様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　５　月　３０　日現在

機関番号：　１３９０１
研究種目：　基盤研究(B)（一般）
研究期間：　2012～2015
課題番号：　２４３０００５２

研究課題名（和文）平易な日本語表現への工学的アプローチ

研究課題名（英文）An Engineering Approach to Plain Japanese

研究代表者
佐藤　理史（Sato, Satoshi）

名古屋大学・工学（系）研究科（研究院）・教授

研究者番号：３０２０５９１８

交付決定額（研究期間全体）：（直接経費）　　13,200,000円

研究成果の概要（和文）：本研究では、難易度分布に基づくテキストの相対難易度推定システムを実現し、現代日本語書き言葉均衡コーパスの全サンプルに９段階の難易度を付与した。これを利用して、基本語彙を選定するための各種調査を行った。これと並行して、格助詞・副助詞を含む文節、文末文節、節末文節の調査と整理を行い、日本語の文節の大半のパターンを明らかにした。制限言語の一つとして、辞書の定義文を完全な文形式で記述する形式を設計し、実際に160語に対して定義文を試作した。

研究成果の概要（英文）：We developed a system that estimates readability of Japanese text based on readability distribution. By using this system, we assigned one of nine readability score to every sample in Balanced Corpus of Contemporary Written Japanese (BCCWJ), and analyzed word distribution of BCCWJ for selecting basic vocabulary of Japanese language. In addition, we made a list of bunsetsu patterns with case or adverbial particles, sentence-final bunsetsu patterns, and clause-final bunsetu patterns. As a case study of controlled language, we designed a style of full-sentence definition (FSD) of Japanese words, and defined 160 words by using the FSD.

研究分野：　自然言語処理

キーワード：　テキストの難易度　基本語彙　節境界　文節パターン　辞書定義文

１．研究開始当初の背景

(1)ほとんどの知的活動は、言語活動、すなわち、テキストからの情報の取得と情報伝達のためのテキスト生産ぬきでは成り立たない。言語活動の支援と促進は、知的活動の活性化につながると考えられる。
(2) 社会の多様化が進み、情報を平易な日本語で記述することが、これまで以上に強く求められてきている。このためには、ゆるやかな表現の統制が必要である。
(3) 言語表現の基本構成要素は語であり、表現の統制は、使用する語彙の制限として実現できる。しかしながら、膠着語である日本語は、語の単位がそれほど明確ではない。日本語で最も安定している基本単位は文節である。そのため、平易な日本語の基盤の候補は、基本語彙表と基本文節パターンとなる。
(4) 平易な日本語の作成を支援するためには、テキスト作成時に書き手を支援するテキスト自動評価ツールが不可欠である。
(5) 表現の統制が強く望まれる対象は、辞書の定義文である。知らない語を調べる際に利用される辞書においては、その定義文が平易かつ分かりやすいことが必須の条件である。

２．研究の目的

本研究では、「平易な日本語表現」の辞書となる日本語表現バンクを編纂し、これに基づくテキストの難易度評価ツールを実現することを目標とする。前者は、内容語の基本語を示した基本語彙表に加え、文の構成要素となる文節の基本パターンを示した基本文節パターン集を新たに導入する。これは、いわば日本語の骨格を定義するものであり、各種用途の制限言語の設計の基盤となる。一方、後者は、平易なテキストの作成を支援するツールであり、テキスト全体を対象としたマクロな難易度評価に加え、テキスト中の難解部分を自動同定するミクロ難易度評価を実現する。さらに、本研究では、辞書定義文を基本語彙と基本文型のみで記述する方法（制限言語）を設計し、制限言語の効果を実証する。

３．研究の方法

(1) 研究は、およそ図１に示すような形で実施する。

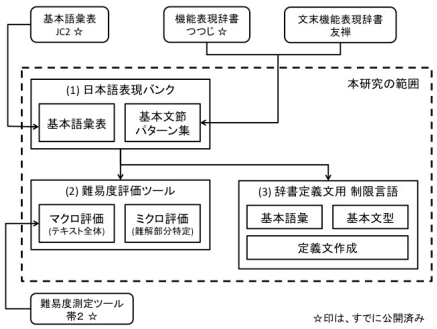

図１：研究の全体像

(2) 国立国語研究所が編纂した「現代日本語書き言葉均衡コーパス(BCCWJ)」を分析し、基本語彙表に収録すべき語の選定を進める。
(3) これと並行して、これまでの言語学の成果や BCCWJ の分析に基づき、日本語の機能表現に対する整理を進め、基本文節パターン集を作成する。
(3) 語彙の難易度等を利用したテキスト難易度推定方法について検討する。
(4) コリンズ・コウビルド辞書で採用されている full sentence definition (FSD)による語の定義法を参考に、日本語の語を定義・記述するための FSD を設計する。

