザーが任意に語彙のレベルを登録できるものを準備した．図 5 は，語彙の難易度を付与した例である．

表 5　語彙の難易度を決定する基準データ

基準データ	説明
教科書コーパス語彙表	BCCWJ[14]に収録されている「教科書コーパス」の語彙の一覧． 2005 年度に使用された小学校・中学校・高等学校の全学年・全教科の教科書 1 種ずつを対象として，それらに出現する約 50000 語の初出学年が，小学校低学年・小学校高学年・中学校・高校の 4 段階で示されている．ここから，記号，固有名詞，数詞を除外した約 39000 語を利用．
主要コーパス語彙表	BCCWJ の 2010 年 12 月 9 日版（非公開）の図書館書籍，出版物書籍，雑誌，新聞，Yahoo!知恵袋，Yahoo!ブログの 6 種を調査対象として得られた語彙の一覧である．約 130000 語が収録されているが，教科書コーパスにおける語彙と重複するものも存在する．ここから，記号，固有名詞，数詞を除外し，Yahoo!知恵袋・Yahoo!ブログの両者にしか出現しない語も除外した．さらに教科書コーパス語彙表との重複分も省いた約 44000 語を利用．
新阪本教育基本語彙	義務教育 9 年間のうちに，どのような範囲・順序で単語を学習させるのが良いかを人手によって定めたものである．約 27000 語が小学校低学年・小学校高学年・中学校の 3 段階で示されており，さらに各段階内での優先度も 1～4 で示されている．ここから格助詞，計助詞，係助詞，間投助詞，助動詞，終助詞，接続助詞，副助詞，連語を除外した約 20000 語を利用．

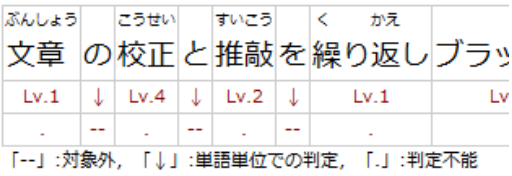

図 5　語彙の難易度付与例

3.2 教科書，白書，新聞，学生のレポート文の選定

学生のレポート文は，2008 年度から 2015 年度までの文章表現法関連の授業で提出された文章を用いる．教科書，白書，新聞データは，「現代日本語書き言葉均衡コーパス（BCCWJ）」の教科書コーパス（OT）と白書コーパス（OW）および新聞コーパス（PN）を用いる．なお，これらのデータから丸括弧を含む文と鍵括弧を含む文は除外してある．また，文として不要な文頭の記号類を除去するなど一定の整形を行っている．

このように得られた文に対して，語彙の難易度を揃えるために中等教育相当以下の文（初等教育のみの語彙で作られた文は除く）を抽出し，さらに文の長さを大まかに 10～14 文節，15～19 文節，20～24 文節の 3 グループに分け，表 6 に示したデータ数が得られた．これを今回の評価対象とする．

表 6　選定した各文のデータ数

グループ	学生	教科書	白書	新聞
10～14 文節	3715 文	4523 文	10477 文	2552 文
15～19 文節	1451 文	1575 文	7692 文	663 文
20～24 文節	464 文	387 文	4341 文	140 文

4. 学生のレポート文の評価

4.1 文の構造の傾向

学生のレポート文が，文の構造上，どういったタイプの文に近いかを判定するために，3 つの文長のグループごとに mySVM[15]による学習および 2 値分類を実施した．これは，分類すること自体を主目的とするのではなく，それぞれの特徴量の mySVM から得られる重みベクトルによって，文のタイプに影響を与える特徴量が何かを調べるためである．

具体的には，教科書と学生のデータ（以降，教 vs 学），白書と学生のデータ（以降，白 vs 学），新聞と学生のデータ（以降，新 vs 学）の 3 パターンにおいて，3 つの文長のグループごとに行い，合計 9 パターンの学習を行った．

