# BCCWJ と学習者作文コーパスを利用した日本語作文支援
# －表記と共起に関する誤用添削プロトタイプ構築－

八木　豊　　（株式会社ピコラボ）†
ホドシチェク・ボル（東京工業大学）
仁科　喜久子　　（東京工業大学）

# Japanese Writing Support System Using the BCCWJ and Learners' Corpus
# － Error Correction Prototype for Misspelling and Misuse of Co-occurrence Representation －

Yutaka YAGI (Picolab Co., Ltd.)
Bor Hodošček (Tokyo Institute of Technology)
Kikuko NISHINA (Tokyo Institute of Technology)

## １．はじめに

近年、BCCWJ に代表される大規模な日本語コーパスが利用可能になったことや、形態素解析器や係り受け解析器といった自然言語処理基盤のソフトウェアがより身近なものになったことで、それらを利用した日本語作文支援システムも数多く見られる。jcorrect は、形態素解析および係り受け解析の結果を利用して、技術文章に含まれる誤りの可能性を指摘し、日本語文章の校正を補助する機能を提供している。Chantokun では、Google 日本語 n-gram や日本語ウェブコーパスといった大量のデータから収集した統計情報を利用して、格助詞誤りをチェックする機能を提供している。

仁科らの日本語作文支援システム「なつめ」では、BCCWJ を含む日本語のコーパスから大量の共起情報を収集し、日本語学習者がそれらを効果的に閲覧できる環境を提供することで作文支援における一定の成果をあげている（仁科他(2011) ）。しかしながら、使用しているコーパスは基本的に正しい日本語として収集したもので、誤用に関する情報は含んでいない。我々は、学習者作文コーパスを利用することで、現行の「なつめ」とは異なる観点からの日本語作文支援機能を実現することを目的として、日本語学習者が書いた作文を対象に誤用タグの付与および誤用の分析を進めてきた（曹他(2010)、八木他(2011b) ）。これらの内容に基づいて、学習者の作文に含まれる誤用を検出・特定し訂正例を提示する誤用添削システムを開発中であり、将来的には「なつめ」の共起表現検索および例文表示機能と組み合わせて一つの日本語作文支援システムとすることを検討している（図　１）。

本稿では、正用データとして BCCWJ および「なつめ」プロジェクトで独自に収集したコーパスを利用し、我々がこれまでに構築した学習者作文コーパスを誤用データとして利用した日本語作文支援機能の中から、学習者の作文に含まれる表記の誤り訂正および共起表現の誤り訂正に焦点を当てて報告する。

† yagi@picolab.jp

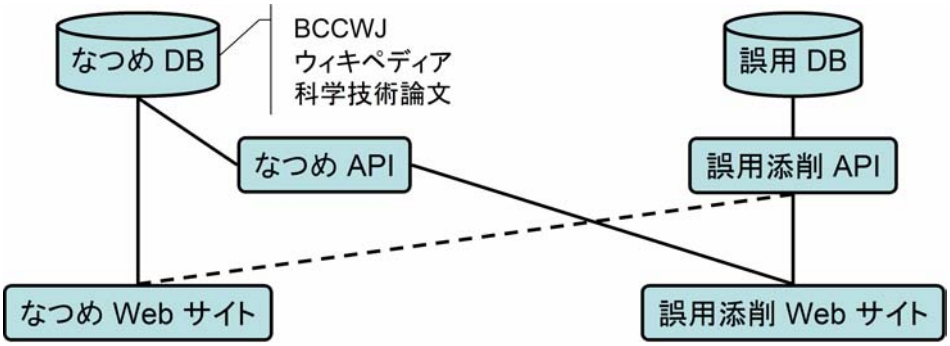


**図 1 「なつめ」プロジェクト**

## ２．利用したデータ

表記の誤り訂正および共起表現の誤り訂正において利用したデータについて、正用データと誤用データに分けて説明する。

### ２．１．正用データ

正用データとしては、BCCWJ に収録されている各種サブコーパスに加えて、ウィキペディアおよび、国立国語研究所の支援により「なつめ」プロジェクトで独自に収集している科学技術論文のデータを利用した。我々は、大学あるいは大学院に在籍する日本語学習者が書くレポートや論文を本システムによる当面の作文支援対象と位置付けており、科学技術論文のデータは、作文支援対象に近い文体のデータを拡充するものである。