４．研究成果

(1) 現代日本語書き言葉均衡コーパス(BCCWJ)のすべてのサンプルに対して、9 段階のテキストの難易度（相対難易度）を付与した。

まず、正式リリース版の BCCWJ を用いて、難易度の分布が正規分布になるよう難易度モデルをブートストラップ法により作成し、このモデルを用いた難易度推定結果が、人間による難易度推定結果の平均と、比較的よく一致することを確認した。この難易度モデルを用いて付与した難易度の分布を図１（固定長サンプル）と図２（可変長サンプル）に示す。

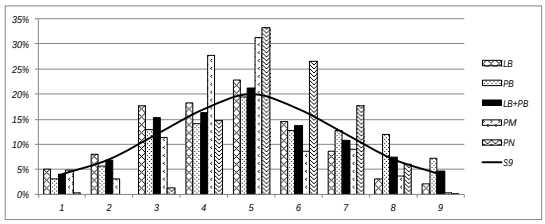

図１：固定長サンプルの難易度分布

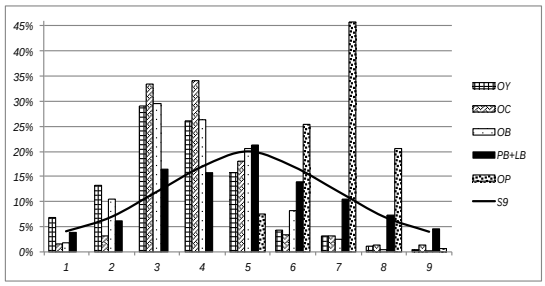

図２：可変長サンプルの難易度分布

図１に示すように、雑誌(PM)や新聞(PN)は、書籍(PB と LB)より難易度の分散が小さい。これは、編集によってテキストの難易度がコントロールされているためと考えられる。これに対して、書籍(PB と LB)の難易度は、幅広く分布する。図 2 に示すように、Yahoo!知恵袋(OC)や Yahoo!ブログ(OY)は比較的平易で難易度は 3 または 4 のものが多い。また、ベストセラー(OB)は書籍全体(PB+LB)と比較して、かなりやさしい方に偏る。このように、それぞれのレジスタ（文書の種類）毎に、テキストの難易度の特徴を把握することが可

能となった。
(2) (1)で作成した難易度付き BCCWJ を用いて、テキストの難易度と語の分布に対する一連の調査を行ない、我々が直感的に考えている仮説のうち、何が成り立ち、何が成り立たないかを明らかにした。主要な結果は、次のとおりである。
① テキストの難易度が高いほど、複合語の割合が増加する。
② テキストの難易度が高いほど、漢語の割合が増加し、和語の割合が減少する。
③ テキストの難易度が高いほど、名詞の割合が増加し、動詞の割合が減少する。
④ テキストの難易度が高くなるほど、単位長さ当たりに出現する語彙数は必ずしも多くなるわけではない。
⑤ テキストの難易度が高くなるほど、単位長さ当たりに出現する複合語のバリエーションは必ずしも多くなるわけではない。ただし、ある難易度のテキスト群全体において観察される複合語のバリエーションは、難易度が高くなるほど多くなる。
これらの結果とともに、図３に示す累計カバー率のグラフを得た。日本語では、語の出現数の 85%をカバーするためには、短単位(SUW)では 3108 語、長単位(LUW)では 8396 語が必要となる。これらの結果を利用して、テキスト中の難しい文を見つける方法を試作した。

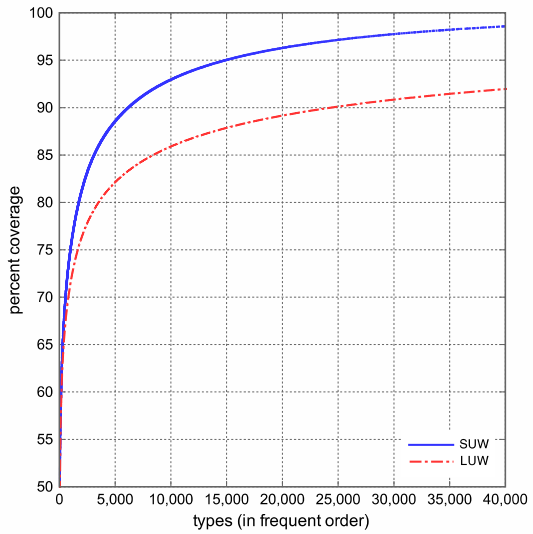

図３：日本語の語の累積カバー率

(3) 基本文節パターン集を作るための準備として、BCCWJ を用いて、格助詞・副助詞の連続出現パターンの調査・整理を行い、文節パターンの形式を明らかにした。その結果が図４である。この図に示すように、格助詞・副助詞は、A群からD群の４種類のグループに分類することができ、かつ、それらに出現順序の制約があることが示された。このモデルにより、体言（名詞）を中心とした文節パターンの大半が明らかになった。