1. それぞれの素材文数の少ない方の半分を基準に，ランダムに文を選択
2. mySVM による学習
3. それぞれの特徴量の重みベクトルを取得

ⓒ 2019 Information Processing Society of Japan

4. 上記の作業を 100 回繰り返し
5. 100 回のうち 7 割以上ベクトルの方向が一致している方を取得
6. 100 回の重み平均を取得

例えば 10〜14 文節の「新 vs 学」であれば，新聞の方が少ない素材文数なので，1276 文を新聞と学生側からランダムに取得し学習を行って，重みベクトルを取得する．これを 100 回繰り返し，その特徴量が新聞寄りのものか，学生寄りのものかを得る．ただし，重みベクトルの方向が一致していても，重み自体が小さい場合は影響を与えていないと思われるため，その点は考慮する．

これらの処理の実験結果を表 7 に示す．各特徴量の値が大きい時に二者間において，どちら寄りに近づくかを表したものである．「×」は重みベクトルが 7 割未満であり，どちら寄りとも判断できなかったものである．また，背景色付きの欄は，重みベクトルが-1〜1 以内のもので重み自体が小さいものである．

なお，訓練データの正解率は，全体として 60%〜70%程度であり，決して高くはない．これは，学生のレポート文すべてが，教科書・白書・新聞と明確な違いがあるというわけではないためと思われる．mySVM にて作成された学習モデルで学生のレポート文を分類し，確信度が 0.9 以上のものに限定した時の正解率を表 8 に示す．

重みの傾向からもわかるように 10〜14 文節の場合，学生のレポート文と教科書・新聞においては，今回挙げた特徴量における明確な違いは見られず，分類もうまくいかない．文節が多くなる，すなわち長文になるにつれて，補足節・連体節が用いられている時のその節の長さや，係り受け距離のばらつきに違いが表れるようである．また，文の長さに関係なく，補足節を多用する傾向が見受けられる．

表 7　文の構造に関する特徴量の重みの傾向

	10-14文節			15-19文節			20-24文節		
	教vs学	白vs学	新vs学	教vs学	白vs学	新vs学	教vs学	白vs学	新vs学
平均係り受け距離（正規化）	学	学	×	学	学	新	×	学	×
係り受け距離標準偏差（正規化）	教	白	学	教	白	学	教	白	×
最終文節への入射集中率	教	×	新	教	学	新	×	学	×
修飾語の順番評価値	教	白	新	×	白	新	学	×	×
補足節占有文節率	学	学	学	学	学	学	学	学	学
連体節占有文節率	教	学	新	×	学	新	×	学	新
副詞節占有文節率	学	学	学	学	学	学	×	学	学
チャンク圧縮率	教	白	新	教	白	新	教	白	新
葉の深さの平均（正規化）	学	×	学	×	×	学	×	×	学
葉の深さの標準偏差（正規化）	学	白	×	学	×	×	学	×	新
葉(補足節・連体節)の深さの平均（正規化）	教	白	新	教	白	新	教	白	新
葉(補足節・連体節)の深さの標準偏差（正規化）	×	学	学	学	白	学	学	×	学
分岐点率	学	学	学	×	学	学	学	学	学

表 8　確信度 0.9 以上における学生の文の分類正解率

グループ	vs 教科書	vs 白書	vs 新聞
10〜14 文節	50%	73%	49%
15〜19 文節	67%	80%	70%
20〜24 文節	74%	80%	81%

4.2 学生のレポート文における典型例

まず，補足節を多用している実例を図 6 に示す．これは「高齢者が増加することで社会保険や年金などの需要が高まるのに対し，少子化が進むことで一人一人の税負担が増すことになるのである」という文である（前述のモデルによって，教科書・白書・新聞のいずれと比較しても，学生の文であると確信度の高い判定がされる）．この文は，「高齢者の増加によって社会保険や年金などの需要が高まる一方，少子化による影響で，一人ひとりの税負担が増加する」のように補足節を用いず書き替えることも可能である（この文は同様に分類判定を行うといずれも学生の文ではないと確信度の高い判定がされる）．にもかかわらず，「こと」に代表される補足節を多用してしまうのは，それに置き換える語彙力が足らず，説明的な要素を補足節としてまとめて，文をつなぎ合わせてしまうことに原因があるのではないだろうか．また，これらの文を木構造として見ると，図 7 となる．多用している側の木構造は，途中要素を「こと」で修飾しているため，分岐率も高めになり，「葉(補足節・連体節)の深さの平均（正規化）」も小さくなる．つまり，チャンキングできる要素が多いため，少ない短期記憶でも把握しやすい．一方，未使用側の木構造は，文末の文節を修飾する形となっており，補足節だけでなく連体節も使用していないため，「増加する」を修飾する複数の情報を脳に長くとどめておく必要がある．どちらのタイプが良い・悪いというわけではないが，補足節が多過ぎるのは修正を指導したい点ではある．

