文章構造に基づく難易度推定と教育への活用方法の検討

青山学院大学　社会情報学部　稲積 宏誠
東京医療保健大学　医療保健学部　大野 博之

研究成果要約

1．研究活動の概要

高等教育の現場でも、アカデミックライティングに類する授業が数多く展開されてきている。しかし、学習者のレベルが多様である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、適切な教材資料の選定や効率的な作文指導などの教育方法の改善が求められている。そこで本研究では、これらの教育方法の改善を実現するために、自然言語処理技術を活用した新しい学習支援・指導支援方法の提案を目的とし、それを実現するためのツールを開発する。

適切な教材資料の選定や効率的な作文指導を実現するには、文章を客観的に評価する環境が必要であるが、そのためには文を構成する語彙の難易度や文の構造を概観できなければならない。そこで、主として教科書コーパスと白書コーパスを分析対象として、自然言語処理技術とICTを活用することによって、①語彙の難易度の決定ルールの作成、②文の節構造の判定と図示化、③語彙と構造の特徴量に基づく文の特徴を表すモデルを作成し、これらの情報から、④各文が教科書タイプか白書タイプかを判定する。ここで、教科書と白書に注目したのは、高校以下で扱われる教科書で用いられている文は、教育現場で「構造的にわかりやすい」表現で理解を促すことを目的としており、一方白書で用いられている文はそのような配慮が少ないと仮定したことによる。この①から④の取り組みをもとに、さらに主観評価を加えることによって学習支援・指導支援を行うためのツールとして専用のテキストエディタを開発する。

2．研究成果の概要

学習支援・指導支援方法の提案とそれを実現させるためのツールを開発するために下記①～④の取り組みを行った。

① 語彙の難易度（語彙レベル）の付与のためのデータ整備とルールの作成
「教科書コーパス語彙表」と「BCCWJ主要コーパス語彙表」をベースとした語彙レベルデータを整備し、文の各形態素に対して、語彙レベルを付与するルールを作成した。

② 文の節構造（補足節・連体節・副詞節）の判定と図示化の実現
文構造の解析のため、形態素解析・係り受け解析のみでは判定することのできない節構造を判定するためのルールを作成し、図示化可能なツールを開発した。

③ 語彙と構造の特徴量に基づく文の特徴を表すモデルの作成

どのような語彙と構造の特徴量を用いることで客観的に文を評価するためのモデルが作成できるのかを検討し、各特徴量を算出するためのツールを作成した。特徴量は、語彙に関するものとして「平仮名・片仮名・漢字」等の含有率や品詞別の含有率、和語や漢語などの語種の含有率など約25種類、構造に関するものとして係り受け距離や節構造など約30種類である。

④ 機械学習に基づく文の教科書タイプ・白書タイプの判定ルールの作成
③で得られたモデルを利用して、対象とする文が教科書タイプか白書タイプか判定するルールを作成した。具体的には、文の特徴を表す特徴量の中の19属性で構成されるSupport Vector Machine（SVM）による分類器を作成した。

⑤ 学習支援・指導支援を行うためのツールの開発
①から④までの結果を活用するための専用エディタを開発した。

3. 成果活用について

本研究を構成する一部については、すでに研究会等で対外発表を行っているが、授業内で利用することを想定して開発した学習支援・指導支援ツールについては、未発表である。これについては、実際の授業実践を踏まえたうえで、その成果について対外的発表を行っていきたい。

4. 今後の研究課題

本研究では、文のわかりやすさについての各種情報を概観できる環境を提供することができた。また、文のわかりやすさの評価についての客観的な分析結果に基づく学習支援・指導支援ツールを開発したが、その有効性については不十分である。今後、検証実験を含めて実際に授業内で利用していくことによって、教師側と学習者側それぞれにとっての効果的な活用の仕方や支援ツールとしての機能強化について検討していく必要がある。

研究成果報告

1．はじめに

高等教育の現場でも、母語としての日本語教育の重要性が広く認められ、アカデミックライティングなどの授業が展開されてきている。しかし、学習者のレベルが多様である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、適切な教材資料の選定や効率的な作文指導などの教育方法の改善が求められている。このためには、教育の負担をいかに軽減し、より質の高い教育を行うことが重要である。そこで、筆者らは、自然言語処理技術を活用した学習支援・指導支援の取り組みを行っている。

校正上の形式チェックについては、ある程度のものは、これまでの筆者らを含めた又平ら（2010）の取り組みで実現しており、どこを修正すべきかを明示することができる。しかし、文のわかりやすさに着目した時、何が原因でわかりづらくなっており、どのように修正すべきかまでは指摘できていない。もし､わかりにくい文の特徴を機械的にとらえることができれば、形式チェックとは違う「推敲」の指導支援につなげることができるようになる。

わかりにくさとは逆であるが、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では、すでにウェブサービスとしても提供されている取り組みとして、次の3つが挙げられる。柴崎・原（2010）の取り組みでは、文中の平仮名の割合、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章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（2011）の取り組みでは、教科書を基に作成した規準コーパスから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（2016）の取り組みでは、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示しており、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。ただし、これらに共通するのは、対象となる文全体に対して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。

そこで本研究では、個々の文を対象にすることも踏まえ、自然言語処理技術を活用した新しい学習支援・指導支援方法の提案を目的とし、それを実現するためのツールを開発する。

