

日本語文章作成支援とテキストマイニングの取組

Support for Academic Japanese Writing and Approach to Text Mining

<プロジェクト・メンバー>

稲積宏誠^{†1} 大野博之^{†1} 又平恵美子^{†1} 多田知子^{†1}

Hiroshige Inazumi^{†1}, Hiroyuki Oono^{†1}, Emiko Matahira^{†1}, Tomoko Tada^{†1}

^{†1} 青山学院大学 社会情報学部

School of Social Informatics, Aoyama Gakuin University

キーワード：ライティング，教科書コーパス，白書コーパス，難易度

Keywords: academic writing, school textbook corpus, white papers corpus, text difficulty

1. 取り組みの概要

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。

本社会情報学部では、学部設置時から、上記の趣旨に基づいて正規科目「コミュニケーション基礎」をスタートさせた。現在では、必須の情報スキルである表計算ソフトの活用と結びつけることで、レポートライティングの総合課題を最終ゴールとして授業が設計されている。

本プロジェクトは、一貫して日本語文章力を高めるための取り組みを、母語話者だけではなく非母語話者への展開も視野に入れて大きく3つのカテゴリで進めている。

一つは、コミュニケーション基礎の科目の下支えを続けていくことである。現在は独自テキストを作成

し、継続的に内容の見直し、練習問題の追加・修正を繰り返している。また、学部教員にもデータを開放し、ゼミ指導や卒業論文作成時にも活用してもらえる環境を作っている。

いま一つは、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である。校正上の形式チェックについては、すでに公開し、一定の評価を得てきた。現在は、文のわかりやすさに着目し、わかりにくい文の特徴を機械的にとらえることによってより高度な「推敲」の指導支援につなげることを目指している。

最後は、日本語を中心としたテキストマイニングの取り組みである。これは、ライティングとは直接関係はしないが、いわゆるテキスト情報を対象とした分析技術を磨く取り組みともいえる。すなわち、文章読解のような定性的な取り組みと、自然言語処理に基づく定量評価可能な取り組みをどのように組み合わせしていくかということをテーマとして扱っていくということである。

本稿では、文のわかりやすさに着目した取り組みを紹介する。現段階では、ライティングをはじめとした教育現場へのフィードバックをはじめとして実践レベルには至っていないが、今後の展望も交えて現状を報告するものである。

2. わかりやすい文とは どのようなものか

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、既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[1]。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[2]。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[3]。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。

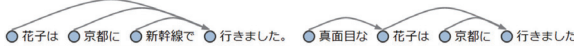
一方、本取り組みでは、個々の文を対象にすることも踏まえ、自然言語処理技術を活用した新しい学習支援・指導支援方法の提案を目的としたものであり、それを実現するためのツールの開発も視野に入れている。

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。ここで紹介するのは、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を活用した次のような取り組みである。

①語彙の難易度の決定ルール作成、②文の節構造の判定と図示化、③語彙と構造の特徴量にもとづく文の特徴を表すモデルの作成を行ったのち、これらの情報を用いて、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。

語彙の難易度を定めるために、2011年に発行された「教科書コーパス語彙表」・「BCCWJ 主要コーパス語彙表」と1984年に発行された「新阪本教育基本語彙」の3つを基準データとして用いる。その結果、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。また、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、「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。

次に特徴量の抽出を行う。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」などである。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」などである。文構造の特徴量の算出例を図1に示す。



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1 係り受けに関する特徴量算出結果例

教科書タイプと白書タイプの分類は、機械学習アルゴリズムの一つであるサポートベクターマシン(以後SVM)を用いる。特に各特徴量の重みを評価するために線形SVMによる学習を行った。実験にあたっては、それぞれ文の構造以外の影響を除くために、語彙に関

する条件をそろえることで、表 1 に示すように利用する文を絞り込んでいる。

表 1 学習に用いた教科書・白書コーパスの文数

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

線形 SVM の学習結果から、各特徴量の分類への影響力は表 2 に示すことができる。ここから、白書のおおよその特徴を推測すると、「接続助詞を多用した文の構成」「修飾語と被修飾語が離れる傾向」「連体節などの長い前置修飾の多用」「独立性の強い 2 文からなる複文構造」などであることがわかる。

表 2：教科書タイプと白書タイプに分類する指標

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
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

同一語彙レベル、同一文節数の教科書タイプの文と白書タイプの文の構造上の特徴を図 2 に示す。教科書タイプ「多量の水をかけて十分に洗い、中和剤で処理した後、さらに水でよく洗う」は、修飾・被修飾の関係が単純であること、白書タイプ「住民の不安に対し適切な対応をとるため、健康診断後、健康相談を実施することが適当とした」は、修飾・被修飾の関係が複雑であることがわかる。

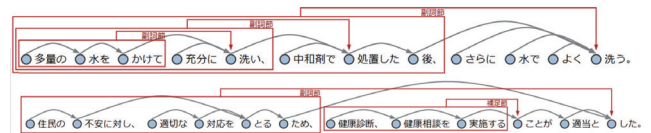


図 2 教科書タイプと白書タイプの文例

3. まとめと今後の課題

今回は、本プロジェクトの一部として、わかりやすい文の要件を探る取り組みについて紹介した。現在、ここで得られた指標の精査、文の難易度評価の客観的基準の一つとして、機械翻訳精度の導入、長文に自動分割によるわかりやすさの向上などの取り組みを行っている。

参考文献

- [1] 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
- [2] 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
- [3] 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