図 6　補足節を多用している例

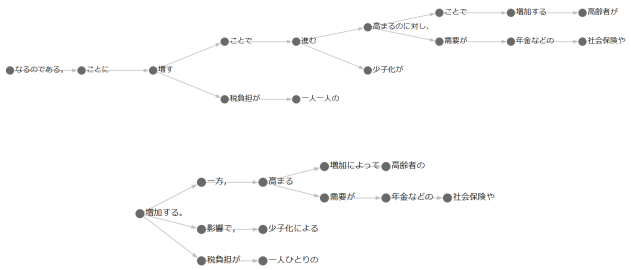

図 7　補足節を多用した文と未使用の文の木構造

次に，分岐点や係り受け距離，「葉(補足節・連体節)の深さの平均（正規化）」に関する実例として，図 8 に学生の文と判定されたもの，図 9 に学生の文とは判定されなかったものを示す．どちらの図も，元は学生の文である．図を見てわかるように，学生の文と判定されるものは，1 文節という短い情報で文節を修飾している傾向が多く見られる．

これにより，分岐点が多くなり，係り受け距離も結果的に長くなってしまう．これらに加えて，補足節や連体節も多く利用するため，それが「葉(補足節・連体節)の深さの平均（正規化）」に数値として現れている．

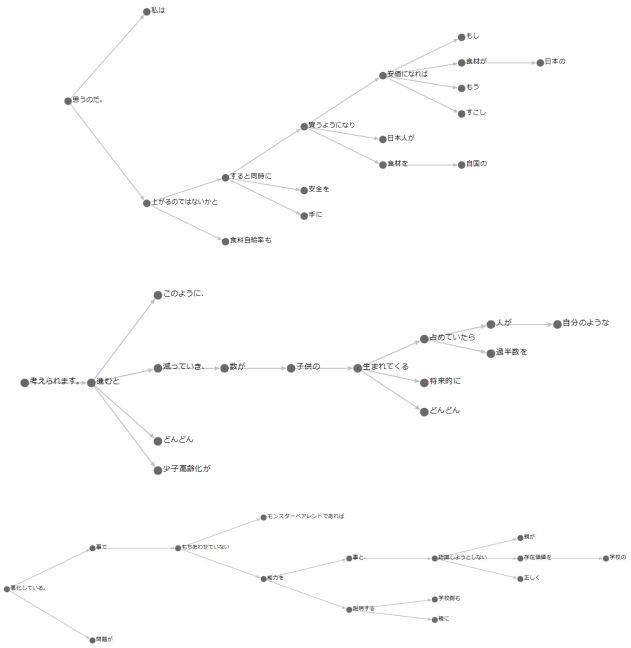

図 8　学生の文が学生のものだと判定された文例

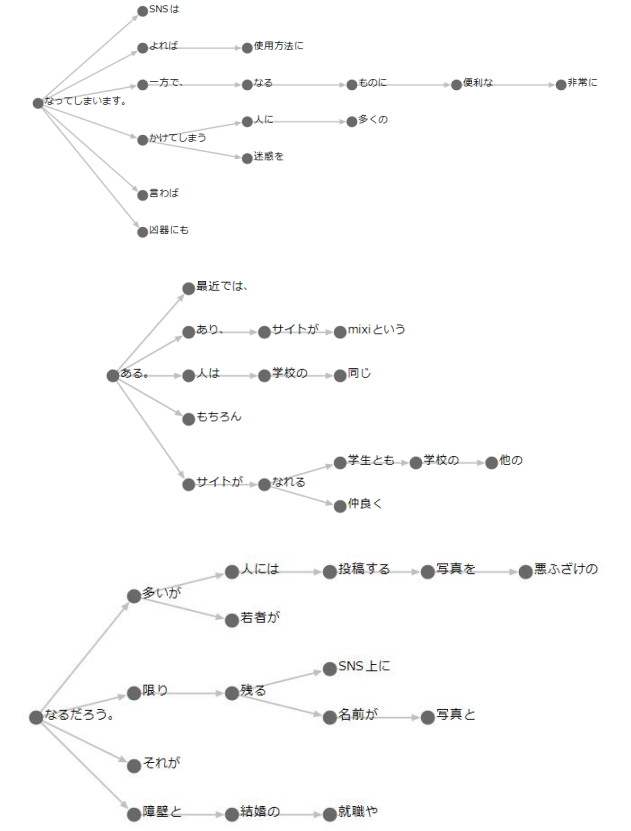

図 9　学生の文が学生のものでないと判定された文例

4.3 文字種・語種に関する特徴量を含めた比較

参考に，文の構造に関する特徴量に加え，文字種・語種の特徴量を含めた場合の文字種・語種の特徴量の重みの傾向を表 9 に示す．文字種・語種としては，教科書・白書・新聞と比べて，学生は平仮名と助詞を多用する傾向があるようである．実際に，例えば 10〜14 文節の新聞データと学生データを比較すると，新聞データの平仮名率の平均と助詞率の平均が約 44%と約 28%であるのに対し，学生データは約 54%と約 31%で多めになっている．なお，訓練データの正解率は，全体として 70%〜85%程度であり，文の構造に関する特徴量のみで実施したときよりも，良い判別結果となった．

表 9　文字種・語種の特徴量の重みの傾向

	10-14文節			15-19文節			20-24文節		
	教vs学	白vs学	新vs学	教vs学	白vs学	新vs学	教vs学	白vs学	新vs学
平仮名率	学	学	学	学	学	学	**学**	学	学
片仮名率	学	学	新	**教**	学	新	教	×	新
漢字率	学	白	新	学	白	新	学	白	新
形容詞率	教	×	新	学	×	×	教	×	学
形容動詞率	学	学	学	学	白	学	学	白	×