適切な教材資料の選定や効率的な作文指導を実現するには、文章を客観的に評価する環境が必要であるが、そのためには文を構成する語彙の難易度や文の構造を概観できなければならない。そこで、主として教科書コーパスと白書コーパスを分析対象として、自然言語処理技術とICTを活用することによって、①語彙の難易度の決定ルールの作成、②文の節構造の判定と図示化、③語彙と構造の特徴量に基づく文の特徴を表すモデルを作成し、これらの情報から、④各文が教科書タイプか白書タイプかを判定する。ここで、教科書と白書に注目したのは、高校以下で扱われる教科書で用いられている文は、教育現場で「構造的にわかりやすい」表現で理解を促すことを目的としており、一方白書で用いられている文はそのような配慮が少ないと仮定したことによる。この①から④の取り組みを基に、さらに主観評価を加えることによって学習支援・指導支援を行うためのツールとして専用のテキストエディタを開発する。

2. 語彙の難易度

語彙の難易度は、誰を対象とするかで変わる。例えば、日本語を母国語としない留学生等を対象とする場合は、留学生向けの日本語教科書が基準となる。また、日本語を母語とする場合でも、分野別専門用語まで含めると、明確に難易度を定義するのは難しい。

そこで、本研究の対象が「日本語を母語とする学習者」であることから、一般的に使用されている語を対象に難易度を付与することを考える。すなわち、初等教育、中等教育および書籍等で一般的に使用されている語を対象とする。以下、語彙の難易度の決定方法について説明する。

2.1. 語彙の難易度を決定する基準データ

語彙の難易度を定めるために、3つの基準データを用いる。コーパス開発センター（参考文献URL参照）にて公開されている「教科書コーパス語彙表」・「BCCWJ主要コーパス語彙表」と、国立国語研究所（2009）「教育基本語彙の基本的研究 増補改訂版」内で示されている「新阪本教育基本語彙」である。

①国立国語研究所コーパス開発センターで配布されている「教科書コーパス語彙表」は、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種ずつを対象とした「教科書コーパス」の語彙の一覧である。教科書に出現する約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語を利用する。

②国立国語研究所コーパス開発センターで配布されている「BCCWJ主要コーパス語彙表」は、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語を利用する。

③「新阪本教育基本語彙」は、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語を利用する。

2.2. 語彙の難易度のレベル表記

3つの基準データを使用することから、語彙の難易度として、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。

固有名詞については、語彙表からのデータ抽出は基本的に省き、形態素解析の段階で固有名詞と判定された時に、一律レベル1を付与することとした。初出年次等を基にレベルを決定すると、例えば、同じ都道府県であるのに「青森」はレベル1で、「秋田」はレベル2ということ

が起きる。このような事態を避けるためである。ただし、「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。

2.3. 語彙の難易度の付与

語彙の難易度を示すレベルを付与するには、形態素解析を行い、基準データとのマッチングを行うことが基本方針である。

準備：

教科書コーパス語彙表･主要コーパス語彙表は、Taku Kudoら（2004）による形態素解析ツール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を利用するため、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。例えば、「取る」は、教科書コーパス語彙表では、基本形は同じであり、活用が異なっている。一方、本研究では工藤ら（2002）による係り受け解析ツールCabochaをIPADIC辞書にて動作させている。そのため、MeCabもこれに合わせIPADIC辞書を使用する必要がある。そこで、このような問題に対処するため、教科書コーパス語彙表と主要コーパス語彙表は、図1に示すように修正が必要なものをある程度機械的に収集し、その後1データずつ人手で修正して利用する。また、品詞については、表1のように変換ルールを作成し対応する。品詞の変換は機械処理を容易にするために記号化している。また、新阪本教育基本語彙は、語の基本形の欄に常用漢字外であることを示す記号や、別表記の羅列などがあり、そのままでは使用できないため、表記を修正した。

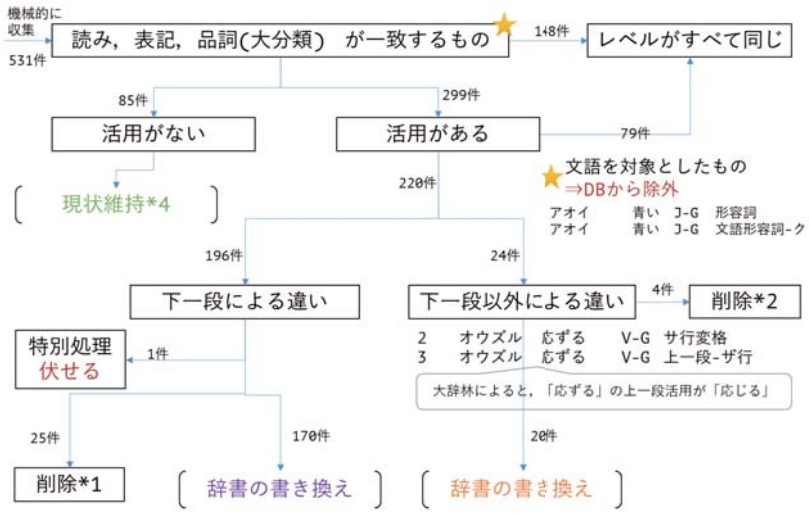


図1　辞書の書き換え作業

最終的には、図2に示すように4つの語彙のデータベース（DB）と、ユーザーが独自に語彙のレベルを登録できる「ユーザーDB」を準備した。教科書コーパス語彙DBと主要コーパス語彙DBと教育語彙DBについては、元データからの語種（和語・漢語・外語など）の情報も一緒に持たせている。