「なつめ」では、名詞が格助詞を介して動詞と係り受け関係にあるもの（以下、「名詞＋格助詞＋動詞」と表記する）などをこれらのコーパスから共起表現として抽出しており、本システムでも「なつめ」に搭載されている共起表現データを利用する。表 1 にコーパスごとの共起表現数を示す。

**表 1 「なつめ」搭載のコーパス（単位：千）**

| コーパス名 | 述べ共起表現数 | 異なり共起表現数 | 文字数 |
|---|---|---|---|
| BCCWJ： | | | |
| 書籍 | 2,955 | 1,608 | 53,801 |
| Yahoo!知恵袋 | 427 | 260 | 9,763 |
| 国会会議録 | 422 | 202 | 8,712 |
| 検定教科書 | 99 | 69 | 1,819 |
| 白書 | 410 | 172 | 8,444 |
| Yahoo!ブログ | 195 | 144 | 5,246 |
| 雑誌 | 21 | 18 | 456 |
| 新聞 | 62 | 51 | 1,188 |
| ウィキペディア | 18,550 | 7,022 | 372,901 |
| 科学技術論文 | 340 | 174 | 6,108 |
| 計 | 23,482 | 8,711 | 468,439 |

## ２．２．誤用データ

誤用データとしては、我々がこれまでに構築した学習者作文コーパスに含まれる誤用タグ付き作文データを利用した。学習者作文コーパスは、複数の日本語教師の協力のもとに、大学あるいは大学院に在籍する日本語学習者が日本語の授業で書いた作文を収集したもので、作文の主な内容は、日本語の授業の中で設定した特定のテーマについてのレポートである。これらの作文に対して日本語教師による添削を行い、誤用箇所ごとに「表記の誤り」や「語の共起（コロケーション）の誤り」など独自に定義した誤用種別および誤用の訂正例をタグ付けしたものが「誤用タグ付き作文データ」である。また、学習者作文コーパスは、誤用タグ付き作文データの他に、作文を行った日本語学習者の情報として、国籍、母語、性別、日本語レベル、学習時間なども保持している。

現在も汎用アノテーションツール Slate（徳永他(2010) ）を用いて誤用タグの付与を実施しているが分析の途中であるため、本稿では、試験的な誤用タグの付与を行った以前のデータを使用した。164 人の日本語学習者から 261 作文（総文数 5,600 文）を収集し、そのうちのおよそ 3,500 文に対して誤用タグの付与を行ったものである。

## ３．誤用添削

誤用添削処理は、まず、学習者作文を CaboCha+UniDic で形態素解析、係り受け解析し、その解析結果および前述の正用データと誤用データに基づいて大まかに以下の流れに沿って行う。

(1)　誤用判定対象箇所の特定
(2)　誤用か否かの判定
(3)　訂正候補の提示

本稿では、「表記の誤り訂正」、「共起表現の誤り訂正」を独立して行った。共起表現の誤りは表記の誤りを含んでいる場合もあるため、本来は双方を関連付けて行うべきところではあるが、現時点でそこまでの連携はとれておらず今後の課題である。以降では、「表記の誤り訂正」、「共起表現の誤り訂正」それぞれについて、上記、誤用添削処理の流れにしたがって説明する。

## ３．１．表記の誤り訂正

(1) 誤用判定対象箇所の特定

形態素解析の結果、未知語とされた文字列を誤用判定の対象箇所とする。

(2) 誤用か否かの判定

八木他(2011a)では、学習者が表記誤りをした場合の多くは編集距離が 2 以内のところに訂正例の文字列が含まれていることを示した。そのことに基づいて、(1)で特定した対象箇所の文字列から編集距離が 2 以内の文字列リストに展開し、その中から正用データに単語として出現している文字列のみを取り出して訂正候補リストとした。訂正候補リストに 1 件以上の単語が含まれている場合に対象箇所を誤用であると判定する。

(3) 訂正候補の提示

正用データから抽出した単語の頻度情報および誤用データから抽出した編集操作の頻度情報に基づいて訂正候補リストに含まれる単語を順位付けし、上位のものを訂正候補として提示する。