副詞率	学	白	×	学	白	×	学	白	学
助詞率	学	学	学	学	学	学	学	学	学
和語率	**学**	学	**学**	**学**	学	×	**学**	学	**学**
漢語率	**教**	**白**	学	**教**	**白**	学	教	白	**学**
外語率	学	**学**	学	学	×	学	学	×	学
動詞・サ変文節率	**教**	**学**	新	×	学	新	×	学	×

5. 推敲支援・指導支援ツール

文の推敲や指導支援を行うツールは，利用者が教員もしくは学生自身であることから，ある程度容易に扱える必要がある．そこで，図 10 のように，基本的には 3 つの手順「①文章を入力する，②チェックボタンを押す，③結果を確認する」で利用できるテキストエディタを試作した．このエディタは，本研究で述べた語彙レベルや語種および文章タイプなど，表 10 に示す情報を提示することで，利用者への情報提供を行うものである．

図 10　推敲支援・指導支援ツール画面

ⓒ 2019 Information Processing Society of Japan

表 10　ツールにおける対象文全体の概要表示項目

表示情報	説明
文章全体の評価	7 つの視点で，全体の評価を表示 「漢字のバランス，語彙の難しさ，文のやわらかさ，文の読みづらさ，文の分割候補，文の長さ，タイプ(教科書・白書・新聞)
漢字含有率	全体の漢字含有率を表示
語彙レベルの分布	全体の語彙レベル 1〜5 までの含有数/含有率を表示
語種の分布	和語・漢語・外来語などの含有数/含有率を表示

5.1 文章全体の評価項目

文章全体の評価としては，それぞれの数値を出しただけではわかりづらいため，以下のような一定のルールで試作段階では表示内容を定めている．ただし，ここで用いている判断基準の一部には，経験則によるものがあるため，今後精密化していく必要がある．

①漢字のバランス

全体の漢字の含有率によって，評価を 5 パターンで表示する．具体的には以下のとおりであり，これらは BCCWJ コーパスにおける新聞・教科書・白書の漢字含有率をもとに定めた．新聞コーパスの 5412 文において漢字含有率は 41.4%であり，教科書コーパスの 8736 文では 34.2%，白書コーパス 32544 文では 47.8%であった．