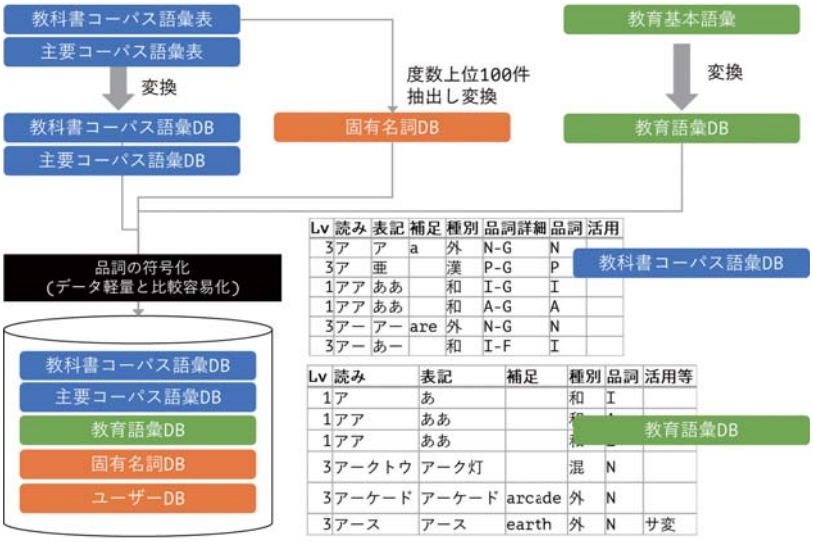
表１　品詞の変換ルール

教科書コーパス・主要コーパス語彙	新阪本教育基本語彙	変換後
名詞	感・名、感・名・ス他、名、名・ス自、名・ス自・形動、名・ス自・代、名・ス自・副、名・ス自・副・形動、名・ス自他、名・ス他、名・ス他・形動、名・ス他・接尾、名・ス他・副、名・感、名・代、名・代・感、名・副、名・副・ス自、名・副・接	N
動詞	サ変自、サ変他、下一自、下一自他、下一他、下一他・接尾、下二自、下二他、五自、五自・接尾、五自他、五他、四自、四他、上一自、上一自他、上一他、上二他	V
副詞	副、副・ス自、副・ス自・名、副・ス他、副・トタル、副・感、副・形動、副・形動・ス自、副・形動・名、副・名、副・名・ス自、副・名・形動	A
接尾辞	接頭・接尾、接尾、造、代・接尾、副・接尾、名・形動・接尾、名・接尾、名・連体・接頭・接尾	S
形状詞	トス自、トス自・副、トタル、トタル・ス自、トタル・名、形動、形動・ス自、形動・接、形動・副、形動・副・ス自、形動・副・名、形動・名、形動・名・ス自、名・形動、名・形動・ス自、名・形動・ス他、名・形動・副、名・副・形動	F
形容詞	形、形ク	J
接頭辞	接頭、接頭・形動、名・接頭	P
感動詞	感、接・感	I
代名詞	代、代・感、代・名	R
連体詞	連体、連体・形動・副	D
接続詞	接、接・副、接・連語	C

教科書コーパス・主要コーパス語彙	変換後
名詞－普通名詞－一般	N－G
名詞－普通名詞－サ変可能	N－S
名詞－普通名詞－形状詞可能	N－F
名詞－普通名詞－副詞可能	N－A
名詞－普通名詞－サ変形状詞可能	N－P
名詞－助動詞語幹	N－V
動詞－一般	V－G
動詞－非自立可能	V－U
感動詞－フィラー	I－F

教科書コーパス・主要コーパス語彙	変換後
感動詞－一般	I－G
形状詞－タリ	F－T
形状詞－一般	F－G
形状詞－助動詞語幹	F－V
形容詞－一般	J－G
形容詞－非自立可能	J－U
接続詞	C－G
接頭辞	P－G
接尾辞－形状詞的	S－F

教科書コーパス・主要コーパス語彙	変換後
接尾辞－形容詞的	S－J
接尾辞－動詞的	S－V
接尾辞－名詞的－サ変可能	S－S
接尾辞－名詞的－一般	S－G
接尾辞－名詞的－助数詞	S－N
接尾辞－名詞的－副詞可能	S－A
代名詞	R－G
副詞	A－G
連体詞	D－G



Lv	読み	表記	補足	種別	品詞詳細	品詞	活用
3	ア	ア	a	外	N-G	N	
3	ア	亜		漢	P-G	P	
1	アア	ああ		和	I-G	I	
1	アア	ああ		和	A-G	A	
3	アー	アー	are	外	N-G	N	
3	アー	あー		和	I-F	I	

Lv	読み	表記	補足	種別	品詞	活用等
1	ア	あ		和	I	
1	アア	ああ				
1	アア	ああ				
3	アークトウ	アーク灯		混	N	
3	アーケード	アーケード	arcade	外	N	
3	アース	アース	earth	外	N	サ変

図２　語彙のレベルを定義するデータベース

手順：

実際に文を解析し、語彙レベルを付与する手順を説明する。文を自然言語解析ツールであるMeCab・Cabochaを用いて、形態素解析および係り受け解析を行い、形態素と文節を得る。また、文節をまたがない範囲で複合語を得る。複合語を得る理由は、例えば「目覚まし時計」は形態素解析では「目覚まし」と「時計」に分かれてしまうが、新阪本教育基本語彙では、「目覚まし時計」で語が登録されており、形態素解析だけでは、レベルを決定するのに不十分なケースがあるためである。したがって、まず複合語の語彙レベルが決定可能かを試み、決定できなかった場合に複合語を構成する形態素ごとにレベルの決定を試みる。複合語を対象とするか、形態素を対象とするかで一部処理が異なるがレベルを決定する手順の概要は次の通りである。最終的に図3のように、各語に語彙レベルが付与されることになる。

