語の共起を効率的に検索できる日本語作文支援システム「なつめ」の紹介

阿辺川 武[†]　　ホドシチェク・ボル[‡]　　仁科喜久子[‡]

[†]国立情報学研究所
[‡]東京工業大学

1　はじめに

第2言語を学習する学習者にとって，ある単語に対して共起する語を適切に選択することは難しい問題である．特に日本語では，同じ意味を表す語でも共起する語が異なったり，同じ語でも状況に応じて共起する語が変ってくる．また中級者以上の学習者でも，ある語に対して意味的に共起する語を想起することはできるが，執筆するジャンルに応じて適切な語を選択することはまだまだ難しい．

そこで我々は，日本語を学ぶ学習者を対象とし，ジャンルに応じて適切な共起語を検索する可能にした作文支援システム「なつめ」を開発している.「なつめ」には，学習者の入力した語に対して共起する語を頻度順に表示する機能を軸に，共起対の頻度をジャンル別に表示する機能，複数の語の共起頻度を一度に比較する機能などがある．

共起語を検索するツールは既にいくつか存在する．コーパス検索ツール Sketch Engine[2] には，共起語を表示する機能があり，2語の入力語に対して共起語の分布を比較する機能 (Sketch Difference) も備わっている．また，赤瀬川による LagoWordProfiler[1] では，ユーザが共起対として抽出したい文法規則を自由に記述でき，文法規則ごとにすばやく文脈を確認することができる．ただ，これらのツールは主に言語研究者や辞書作成者向けであり，外国語学習者にとって必ずしも使い勝手のよいものではなく．その共起対がどのようなジャンルで主に使われているのか，共起対が存在しない場合の代替表現の提示といった機能は存在しない．

我々は，外国語学習者にとって使い勝手のよいインターフェースはどうあるべきかを模索しながら，類義語検索といった言語処理的な機能を加え，作文支援システムにあるべき機能を設計してきた．本稿では最初に本システムの概要を述べ，次に各部のインターフェースについて紹介し，最後に実際に本システムを利用したユーザによる被験者実験の結果について述べる．

2　システムの概要

システムのスクリーンショットを図1に掲載する．画面上部には学習者がキーとなる語を入力するテキストボックス，共起語のソート順を指示するリストボックスが並んでいる．現在では入力語の品詞は名詞もしくは動詞となっている．画面中部では，入力語に対して共起する語が格助詞別にリスト表示されている．初期設定では頻度順である．画面下部は，共起語の頻度がジャンル別に表示される．学習者が指定した共起語だけでなく，格の異なる語や類義語の頻度も同時に表示することができる．

3　データについて

3.1　コーパス

本システムで使用しているコーパスは，国立国語研究所が構築している現代日本語書き言葉均衡コーパス[3](以下，BCCWJ)，および我々独自が収集しテキスト化した科学技術論文集，そして Wikipedia 日本語版である．BCCWJ では，複数のジャンルから構成されており，我々は内容の同一性から考慮して次の7ジャンルとしてまとめた，書籍（流通書籍，ベストセラー，生産書籍，雑誌）Yahoo!知恵袋，国会会議録，検定教科書，白書，Yahoo!ブログ，新聞である．

科学技術論文の使用は,「なつめ」が対象としているユーザは理工系の大学に通う留学生であり，科学技術論文の執筆の際に使用することを想定したが，このジャンルは BCCWJ に含まれていないからである．また Wikipedia を使用した理由は，上記2つのコーパスだけでは共起情報が十分に収集ことができず Wikipedia に含まれる膨大な文により補うためである．もちろん Wikipedia では校正が不十分である文も多く存在し，学習の妨げとなる可能性も考えられるが，その欠点を補うだけの十分な量があると考えている．表1に現在のシステムで使用している各ジャンルの共起数を示す．