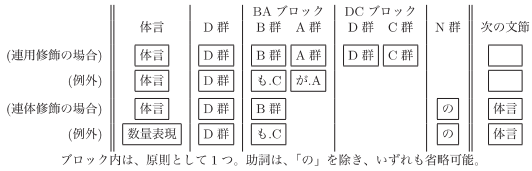

図４：体言文節のパターン

これと並行して、文末機能表現の調査・整理を行い、意味ラベルを付与した文末機能表現シソーラス(45,948 エントリ)を作成した。これは、文末述語文節のパターンの整理に相当する。さらに、このシソーラスを用いた述語正規化システムを試作した。

(4) 残った文節の大半は、文中の述語文節、すなわち、節末文節である。節末機能表現という概念を導入して節末文節の整理を進め、それに基づいて節境界検出システム Rainbow を作成した。このシステムでは、節境界は必ず文節境界でもあるという点に着目し、文中の文節境界を認定したのち、それらが節末境界であるかどうかを判定するという２段階判定方式を採用した。これにより、節末境界認定ルールが簡潔に記述することが可能となった。Rainbow は節末境界認定と同時に、節のタイプを 57 種類のいずれかに分類する。実験により、これまでに提案されてきた CBAP と遜色のない推定精度が得られることを確かめた。

(5) 語の定義を完全な文として記述する日本語 FSD の基本設計を行い、FSD の作成ガイドライン（マニュアル）と、160 項目の記述例を作成した。
FSD の基本コンセプトは、以下のとおりである。
① FSD は前件部と後件部から構成する。
② 前件部では、対象となっている語の表現形式・用法・語義の、典型的な文型を示す。
③ 後件部では、前件部と意味的に等価な言い換え表現を平易な形で示す。
以下に例を示す。
【心強い：こころづよい】イ形容詞
［〈何か〉があって・〈誰か〉がいて］《心強い》とは、［頼ることができる｛〈何か〉があって・〈誰か〉がいて｝］安心だと｛思う・感じる｝ということ。
《心強い》〈味方〉とは、頼りになる〈味方〉のこと。
［〈誰か〉が〈何か〉を］《心強く》〈思う・感じる〉とは、［〈誰か〉が〈何か〉を］頼もしく〈思う・感じる〉ということ。

(6) 研究開始当初は予定していなかったが、日本語テキストの含意認識の研究を実施し、RITE2 において、好成績(BC タスクでは参加 42 システム中 3 位、MC タスクでは参加 21 システム中 1 位)を収めた。

日本語のテキストは、表意文字である漢字をかなり含むため、文字の一致度や文字 bigram の一致度を利用して含意関係を推定する方法が、比較的うまく機能する。このことを公開性能評価型ワークショップ RITE2 に参加して実証した。

さらに、この技術を利用して、大学入試問題のセンター試験「国語」現代文の評論読解問題を解くシステムを作成した。このシステムの基本戦略は、本文のある部分を照合領域として切り出し、それともっともよく一致する（文字のオーバーラップ率がもっとも高い）選択肢を選ぶというものである。あらかじめ５つの選択肢のうちの１つを除外するなどの工夫を追加することにより、センター試験「国語」評論読解問題の過去問の約半分に対して、正しい選択肢を選択できることを示した。さらに、(4)で作成した節境界検出システム Rainbow を用いて、本文と選択肢をそれぞれ節に分割し、節単位の照合に基づいて本文と最もよく一致する選択肢を選ぶという節照合法により、過去問 40 問中 28 問に正解することを示した。

大学入試において、より直接的に事実判定を問う科目は歴史である。事実判定は、ある種の含意認識の拡張であり、これを対象とした公開性能評価型ワークショップ RITE-VAL に参加した。成績は、参加 30 システム中 6 位であった。このシステムを元に、センター試験「世界史」のソルバーを作成した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

① 服部昇平, 佐藤理史, 駒谷和範.
表層類似度に基づく日本語含意認識. 人工知能学会論文誌, Vol.20, No.4, pp416-426, 2014, 査読有, http://doi.org/10.1527/tjsai.29.416

② 佐藤理史, 加納隼人, 西村翔平, 駒谷和範.
表層類似度に基づくセンター試験『国語』現代文傍線部問題ソルバー, 自然言語処理, Vol.21, No.3, 2014, pp465-483, 査読有, http://doi.org/10.5715/jnlp.21.465

〔学会発表〕（計２３件）

① 服部昇平, 佐藤理史, 松崎拓也.
RITE-VAL タスクを対象とした表層類似度に基づくテキストの真偽判定.
情報処理学会第 77 回全国大会, 5Q-03, 2015.3.19, 京都大学（京都市・京都府）.