①固有名詞DBを検索し、該当するものがあれば、そのレベルに決定
②「①」以外の固有名詞であった場合は、レベルを1とし処理を終了
③助詞や助動詞、記号などであれば、処理を終了
④ユーザー DBを検索し、該当するものがあれば、そのレベルに決定
⑤残りの3つの辞書を検索し、語彙レベル決定のための候補を抽出
　※元の調査対象語に漢字が使用されている場合は、それによってマッチング候補を絞り込むことで、同音異義語による誤判定を極力避けるようにする。
⑥「⑤」で得られた候補から、以下のルールで語彙レベルの決定を試行
　・候補が1つの場合は、そのレベルに決定
　・候補がすべて同じレベルであれば、そのレベルに決定
　・上記以外は、表記・読み・品詞の一致度で点数化し、優先度の高いもののレベルに決定
⑦読みの一致のみで、候補を抽出し、語彙レベルの決定を試行
⑧教科書コーパス語彙DB内の接頭辞と接尾辞を他の言葉につなげ、決定できないか試行
⑨カタカナ語であれば、語尾に長音を付加もしくは除去して、決定できないか試行

⑩「①～⑨」で決定できなかった場合は、レベル付与に失敗

図３　語彙レベルの付与例

3. 文のタイプの判定

1文が「教科書タイプ」か「白書タイプ」かを判定するためには、教科書の文と白書の文が必要になる。そこで「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）を利用することとした。

3.1. BCCWJからの文章抽出

BCCWJには、教科書コーパス（OT）は412サンプル、白書コーパス（OW）は1500サンプルあり、XML形式で提供されており、1文がsentenceタグで囲まれているのが基本形式である。しかし、文ではないものにもsentenceタグが付いていたり、鍵括弧を伴った文中文があると、このタグが単純に1文に対応していない場合もある。また、箇条書きの先頭記号や空白等も混在しており、きれいな文の素材ではない。

そこでまずは、1文単位で整形された文を取得することを考える。sentenceタグで囲まれており、かつ、句点（。か。）で終わっているものを機械的に抽出し、OT約38000文、OW約91000文を得た。そして、文頭の接続詞と副詞と感嘆詞を削除しつつ、以下のように使用する文を絞り込んだ（文頭を一部削除したのは、のちに文の指標値を取得する際に、係り受け距離に大きく影響を与えてしまうからである）。

・10文節未満または30文節を超える文を除外
・箇条書き記号の除去や鍵括弧を含む文を除外
・％を3つ以上含む文を除外
・数字もしくは英字を10文字以上含む文を除外

最終的には、OT約8700文、OW約32000文を得た。これを語彙レベルn（1≦n≦5）以下のみを含む文に表2のようにグループ分けした。例えば、レベル3のグループは、レベル4以

表２　レベルグループごとの教科書・白書コーパスの文数

グループ	教科書コーパスから	白書コーパスから
レベル1	238	4
レベル2	1560	732
レベル3	3602	10847
レベル4	2933	14001
レベル5	403	6961

上の語彙を含まず、レベル3以下の語彙のみで構成されることを意味し、そのような文が教科書コーパスには3602文、白書コーパスには10847文があったということになる。

3.2. 文の構造的指標

語彙レベルが付与された各文に対して、その特徴量を求めた。特徴量としては、以下の語彙に関するものと文構造に関するものを定め、算出した。算出には専用のツールを作成し行った。また、補足節・連体節・副詞節を判定する必要があるため、益岡・田窪（1992）による「基礎日本語文法・改訂版」に基づいて判定ルールを実装した。

3.2.1. 語彙に関する特徴量

語彙に関する特徴量としては、文節数、①「語彙Lv1率～語彙Lv5率」、②「平仮名率、片仮名率、漢字率、文字他率」、③「動詞含有率、形容詞率、形容動詞率、副詞率、助詞率、読点率」、④「和語率、漢語率、外語率」、⑤同音異義語率、⑥「漢1率、漢2率、漢3率、漢4率、漢5率、漢6率、漢7率、漢他率」を算出した。

①については、本研究の一部である語彙の難易度決定方法によって得た。②の「文字他率」とは、文字種が平仮名でも片仮名でも漢字でもない率である。③は、MeCabで判定された品詞による。④は、「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を基に求めた。⑤は、各種コーパスを利用して作成した独自の同音異義語辞書による。⑥の「漢1率」とは学習年次が小学1年生の漢字の割合である。「漢7率」は中学生で学習すべき漢字、「漢他率」はJISX0208で規定されている第二水準以下の漢字で、小学～中学までの学習漢字に含まれないものの率である。

3.2.2. 文構造に関する特徴量

文構造に関する特徴量としては、⑦「動詞・サ変文節率、用言文節率、受動態含有率、使役率、否定率、接続助詞（が）率、指示詞率、AのB率」、⑧「平均係受距離、最大係受距離、係受距離標準偏差、最終文節への入射集中率、各文節への入射数の標準偏差」、⑨「補足節占有文節率、連体節占有文節率、副詞節占有文節率、補足節最大係受距離、連体節最大係受距離、副詞節最大係受距離、補足節係受距離平均、連体節係受距離平均、補足節係受距離累積、連体節係受距離累積、副詞節係受距離累積、節係受距離標準偏差」、⑩「チャンク圧縮率A、チャンク圧縮率B、チャンク圧縮率C、チャンク圧縮率D」を算出した。