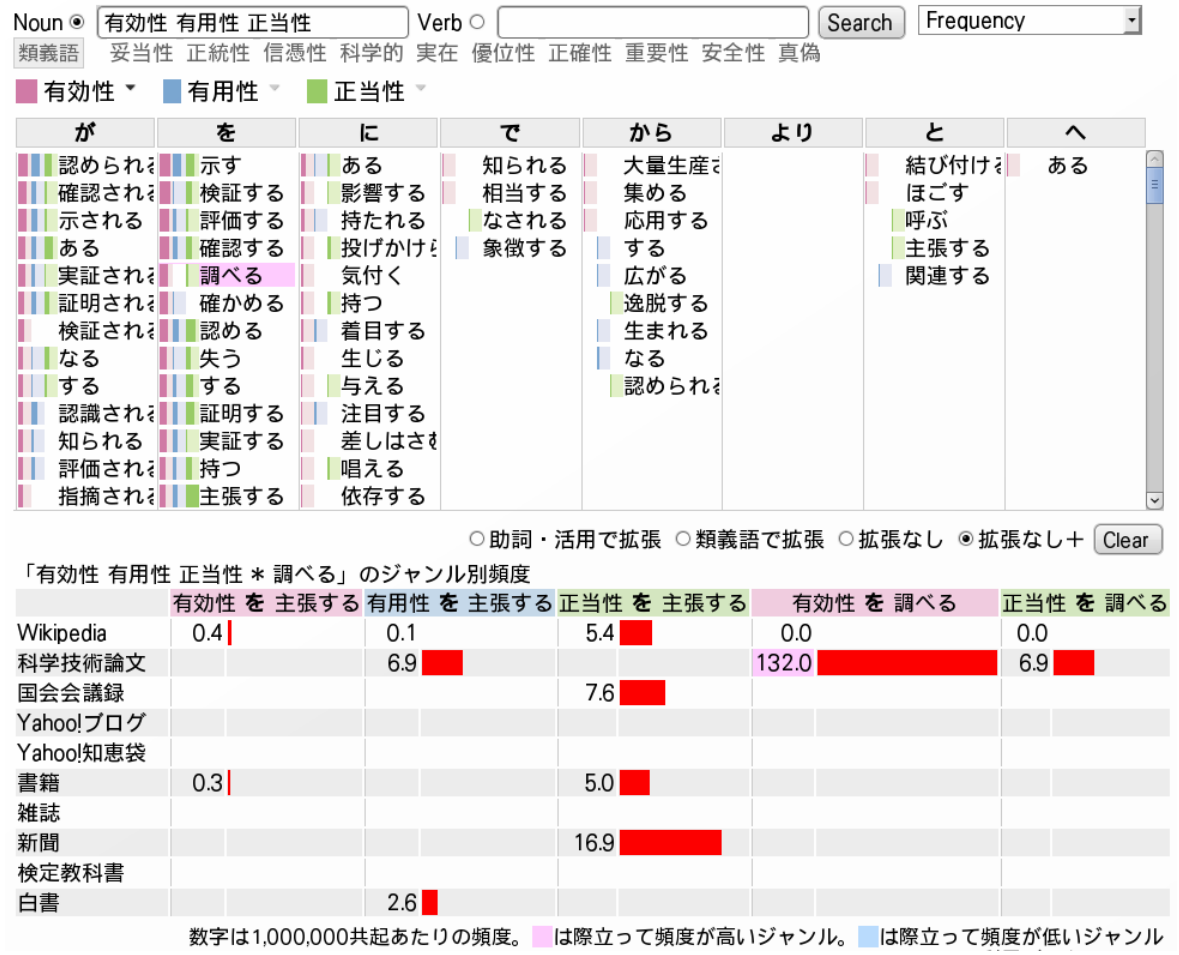

図 1: 「なつめ」の画面

表 1: ジャンル別共起数

ジャンル名	述べ共起数	異なり共起数
Wikipedia	1,753,886	16,710,442
科学技術論文	143,843	75,055
国会会議録	390,961	187,822
Yahoo!ブログ	178,779	132,125
Yahoo!知恵袋	382,135	234,275
書籍	2,749,294	1,501,642
雑誌	19,535	16,613
新聞	58,841	48,436
検定教科書	93,037	64,870
白書	380,223	161,518
総計	21,935,509	8,230,158

3.2 共起情報

前節で紹介したコーパスに対して文区切りを行ない MeCab, Cabocha を用いて係り受け解析を行なった．解析結果から格要素と動詞の係り受け関係を抽出し，「名詞句 格助詞 動詞」の三つ組として共起情報を収集した．このとき動詞については，受動・使役といった格が変わる態,「書き込む」といった複合動詞,「書き始める」のような補助動詞が後接した表現などは，格要素の格が代わる可能性があるので，すべて別々の動詞としてデータベースに格納している．

3.3 類義語の検索

名詞・動詞の共起を考えたとき，行を名詞，列を動詞 (またはその逆) とした共起行列が作成できる．本システムではこの行列を利用して，与えられた語に対する類義語を計算する機能を有している．基本的な考え方は，頻度を tf・idf などで重み付けしたベクトルに対して，コサイン類似度などの関数を用いて類似度順にランキングする．情報検索の分野では，文書・単語から生成された行列に対し，高速に関連文書・単語を検索するライブラリが数多く公開されており，これらのライブラリの持つ機能は名詞・動詞の共起行列にも応用できる．本システムでは，行列の双対性を利用し，行・列どちらからでも高速に類似度計算のできる汎用連想計算エンジン GETA[4] を使用し，類似度尺度には，情報検索でよく用いられている Okapi BM25[5] を