② 佐藤理史, 夏目和子
新しい日本語辞書定義文型の策定に向けて（第二報）. 第 7 回コーパス日本語学ワークショップ, 2015.3.10, 国立国語研究所（東京都・立川市）.

③ 加納隼人, 佐藤理史, 松崎拓也.
節境界検出を用いたセンター試験『国語』評論傍線部問題ソルバー. 情報処理学会自然言語研究会, NL-220-8, 2015.1.20, 九州大学（福岡県・福岡市）.

④ Shohei Hattori and Satoshi Sato.
A Surface-Similarity Based Two-Step Classifier for RITE-VAL. NTCIR-11, 2014.12.11, 国立情報学研究所（東京都・千代田区）.

⑤ 加納隼人, 佐藤理史.
日本語節境界検出プログラム Rainbow の作成と評価. 第 13 回情報科学技術フォーラム(FIT2014), E-005, 2014.9.4. 筑波大学（茨城県・つくば市）.

⑥ 刀山将大, 佐藤理史, 近藤秀, 吉田達平.
日本語の文の平均像を体現した文を探す (1) 文の特徴量の抽出. 第 13 回情報科学技術フォーラム(FIT2014), E-006, 2014.9.4, 筑波大学（茨城県・つくば市）.

⑦ 近藤秀, 佐藤理史, 刀山将大, 加納隼人.
日本語の文の平均像を体現した文を探す (2) 平均からの距離. 第 13 回情報科学技術フォーラム(FIT2014), E-007, 2014.9.4, 筑波大学（茨城県・つくば市）.

⑧ Satoshi Sato. Text Readability and Word Distribution in Japanese. The Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014.5.29, Reykjavík (Iceland).

⑨ 佐藤理史, 夏目和子.
新しい日本語辞書定義文型の策定に向けて. 第 5 回コーパス日本語学ワークショップ, 2014.3.6, 国立国語研究所（東京都・立川市）.

⑩ 佐藤理史, 加納隼人, 西村翔平.
代ゼミ模試に挑戦 2013---『国語』現代文. 情報処理学会自然言語処理研究会, NL-215 No.10, 2014.2, 国立情報学研究所（東京都・千代田区）.

⑪ 佐藤理史. テキストの難易度と語の分布.
情報処理学会自然言語研究会, NL-213 No.6, 2013.9.12, 山梨大学（山梨県・甲府市）.

⑫ 佐藤理史, 加納隼人, 西村翔平, 駒谷和範. センター試験『国語』現代文の傍線部問題を解くベースライン法. 情報処理学会自然言語研究会, NL-212 No.5, 2013.7.18, はこだて未来大学（北海道・函館市）.

⑬ Shohei Hattori and Satoshi Sato.
Team SKL's Strategy and Experience in RITE2. The 10th NTCIR Conference, 2013.6.20, 国立情報学研究所（東京都・千代田区）.

⑭ 伊藤美咲姫, 佐藤理史, 駒谷和範.
難しい日本語文の自動検出のための基礎調査. 言語処理学会第 19 回年次大会, 2013.3.15, 名古屋大学（愛知県・名古屋市）.

⑮ 松木久幸, 佐藤理史, 駒谷和範.
文末機能表現シソーラスの網羅性の検証．情報処理学会第 75 回全国大会, 2013.3.6, 東北大学（宮城県・仙台市）.

⑯ 佐藤 理史.
格助詞・副助詞類の連続出現パターン. 第 3 回コーパス日本語学ワークショップ, 2013.2.28, 国立国語研究所（東京都・立川市）.

⑰ 佐藤理史.
現代日本語書き言葉均衡コーパスに対する難易度付与. 第 2 回コーパス日本語学ワークショップ, 2012.9.7, 国立国語研究所（東京都・立川市）.

⑱ 松木久幸, 佐藤理史, 駒谷和範.
文末機能表現シソーラスと述語正規化システム. 第 2 回コーパス日本語学ワークショップ, 2012.9.7, 国立国語研究所（東京都・立川市）.

〔その他〕
ホームページ等
http://kotoba.nuee.nagoya-u.ac.jp

６．研究組織
(1)研究代表者
佐藤　理史（SATO, Satoshi）
名古屋大学・大学院工学研究科・教授
研究者番号：３０２０５９１８