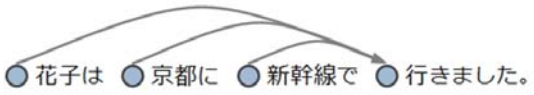
⑦の「動詞・サ変文節率」は、動詞および『する』と接続して動詞的な使われ方をしているサ変名詞から求めている。「用言文節率」は、これに加え、形容詞と『名詞＋だ』も追加して得る。「AのB率」は、『研究者の教育』のように連体助詞「の」を伴うものである。⑨は、補足節・連体節・副詞節を構成する文節が、文に対してどの程度占有しているかを表す。それぞれの節が入れ子になっている場合は、重複してカウントする。⑧の係り受けに関する特徴量と⑩のチャンクに関する特徴量は後述する。

3.2.3. 文構造の係り受けに関する特徴量

係り受けに関する特徴量は、以下に示す通りとする。具体的な例を図4に示す。

平均係受距離：文中での係り受け距離の平均値

最大係受距離：文中での係り受け距離の最大値
係受距離標準偏差：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
各文節への入射数の標準偏差：係り受け数を分母とした時の最初の文節以外への入射数の標準偏差



項目	値
平均係受距離	2.00
最大係受距離	3
係受距離標準偏差	0.82
最終文節への入射集中率	1.00
各文節への入射数の標準偏差	1.41

項目	値
平均係受距離	1.33
最大係受距離	2
係受距離標準偏差	0.47
最終文節への入射集中率	0.67
各文節への入射数の標準偏差	0.82

図４　係り受けに関する特徴量算出結果例

「係受距離標準偏差」が大きいほど、文節ごとに係り受け先までの距離が大きく異なることを意味している。「最終文節への入射集中率」が大きいほど、それぞれの文節が最後の文節を修飾するシンプルな係り受けになっていることを示す。「各文節への入射数の標準偏差」の値が大きいほど、特定のいくつかの文節に係り受けが集中していることを示す。

3.2.4. チャンクに関する特徴量

阿部ら（1994）によると、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行う際に、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。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。ここでのチャンクとは、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」を指す。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。

この考えに基づいて、1文節1チャンクを基本とし、4つの方法（チャンクA～チャンクD）

チャンクA	「助詞-連語」　例）AによるB 「助詞-連体化」　例）AのB 「連体詞」　例）あのA，そのB を1チャンクとする	下記例で 6チャンク
チャンクB	チャンクAに加え、節があった場合はそれをまとめて1チャンクとする (入れ子の節は一番外側でカウント)	下記例で 2チャンク
チャンクC	チャンクBにおいて、入れ子の節は入れ子ごとに1チャンクとする	下記例で 3チャンク
チャンクD	チャンクCにおいて、節をチャンクではなく2チャンクとカウントする	下記例で 5チャンク

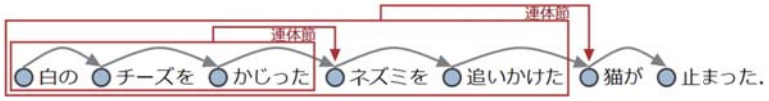


図５　チャンキングによるチャンク数カウント例

でチャンキングを行い、それによってチャンキングしない場合と比べて、どれだけチャンクが圧縮されたかを示したのが「チャンク圧縮率」である。具体的なチャンキング方法と例を図5に示す。

3.3. 教科書タイプと白書タイプの文の分類

文の構造的指標を特徴量として、教科書タイプと白書タイプに分類する予備実験を行う。ここでは、言葉の難しさや文の長さに依存しないように、表2における語彙レベル3以下（3602文と10847文）とレベル4以下（2933文と14001文）のデータの中から、文節数が12 ～ 24までの文を扱うこととした。極端な短文と長文は、議論する意味が薄れるからである。また、12 ～ 24文節も文長の差が大きいため、さらに表3のように3グループに分割した。これらのデータを用いて、予備実験を何度か繰り返し、分類に使用する指標を取捨選択する。最終的に選択した指標で分類器を作成することが目的である。

表3　レベルグループごとの教科書・白書コーパスの文数

グループ	文の数	内、教科書	内、白書	文長の平均
12文節～ 14文節	8482文	2305文	6177文	60.2文字
15文節～ 19文節	9228文	1575文	7653文	77.4文字
20文節～ 24文節	4699文	386文	4313文	100.6文字

3.3.1. 予備実験

教科書タイプと白書タイプの文の分類を行うために、SVMを使用する。本研究では、Stefan Ruping（2000）によるmySVMを利用した。この予備実験の目的は、分類すること自体を目的とするのではなく、それぞれの指標の重みベクトルによって、各指標がどちらのタイプ側に影響を与えるのかを知るためである。なお、0 ～ 1の値となっていないものは、正規化してから利用する。

「12文節～ 14文節」グループを例に説明する。このグループは、教科書の文は2305文あり、白書の文は6177文あり、素材文数が大きく偏っている。そこで、少ないほうのデータ数の半分を訓練データとして、ランダムに取り出し、それと同数のデータをもう片方からもランダムに取得する。すなわち、このグループでは、それぞれ1152文、合計2304文を訓練データとする。残りのデータは、すべてテストデータとする。これをmySVMにかけて、それぞれの指標の重みベクトルを取得する。この作業を100回繰り返し、重みの分布を俯瞰する。

例えば、漢字の学習年次が1年と6年の項目に対して、SVMによる重みベクトルを調べると図6のようになる。横軸は試行回数、縦軸は重みベクトルである。マイナス値は、教科書寄りであることを示し、プラス値は白書寄りであることを示す。当然の結果ともいえるが、低学年の漢字は教科書寄り、高学年の漢字は白書寄りに影響を与えていることが確認できる。このように、各指標がどちらのタイプに影響するかを調べる。