0〜10%未満：漢字が少なすぎる
10〜20%未満：漢字が少ない
20〜45%未満：ちょうど良い
45〜50%未満：漢字が多い
50%以上　　：漢字が多すぎる

②語彙の難しさ

語彙レベルの分布をもとに，評価を 5 パターンで表示する．具体的には，以下の通りで，上のレベルが優先されて表示される．

語彙レベル 5 以上が 10%以上：難しめ
語彙レベル 4 以上が 10%以上：少し難しめ
語彙レベル 3 以上が 10%以上：ふつう
語彙レベル 2 以上が 10%以上：やさしめ
その他　　：かなりやさしめ

③文のやわらかさ

語種の和語と漢語の分布をもとに，評価を 9 パターンで表示する．具体的には，和語と漢語の割合の差が，10 以下の場合は「ふつう」，20 以下の場合は「少し やわらかい / かたい」，30 以下の場合は「やわらかい / かたい」，40 以下の場合は「けっこう やわらかい / かたい」，それ以外は「かなり やわらかい / かたい」と表示される．

④文の読みづらさ

係り受け距離の標準偏差の平均によって，評価を 3 パターンで表示する．係り受け距離の標準偏差値が小さいほど，係り受けの距離がばらけていないため，つまりは係り受け先が遠く離れていない傾向があることを示している．具体的には，1.76 以下は「読みやすい」，4.4 以下は「ふつう」，それより大きい場合は「読みづらい」と表示される．

⑤文の分割候補

長文の場合，節があると分割しやすい．そこで，節が基準の文節を含めて 5 文節以上あり，かつ，それを除いた文節が 4 文節以上ある場合には，その文は，分割可能であると判定する．この判定をすべての文に対して行い，分割可能と思われる文が全体のどれぐらいを占めるかで，評価を 6 パターンで表示する．具体的には，0%なら「ない」，10%以下なら「ほとんどない」，30%以下なら「少しある」，50%以下なら「ある」，70%以下なら「けっこうある」，70%を超えるなら「かなりある」と表示される．

⑥文の長さ

50 文字以下の文を除外し，残った文の平均をもとに評価を 4 パターンで表示する．60 文字以下であれば「許容範囲内」，80 文字以下であれば「少し長め」，100 文字以下であれば「長め」，それ以上であれば「長過ぎ」とした．なお，この値は BCCWJ コーパスにおける新聞の平均文長をもとに決めた．おおよそ 60 文字以内が標準だと思われる．

新聞コーパス　：5412 文 (最小 30 文字〜最大 146 文字, 平均文長 56.3 文字)
教科書コーパス：8736 文 (最小 26 文字〜最大 115 文字, 平均文長 55.8 文字)
白書コーパス　：32544 文 (最小 39 文字〜最大 221 文字, 平均文長 79.5 文字)

⑦タイプ

本稿で述べた学生の文を教科書タイプ・白書タイプ・新聞タイプに分類するためのモデルを用いた判定方法が利用される．文章全体を構成する各文が，どういったタイプの文かを表示する．

5.2 文ごとの各種情報

文ごとに提示すべき内容は，個々の文に対して，推敲して書き替えるべきかどうかを提案するものである．そこで，図 11 に示すように，まず語彙レベルと語種ごとに色分け表示することで，文を構成する語彙の難易度を視覚的に把握することができるようにした．語彙レベルは，レベル 1

ⓒ 2019 Information Processing Society of Japan

（青系色）からレベル 5（赤系色）で色付けされる．語種では和語は青色，漢語は赤色で色付けされる．和語は比較的，教科書で使用されることが多く，文章が長くなりがちになってしまう反面，印象が柔らかくなる．一方，漢語は文章が短くまとまりやすくなる半面，印象が堅くなる．したがって，使用バランスが重要であるが，こういったツールによって視覚的に使用状態を判断できる．