### ３．２．共起表現の誤り訂正

(1) 誤用判定対象箇所の特定

係り受け解析の結果から特定の係り受けパターンを抽出して対象箇所とする。本稿では、「名詞＋格助詞＋動詞」、「名詞＋格助詞＋形容詞」、「形容詞＋名詞」を対象の係り受けパターンとして対象箇所を抽出した。

(2) 誤用か否かの判定

下記(a)、(b)の二通りの判定を試みる。

(a)　誤用パターンとの一致

(b)　レジスターの妥当性確認

まず(a)では、対象箇所が誤用データに含まれる実際の誤用例あるいは誤用パターンに一致するか否かを確認し、一致した場合に誤用であると判定する。ここでいう誤用パターンとは、実際の誤用例から日本語 WordNet（Bond 他(2009) ）および正用データの頻度情報に基づいて拡張したものである（八木他(2011a) ）。本稿では、誤用データから抽出した「名詞＋格助詞＋動詞」、「名詞＋格助詞＋形容詞」、「形容詞＋名詞」の誤用例 117 個とそこから拡張した誤用パターンを使用した。

次に(b)では、使用される表現や記述内容が、作文支援対象として位置付けているレポートや論文に最も近いと思われる科学技術論文および BCCWJ の白書を準正用データ、口語が含まれておりレポートや論文とは最も遠いと思われる BCCWJ の Yahoo!知恵袋、Yahoo!ブログ、国会会議録を準誤用データとして、対象箇所の共起表現の出現分布を元にカイ二乗検定を行い、準誤用データにおける出現が有意に多い場合にその共起表現はレポートや論文のレジスターとしてふさわしくないものとして誤用であると判定する。ホドシチェク(2011)では、人手でレジスターの誤用であると判定したもののおよそ 8 割がこの手法で自動的に判定可能であることを評価実験により明らかにしている。

(3) 訂正候補の提示

上記(a)の誤用パターンの一致に該当した場合は、誤用データに記載されている訂正例を訂正候補として提示する。上記(b)のレジスターの妥当性確認でレジスターとしてふさわしくないとなった場合は、正用データに含まれるコーパスから準正用データ、準誤用データに分類して利用しており、訂正例を含んでいないため訂正候補の提示をなしとした。

## ４．適用実験および実験結果の考察

学習者作文コーパスに含まれる作文データのうち、誤用タグを付与していない 36 作文（476 文）に対して誤用添削処理を適用する小規模な実験を行った。

### ４．１．表記の誤り訂正結果

表記誤りの訂正では、異なりで 17 語（延べ 26 語）を誤用判定対象箇所として特定し、うち 15 語を誤用であると判定した。誤用であると判定したものに対して適切な訂正候補を提示できたのは、「フットサール（→フットサル)」、「アシア（→アジア)」、「ビビムバ（→ビビンバ)」など 6 割ほどであった。しかし、誤用判定対象箇所の特定で取り漏らしている表記誤りも多く存在する。まずはこの点について改善が必要である。特に今回は、形態素解析で未知語とされた文字列を誤用判定の対象箇所としたのみであったが、例えば「うとんを食べる」のような文では、「うとん」の部分が「う（感動詞)」＋「とん（飛ぶの連用形撥音便)」として形態素解析され未知語にはならないため、誤用判定対象箇所から漏れて

表 2 レジスターの妥当性確認の結果

| 判定結果 | 名詞＋格助詞＋動詞 | 名詞＋格助詞＋形容詞 | 形容詞＋名詞 |
|---|---|---|---|
| 共起データなし | 598 | 16 | 11 |
| 誤用 | 41 | 6 | 17 |
| 判定不可（有意差なし） | 468 | 14 | 26 |
| 正用 | 3 | 1 | 2 |
| 計 | 1,110 | 37 | 56 |

しまう。このような場合に対応するために、正用データに現れにくい品詞の並びになっている箇所や、レポートや論文では使用することの少ない感動詞の周りなど、レジスターに特化した形での特定方法を検討している。