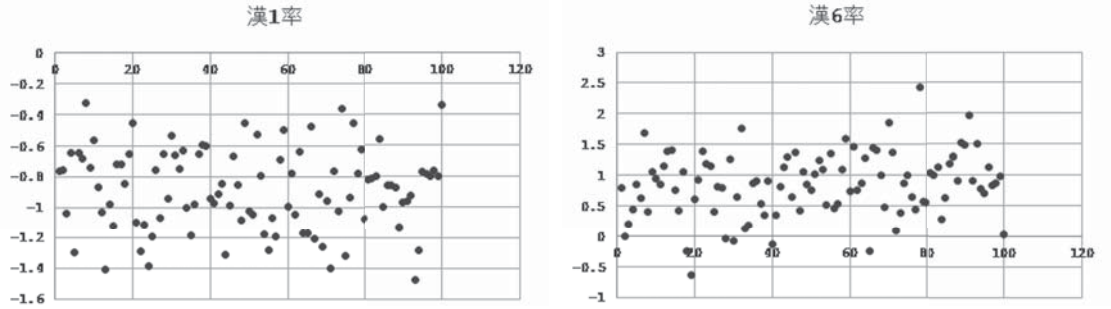


図6　SVMの100回試行による重みの分布例

教科書タイプ・白書タイプの判別の目的は、語彙の難易度などではなく、文の構造的指標での差異を発見したいことから、最初の予備実験では使用したが、語彙の難易度レベルの各含有率や漢字の学習年次の含有率などは、最終的には指標から省いた。また、予備実験によって、教科書寄りとも白書寄りともいえない重みベクトルだったものを除外しながら、2回ほど同様の作業を繰り返し、最終的に表4に示す19個の指標を使用することとした。各値が高いほうが、表に示された側のタイプであることを意味している。△が付いているものは、明確に差異が出ないこともあった指標である。「補足節占有文節率」については、文節の長さでどちらのタイプ寄りか変わるが、その他の指標は概ね同じ方向性を指している。

表4　教科書タイプと白書タイプに分類する19指標

指標	12-14文節	15-19文節	20-24文節
形容詞率	教科書	教科書	教科書
形容動詞率	白書	白書	白書
副詞率	教科書	教科書	教科書
助詞率	教科書	教科書	教科書
動詞・サ変文節率	教科書	教科書	教科書
使役率	教科書	教科書	教科書
否定率	教科書	教科書	教科書
接続助詞（が）率	白書	白書	△白書
指示詞率	教科書	教科書	教科書
係受距離標準偏差	白書	白書	△白書
最終文節への入射集中率	教科書	教科書	教科書
各文節への入射数の標準偏差	教科書	教科書	教科書
補足節占有文節率	教科書	教科書	白書
副詞節占有文節率	教科書	教科書	教科書
連体節最大係受距離	白書	白書	白書
連体節係受距離平均	△教科書	教科書	教科書
副詞節係受距離平均	白書	白書	白書
連体節係受距離累積	教科書	教科書	教科書
チャンク圧縮率D	教科書	教科書	教科書

3.3.2. 分類実験

教科書コーパスの文が、白書寄りであると判定されるものは、それほど多くはないが、白書コーパスの文が教科書寄りと判定されるものは、多く存在すると予想される。白書コーパスの中にも、相応の教科書的な表現が含まれていると思われるためである。したがって、まずは白書コーパスから、教科書寄りの文を除外することを考える。

まず表4の19個の指標を使用し、予備実験と同様に100回の試行を行い、訓練データおよびテストデータの精度の合計値がもっとも高い時の分類器を採用した。次に、この分類器で教科書コーパスと白書コーパスを分類し、白書コーパスの文で教科書タイプと判定された確信度0.9以上の文を除外した。そして再び、残ったデータでこの作業と同じことを繰り返し3回行い、白書コーパスから教科書タイプ寄りの文を除外した。

そして、教科書コーパス文と残った白書コーパスらしい文を用いて、分類器を作成した。これもまた、予備実験と同様に100回の試行を行い、訓練データおよびテストデータの精度の合計値がもっとも高い時の分類器を最終的なものとした。つまり、これを12 ～ 14文節、15文節 ～ 19文節、20 ～ 24文節の3グループに対して実施したので、3つの分類器が作成できたことになる。

この分類器によって、分類される一例を図7に示す。いずれも12文節の文であり、上2つが教科書タイプ、下2つが白書タイプである。19個の指標によって判定されているため、一概に特徴を述べることはできないが、白書タイプへの判定に影響する副詞節の係受距離平均や形容動詞率が高い。

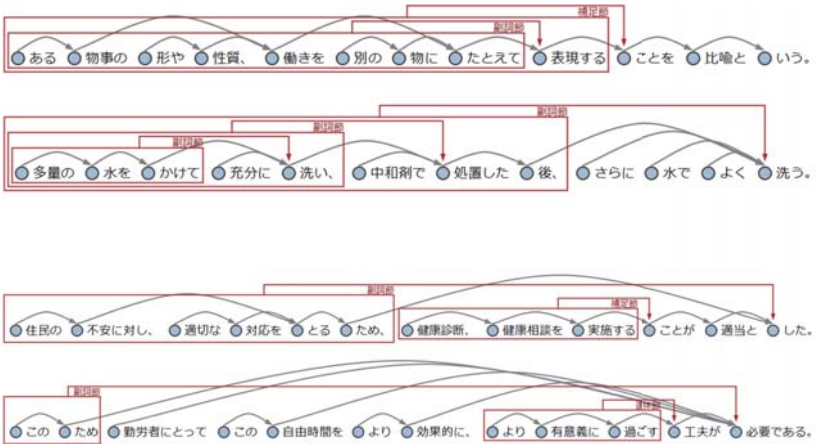


図7　教科書タイプの２文と白書タイプの２文

4. 学習支援･指導支援のための文章難易度表示テキストエディタ(仮称「テキストめもぱ」)