図 11　語彙レベルごと，語種ごとの色分け表示例

また，補足節や連体節の多用や長文に対しては，文の分割処理を促す必要がある．現在は試作段階のため，節が基準の文節を含めて 5 文節以上あり，かつ，それを除いた文節が 4 文節以上ある場合には，分割可能であると定め，図 12 のように文字数表記（図の例では 61）の下に分割候補記号を表示し推敲に向けた案内をしている．また，図 13 のように分割候補を表示することも現段階では考えており試作中である．現在は分割後の文の長さなどを考慮せず，節を基準に分断し，言葉を補っているため，図内の候補 01 のように少し不自然な分割候補も出てしまうことがある．

図 12　文の分割案内

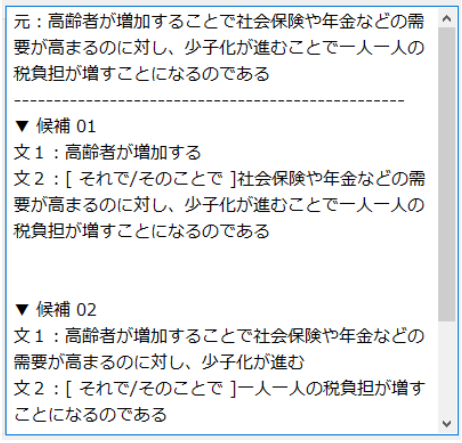

図 13　文の分割修正候補例

6. おわりに

本研究では，ICT を利用した学生のレポート文の推敲支援・推敲指導を目的に，文の各種特徴量を定め，それらを概観することが可能なツールを作成した．また，この特徴量をもとに，学生の書く文と教科書・白書・新聞を比較した．

学生の文の特徴として，実験から補足節と連体節を多用してしまう傾向があることがわかり，特に補足節を 1 つの文内に何度も使用してしまう点が実データからも散見された．また，短い修飾によって小刻みに文の前方から文をつなぎ合わせていくため，最終文節にたどり着くまでに何度も形式名詞や普通名詞で内容が置き換えられている傾向がわかった．

これらの修正案としては，①補足節を連体節に置き換えて単純化，②補足節や連体節で述べられている語を 1 つの語彙で置き換え，③文を分割し短文化，などが考えられる．しかし，①や②の方法は，語彙力が必要となってくるため，③の文の短文化に向けて，試作を行った． 今後は，文の分割ルールを明確にし，実用段階を目指していきたい．

参考文献

[1] 中島利勝，塚本真也．知的な科学・技術文書の書き方．コロナ社, 1996
[2] 塚本真也．知的な科学・技術文書の徹底演習．コロナ社, 2007.
[3] 又平恵美子，竹内純人，大野博之，稲積宏誠．文章作成支援ツールによる日本語文章力育成．私立大学情報教育協会 ICT 活用教育方法研究 第 13 巻 第 1 号 p.16-20, 2010.
[4] 横林博，菅沼明，谷口倫一郎.係り受けの複雑さの指標に基づく文の書き換え候補の生成と推敲支援への応用．情報処理, Vol.45, No.5, pp.1451-1459, 2004.
[5] 柴崎秀子，原信一郎．12 学年を難易尺度とする日本語リーダビリティ判定式．計量国語学, 27-6, pp.215-232, 2010.
[6] 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.
[7] 李在鎬．日本語教育のための文章難易度研究．早稲田日本語教育学, Vol.21, pp.1-16, 2016.
[8] 教科書コーパス語彙表・BCCWJ 主要コーパス語彙表．http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html, (参照日 2017.1.4).
[9] 国立国語研究所，教育基本語彙の基本的研究 増補改訂版．明治書院, 2009.
[10] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004.
[11] 工藤拓，松本裕治．チャンキングの段階適用による係り受け解析．情報処理, Vol 43, No. 6, pp.1834-1842, 2002.
[12] 益岡隆志，田窪行則．基礎日本語文法・改訂版．くろしお出版, 1992.
[13] 阿部純一，桃内佳雄，金子康朗，李光五．人間の言語情報処理-言語理解の認知科学．サイエンス社, 1994.
[14] 現代日本語書き言葉均衡コーパス(BCCWJ). http://www.ninjal.ac.jp/corpus_center/bccwj/ , (参照日 2016.6.24)
[15] Stefan Rüping mySVM-Manual, University of Dortmund, Lehrstuhl Informatik 8. http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/ , (参照日 2018.1.24)

ⓒ 2019 Information Processing Society of Japan