### ４．２．共起表現の誤り訂正結果

誤用パターンとの一致では、「私＋が＋思う」をレポートらしく「私＋が＋考える」に訂正するなど、マッチしたものは数件しかなかった。

レジスターの妥当性確認では、そのほとんどが「名詞＋格助詞＋動詞」の係り受けパターンではあるが、1,203 件の共起表現が妥当性確認の対象となった。係り受けパターンごとの判定結果を表 2に示す。判定結果の大半は共起データが全くないかあるいは、準正用データと準誤用データとの間で出現分布に有意な差がみられず判定不可となったものであったが、全体の 5%ほどをレジスターとしてふさわしくない誤用であると判定することができた。

正しく誤用であると判定できた共起表現としては以下のようなものが挙げられる。こうした表現は日本語として誤っているわけではないがレポートや論文の場合には別の表現に書き換えたほうがよく、まさにレジスターに関する誤用であるといえる。

- ことがある
- 問題が起きる
- 結論を出す
- 一緒にする
- いい経験

反対に誤って誤用であると判定してしまった共起表現としては以下のようなものが挙げられる。判定の際、科学技術論文および BCCWJ の白書を準正用データとして使用したが、その中にこういった話題に関する文章が少なかったことが、誤用であると判定された要因として考えられる。こうした表現はレポートや論文のテーマによって通常使用するものであるので、準正用データとして適切なコーパスを選択し、それが十分に大きいものであれば出現する可能性があると思われる。

- 子供がいる
- 仕事をする
- 大学に行く

### ５．まとめ

本稿では、BCCWJ を含む正用データと誤用データを利用して表記の誤り訂正および共起

表現の誤り訂正を行う誤用添削処理を提案し、提案手法を用いた実験結果を報告した。

レジスターの妥当性確認を除く誤用添削処理では正用データ全体をそのまま利用したが、作文支援対象と位置付けているレポートや論文に対してより適切な作文支援をするためには、レジスターの妥当性確認と同様に正用データに含まれるコーパスを内容に応じて使い分けることが望ましい。一方で、誤用データを利用した誤用パターンとの一致ではマッチしたものが数件と少なく、誤用データの質量ともに強化するために引き続き誤用タグの付与を実施していく予定である。

また、こうした誤用添削の結果を学習者に対して効果的に提示するためのユーザインタフェイスを構築する必要がある。

## 文献


曹紅荃、黒田史彦、八木豊、鈴木泰山、仁科喜久子(2010)「学習者作文支援システムのための誤用データベース作成－動詞の誤用分析を中心に－」世界日語教育大会論文集, pp.1571-1-1571-9.

徳永健伸、Dain Kaplan、飯田龍(2010)「Slate - A multi-purpose annotation tool」情報処理学会自然言語処理研究会報告, 情報処理学会, NL-199, 19.

仁科喜久子、村岡貴子、因京子、Joyce Terence Andrew、鎌田美千子、阿辺川武(2011)「バランス・コーパス利用による日本語作文支援システム『なつめ』の構築と評価」特定領域研究日本語コーパス平成 22 年度公開ワークショップ（研究成果報告会）予稿集, pp.215-224.

Francis Bond, Hitoshi Isahara, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki. (2009)「Extending the Japanese WordNet」言語処理学会第 15 回年次大会発表論文集, 言語処理学会.

ホドシチェク・ボル、仁科喜久子(2011)「作文支援システムにおけるレジスターの扱い」世界日本語教育研究大会　異文化コミュニケーションのための日本語教育 2, pp.522-523.

八木豊、鈴木泰山、仁科喜久子(2011a)「BCCWJ と誤用コーパスを利用した日本語作文支援に関する一考察」特定領域研究日本語コーパス平成 22 年度公開ワークショップ（研究成果報告会）予稿集, pp.119-124.

八木豊、鈴木泰山、仁科喜久子(2011b)「学習者作文コーパスの構築および BCCWJ と併用した日本語作文支援」現代日本語書き言葉均衡コーパス（完成記念講演会）予稿集, pp.119-124


## 関連 URL

日本語作文支援システム「なつめ」：http://hinoki.ryu.titech.ac.jp/

jcorrect を利用した技術文章校正のヒント：http://www.ispl.jp/~oosaki/research/tips-jcorrect/

Chantokun -統計的日本語校正-：http://cl.naist.jp/chantokun/