文章の難易度推定を学習支援・指導支援に利用するためには、利用者が教員もしくは学生自身であることから、ある程度容易にツールを扱える必要がある。そこで、図8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開発した。このエディタは、本研究で述べた語彙レベルと文章タイプの結果だけでなく、表5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。本ツールは、仮称として「テキストめもぱ」と呼ぶことにする。

図８ 「テキストめもぱ」利用画面

表５ 教科書タイプと白書タイプに分類する19指標

表示情報	説明
文章全体の評価	7つの視点で、全体の評価を表示する 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
漢字含有率	全体の漢字含有率をグラフで表示する
語彙レベルの分布	全体の語彙レベル１～５まで割合およびグラフと、個数を表示する
語種の分布	和語・漢語・外来語・混合語の割合およびグラフと、個数を表示する
文ごとの各種情報	・教科書タイプか白書タイプかの表示 ・語彙レベルごと、語種ごとの色分け表示 ・文字数、分割可能性、係り受けの距離がばらけているもの ・係り受け関係の図示、形態素・文節情報、語彙レベル情報

4.1. 文章全体の評価

7つの視点で、文章全体を評価した結果を表示する。全体の様子を把握するために利用する。7つの視点の判定基準と評価結果は、以下の通りである。

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、10%未満は「漢字が少なすぎる」、20%未満は「漢字が少ない」、45%未満は「ちょうど良い」、50%未満は「漢字が多い」、それ以上は「漢字が多すぎる」と表示される。なお、この値はBCCWJコーパスにおける新聞・教科書・白書の漢字含有率を基に決めた。

新聞コーパス　：5412文（漢字含有率　平均41.4%）
教科書コーパス：8736文（漢字含有率　平均34.2%）
白書コーパス　：32544文（漢字含有率　平均47.8%）

②語彙の難しさ

語彙レベルの分布を基に、評価を5パターンで表示する。具体的には、語彙レベル5以上が10%以上含む場合は「難しめ」、語彙レベル4以上が10%以上含む場合は「少し難しめ」、語彙レベル3以上が10%以上含む場合は「ふつう」、語彙レベル2以上が10%以上含む場合は「やさしめ」、それ以外は「かなりやさしめ」と表示される。

③文のやわらかさ

語種の和語と漢語の分布を基に、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、残った文の平均を基に評価を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基に決めた。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55.8文字）
白書コーパス　：32544文（最小39文字～最大221文字、平均79.5文字）

⑦タイプ

本稿で述べた教科書タイプ・白書タイプの判定方法が利用される。文章全体を構成する各文が、教科書タイプか白書タイプのどちら側に判定されるかで、全体のタイプを9パターンで表示する。教科書タイプか白書タイプが同数の場合や判定ができる情報がない場合は、「バランス型」と表示される。具体的には、教科書タイプの割合と白書タイプの割合の差が、10%以下であれば「バ

ランス型」、20%以下であれば「少し教科書/白書タイプ」、30%以下であれば「教科書/白書タイプ」、40%以下であれば「けっこう教科書/白書タイプ」、40%を超えた場合は「かなり教科書/白書タイプ」と表示される。ただし9文節以下は、文が短いものとなるため、判定不能とした。

4.2. 文ごとの各種情報

ひとつひとつの文について、各種情報を表示し、どのような文なのかを検討できる画面である。以下に提示できる情報を説明する。

A. 教科書タイプか白書タイプかの表示

結果の画面に、教科書タイプの文と判定された場合は青、白書タイプの文と判定された場合は赤、判定不能な場合は黄色で、各文ごとに線が引かれて、一目で確認できるようになっている。図9を例にすると、4つの文があるが、上2つが黄色線なので判定不能であり、下2つが青色なので教科書タイプであると判定されたことになる。

B. 語彙レベルごと、語種ごとの色分け表示

図9に示すように、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。語彙レベルは、レベル1（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。

図９　ノーマル表示・語彙レベルごとの色分け表示・語種ごとの色分け表示

C. 文字数、分割可能性、係り受けの距離がばらけているもの

図10に示すように、3つの情報を表示する。文字数は、長文かどうかを把握しやすくするために表示している。また、分割可能性は、節が基準の文節を含めて5文節以上あり，かつ，そ

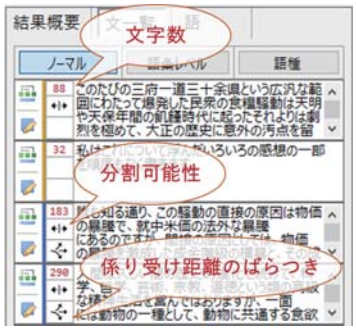


図１０　文の文字数、分割可能性、係り受け距離のばらつき表示

れを除いた文節が4文節以上ある場合には、表示している。このマークが表示されている場合は、文を分割して短文化できる可能性があることを示している。係り受け距離のばらつきについては、係り受け距離の標準偏差が4.4を超える場合にマークを表示する。図11の例のように、係り受け距離の標準偏差が高いと、係り受け先が遠く離れている傾向があり、わかりづらくなる。