選択した．

4　インターフェース

作文支援シテテムを構築するにあたり重視した点は，学習者が知りたい共起対のみを検索するだけでなく，多様な視点からその共起対が正しく使えるものかを確認することである．ジャンル別に共起頻度を閲覧したり，類義語で似た意味の共起対がないかどうかなどを確認したりして，最終的な判断の根拠となる情報を提示することを目的としている．

4.1　入力語

学習者は，キーとなる入力語として複数の語を入力することができる．例えば動詞「走る」について共起語を調べたいが，似た動詞「走行する」についても見てみたい，といった要求を満たすことができる．ここでは「類義語」ボタンを押すことにより，3.3 節で説明した計算から類義語を表示し，選択できる (図 1 上部)．

4.1.1　サジェスト機能

ブラウザのテキストボックスに入力する際に，1 文字入力するごとに候補が表示される．漢字・仮名だけでなくアルファベットを用いたローマ字入力にも対応する．サジェスト候補は，現時点までに入力文字列を接頭語とする語を頻度順にソートしたときの上位 10 語である．動詞の場合，語幹を入力すれば異なる活用形を持つ語や複合動詞がサジェストされ，基本動詞以外の動詞があることに気づく．また韓国語を母語とする人のように日本語を音で覚えている学習者にとっては，漢字を入力することなく，ローマ字で音のまま入力できるので便利である．

4.2　共起語リスト

学習者が入力した入力語に対する共起語を，格助詞別に表示する (図 1 中部)．このとき共起語の左側に小さい棒グラフが表示され，一目でどの語がどのくらいの頻度で共起しているかが認識できる．この表示方法は，複数の入力語があるとき特に有効で，それぞれの入力語に対する共起頻度が簡単に比較できるとともに，共起しない場合は空白で示され，また，どの入力語を基準としてソートするかを指定でき，指定した入力語のみと共起し，他の入力語とは共起しない語というのがすぐに把握できる．

4.2.1　共起語のランキング

共起語の表示順の初期値は頻度順であり，学習者にとっては頻度順で十分であると考えている．しかし，日本語の辞書を作成する研究者などから，特徴的に共起する語が知りたいなどの要望があり，本システムでは以下の尺度によるランキングを選択することができる．Dice 係数，T スコア，Jaccard 係数，対数尤度比，χ 二乗係数，相互情報量．

4.3　ジャンル別共起頻度の比較

学習者が共起語リスト中の語をクリックすると，共起語の頻度をジャンル別に表示し，各ジャンル間の共起頻度を比較することができる (図 1 下部)．多くの共起語検索ツールでは複数の種類のコーパスがあることを想定しておらず，あるジャンルでは共起するが，別のジャンルでは共起しないといったことがわからない．一方で本システムでは，予めコーパスにジャンルのタグを振っておけば，そのタグごとに頻度をカウントすることで，それぞれのタグ内の頻度を計算できる．

表示方法は棒グラフ表示で，学習者が一目でジャンルごとの頻度の差異が視認できるように工夫している．このとき各ジャンルでは，総共起数が大幅に異なるため，頻度グラフでは 100 万共起あたりの頻度に正規化して表示している．また表示の際は χ 二乗検定を行ない，全ジャンルの平均頻度より有意に頻度が高い/低いジャンルについては，その旨を色分けして表示している．

4.3.1　共起語の比較

ジャンル別共起頻度の比較では，学習者が選択した共起語だけでなく，共起語に関連する他の語も表示できる．

1) 格助詞で拡張
   画面中部のリストで選択した「格助詞 共起語」に対し，格助詞が異なる共起対も同時に表示する．例えば「学校 へ 行く」を指定したとき「学校 に 行く」も表示され，両者の頻度を比較できる．また共起語が動詞の場合，指定した動詞に対して態の異なる動詞を同時に表示する．例えば「問題 を 解決する」に対しては,「問題 が 解決される」「問題 を 解決させる」が表示される．さらに動詞の場合は態だけでなく，可能動詞「(遊ぶ→) 遊べる」，使役動詞「(遊ぶ→) 遊ばす」も存在すれば表示する．
2) 類義語で拡張
   この機能が選択されたとき，共起語の類義語を同時に表示する．4.1 節で計算した類義語と異なり，ここでは入力語 A を考慮した上での共起語 B に対する類義語集合 b_n を提示する．類義語の計算方法は以下の通りである．最初に 4.1 節と同様に共起語 B に対する類義語集合候補 $b_i(1 < i < n)$ を計算する．次に以下の $score$ を元に，候補をランキングして提示する．

   $$score_{A,b} = Sim_{BM25}(V_B, V_b) \times Sim_{Cosine}(V_a, V_b)$$

   ここで，V_a は，入力語 A に対する類義語集合 $a_i(1 < i < m)$ を求め，これらを要素としてベ

表 2: 入力語を考慮した共起語の類義語の例

入力語	共起語	共起語の類義語
海岸 (を)	歩く	散歩する，うろつく，さまよう，歩き始める
山道 (を)	歩く	歩ける，歩きまわる，駈ける，散策する
廊下 (を)	歩く	駈ける，駆け出す，走り回る，歩き出す