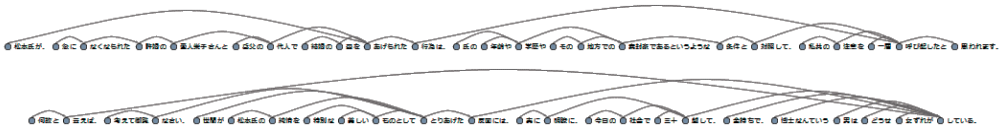


図11　係り受け距離の標準偏差が低い文（上）と高い文（下）

D. 係り受け関係の図示、形態素・文節情報、語彙レベル情報

1文の詳細を確認するために①係り受け関係の図示、②形態素・文節情報、③語彙レベル情報を表示することができる。図12のように、これらの情報は、ボタンで切り替えることができる。

①係り受け関係の図示では、形態素解析・係り受け解析から得られた情報から図示する。また、独自の判定処理にて認識した補足節・連体節・副詞節も、併せて表示することができる。これは、節の入れ子状態の確認や、節を基準に文を分割しようと考える際の手助けとなる。

②形態素・文節情報では、MeCabとCabochaによる解析結果および独自に判定している複合語情報を確認することができる。

③語彙レベル情報では、対象の文で使用されている語彙レベルの分布と、それぞれの語彙ごとのレベルが表示される。

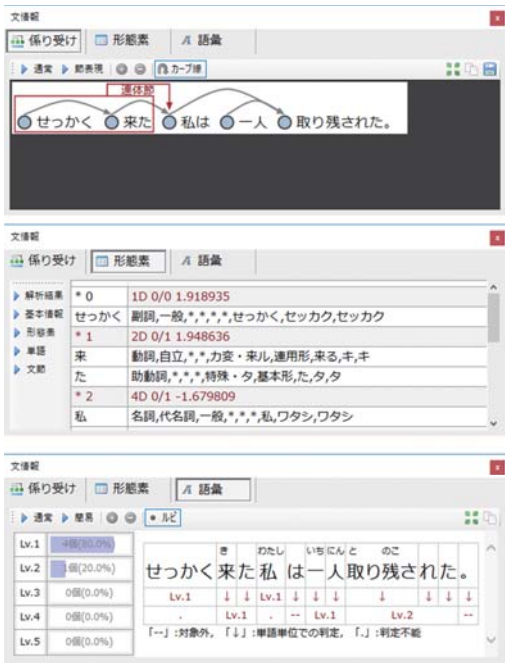


図12：係り受け関係の図示、形態素・文節情報、語彙レベル情報

5. おわりに

本研究では、文章構造に基づく難易度推定と教育への活用方法を検討することを目指して、文章に語彙の難易度を付与し、「構造的にわかりやすい」と思われる教科書タイプの文かどうかを判定する試みを行った。そして、これらを実際の教育現場で使用できるようテキストエディタを開発した。本エディタにより、対象文章の全体的なバランスと個々の文章の特徴を容易に確認することができるため、教員が教材資料の選定に活用したり、学生自身に使用させて、自らの文章に対する客観的な評価を与えるような利用も可能となる。

既存の研究では、漢字含有率や平仮名含有率、平均文長などによって、読みやすさが上がり難易度が下がることが示されている。確かに本稿では述べていないが、そういった傾向はデータからは見受けられた。しかし、文章指導の立場からすれば、用語を平仮名化させて、漢字含有率を下げたり、むやみに文を短くして短文化させるだけでは不十分である。このような単純な文字の置き換えではなく、和語・漢語の言い換えや、文の構造的な推敲に結び付く指導支援の足がかりとなる環境が本研究によって構築できたのではないだろうか。

なお、本研究の過程で、成果の一部として以下の対外発表を行っている。

(1)“文構造に基づく文の難易度を示す評価指標導出の試み”，大野博之，稲積宏誠，計量国語学会第六十一回大会

(2)“文の構造的指標に基づく分かりにくい文の分類方法の検討”，大野博之，稲積宏誠，電子情報通信学会　技術研究報告　教育工学研究会，ET2017-88

6. 今後の課題

文のわかりやすさの評価についての客観的な分析結果に基づく学習支援・指導支援ツールを開発したが、その有効性については不十分である。今後、検証実験を含めて実際に授業内で利用していくことによって、教師側と学習者側それぞれにとっての効果的な活用の仕方や支援ツールとしての機能強化について検討していく必要がある。

また、文を教科書タイプか白書タイプかを判定した結果について、それが妥当かどうかの評価については、6名の学生を対象に実施したものの、その検証が不十分であり、本研究期間内では報告できなかったため、今後も継続して検証を実施したい。

参考文献

又平恵美子、竹内純人、大野博之、稲積宏誠（2010）「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　pp. 16-20

柴崎秀子、原信一郎（2010）「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

佐藤理史（2011）「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

李在鎬（2016）「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

コーパス開発センター、http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html参照日2017/1/4

国立国語研究所（2009）『教育基本語彙の基本的研究　増補改訂版』明治書院

Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto（2004）“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

工藤拓、松本裕治（2002）「チャンキングの段階適用による係り受け解析」情報処理学会論文誌、Vol 43、No. 6、pp. 1834−1842

阿部純一、桃内佳雄、金子康朗、李光五（1994）『人間の言語情報処理―言語理解の認知科学』サイエンス社

Stefan Rüping (2000): mySVM-Manual, University of Dortmund, Lehrstuhl Informatik 8, http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/. 参照日2018/1/24

益岡隆志、田窪行則（1992）『基礎日本語文法・改訂版』くろしお出版