クトルとしたもの．V_B, V_b は，それぞれ共起語 B，類義語候補 b_i について，共起行列から共起ベクトルを抽出したものである．入力語を考慮した共起語の類義語の例を表 2 に掲載する．

5 被験者実験

「なつめ」における共起検索の有効性について，学習者評価実験を試みた．被験者は，理系学部 1 年生，2 年生 40 名で，学部生は日本語能力試験 1 級保持者である．実験は与えられた文および文章を論文調に書き換える課題を設定した．作題文は，それぞれ 1 級から 4 級および級外までの語彙がほぼ均等になるように配置し，論文のためには書き換えが必要な項目が均等に含まれる問題セットを A，B の 2 種類準備した．

最初に被験者を 2 グループに分割し，片方のグループには問題 A を，もう片方のグループには問題 B を配布し,「できるだけ論文らしい表現になるように」と指示し筆記で解答させた．所用時間は 60 分程度であり，電子辞書の使用は許可とした．次に「なつめ」を利用して，2 グループに対して問題を交換して課した．指示は筆記テストと同様にできるだけ論文らしい文を作成するように指示した．その際，できるだけシステム上の「類義語」を参照し,「科学技術文」など論文に近いコーパスに高い頻度があるものを選択するように指示した．

採点方法は,「なつめ」で検索可能な共起語が正しく書き換えられた箇所を配点 2 点 (23 箇所)，科学技術論文のレジスターとして必要と思われる副詞，形容詞，文末モダリティなどの表現項目の問題を配点 1 点 (20 箇所)，計 66 点満点とした．

被験者 40 名について上記の評価方法によって得点を集計した．A，B それぞれのグループの得点を集計し，筆記実験のグループ別平均値の差を検定した結果，A，B 間の有意差は認められず (p 値=0.8851)，ほぼ均一レベルの学習者に対して問題文の難易度の差がないことがわかった．

表 3 には，筆記試験の得点により上位，中位，下位の 3 群に分けたときの,「なつめ」利用時における得点の増減を集計した結果を掲載する．この結果，筆記上位群よりも，筆記下位群の方が「なつめ」利用時における得点の増加点数が大きいことから，日本語能力試験 1 級合格者の中でも中位群，下位群において「なつめ」利用の効果が高いことが明らかになった．

表 3: レベル別の得点集計結果

	上位群	中位群	下位群
対象学習者	8 名	20 名	12 名
筆記 得点	58∼36	35∼16.5	14.5∼0
平均点	43.75	26.78	9.29
システム利用 得点	62∼23	54∼23	40∼15
平均点	45.88	37.36	24.57
差分範囲	19∼-16	19∼-5.5	26∼7.5
差分平均点	0.75	10.41	15.29

6 おわりに

本稿では，日本語作文支援システム「なつめ」について機能を中心に紹介した．現在「なつめ」が持つ共起情報は「名詞 格助詞 動詞」であるが，被験者からの要望が多かった「形容詞 名詞」(装定用法) や，他にも「名詞 が 形容詞」(述定用法) や「副詞 動詞」,『きっと ∼だろう』のような「副詞 モダリティ」の共起の利用についても準備している．また，共起語リストのソート機能でも触れたように，本システムは外国語学習者だけでなく，日本語研究者や辞書執筆者，しいては，はじめて論文を執筆するような日本人学生に対しても有用であると思われ，そのような人へ向けたインターフェースも設計していきたい．

参考文献

[1] 赤瀬川史朗. Wordprofiler, 2010. http://www.lagoinst.com/WordProfiler/.

[2] Kilgarriff A, Rychlý P, Smrž P, and Tugwell D. The sketch engine, 2004. In: Williams G, Vessier S, editors. Proceedings of the Eleventh EURALEX.

[3] 前川喜久雄. Kotonoha『現代日本語書き言葉均衡コーパス』の開発. 日本語の研究, Vol. 4, No. 1, pp. 82–95, 2008.

[4] 西岡真吾. 汎用連想計算エンジン GETA. コンピュータソフトウェア, Vol. 26, No. 4, pp. 87–106, 2009.

[5] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *Third Text REtrieval Conference (TREC 1994)*, pp. 109–126, 1994.