日本語学習者のためのコロケーション検索システムの開発

中溝　朋子[*1]・坂井　美恵子[*2]・金森　由美[*3]・大岩　幸太郎[*4]
Email: tomokon@yamaguchi-u.ac.jp

*1：山口大学留学生センター
*2：大分大学国際教育研究センター
*3：大分大学国際教育研究センター
*4：大分大学

◎Key Words コロケーション，コーパス，日本語学習

1. はじめに

近年、日本語学・日本語研究においても、コーパスの開発が進み、その成果を日本語学習に活用するための研究や教材作りが活発に行われるようになった。日本語のコロケーションに関しても、こうしたコーパスを活用した検索システム「NINJAL-LWP」(1)、「筑波 web コーパス」(2)、日本語作文支援システム「なつめ」(3)などが開発されている。これらのシステムでは、コーパスを用いて抽出されたコロケーションが統計指数別に例文も併せて閲覧表示可能となっているが、日本語学習者にとっては①コロケーションが漢字の読みや意味の説明などがなく一度に多数表示されるため、どのコロケーションを選択すれば良いかの判断が難しい、また②表示される例文がコーパスの原文そのままであるため、文脈の不足、漢字の読み・語彙・文法の難しさなどの点で理解が容易ではないなどの場合があり、実際の使用において難しい面もあった。

そこで筆者らは、日本語学習者を対象とした「日本語学習者のためのコロケーション検索システム 《かりん》＝よく一緒に使われることば＝」(以下、「かりん」)を開発した。「かりん」では、名詞を中心語として共起語である修飾語、および動詞とのコロケーションを検索することが可能となっている。本発表では、この検索システムの概要と、発表者らの大学で試行を行った結果についてまとめ、今後の課題について述べる。

2. 本検索システムの概要

2.1 データ

「かりん」では、旧日本語能力試験（以下、旧 JLPT）１級および２級の名詞と、それらと共起する語（修飾語と動詞）を「現代日本語書き言葉均衡コーパス」（国研 2011、以下、BCCWJ）より抽出したデータを用いている。BCCWJ は「現代日本語の書き言葉の全体像を把握するために構築したコーパス」であり、「書籍全般、雑誌全般、新聞、白書、ブログ、 ネット掲示板、教科書、法律などのジャンルにまたがって１億 430 万語のデータを格納」している(4)。本システムでは、上述の名詞、およびこれらと共起する修飾語と動詞コロケーションを正規表現によって検索・抽出し、共起頻度、および共起強度を示すダイス係数を計算、その結果を表示している。表示する範囲は、共起頻度５以上とし、ダイス係数順に表示している。これらの共起頻度とダイス係数は、画面上では数値とともに棒グラフで視覚的に示されている。

2.2 表示画面

表示画面（図１）は、日英２か国語から選択でき、トップ画面の左側「確認事項」等を読んだ後、検索が可能となる。本システムへのユーザのフィードバックは、任意のアンケートからのみ収集し、ユーザの個人情報は収集していない。

また学習者への利便性を高めるため、表示画面においていくつかの機能を付与している。以下、2.3 でこれらの機能について述べる。

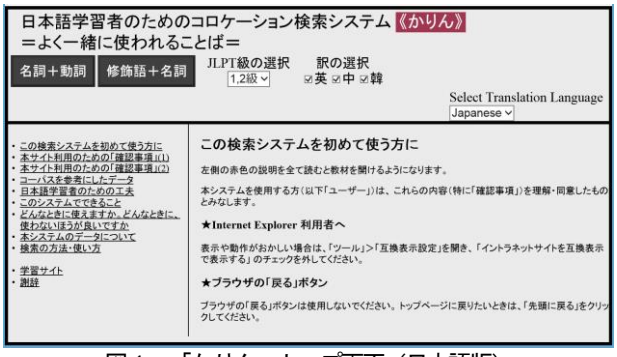

図１　「かりん」トップ画面（日本語版）

2.3 「かりん」の特徴

2.3.1 旧 JLPT 語彙レベル、および学習指標値（徳弘 2006）(5)の表示

日本語学習の参考になるよう、コロケーションを構成する語（名詞、動詞、修飾語）には、すべて旧 JLPT のレベルおよび学習指標値（徳弘 2006）を示している。徳弘（2006）が提案した学習指標値とは、新聞データにおける頻度と単語の親密度を基に計算されている。

例えば、図２の名詞「開発」を検索した結果では、図右上「開発」の右側[　]内に、名詞「開発」の[JLPT 1－学習指標値 10]という値が示されている。同様に「開発」と共起する各動詞の JLPT と学習指標値について、表中の「動詞」右側の列「動詞レベル」に「JLPT-学習指標値」の順で示している。これにより、日本語学習者の学習の目安となるようにした。

図２　名詞「開発」の検索結果画面

2.3.2　動詞の意味分類の表示

「名詞＋動詞」のコロケーションにおいては、動詞の意味分類を表中「意味」の列に記入、表示している。

この動詞分類のひとつは、村木(1991)の「機能動詞」の概念を用いている。村木(1991)の機能動詞とは「実質的な意味を名詞にあずけて，みずからはもっぱら文法的な機能をはたす」動詞のことである[6]。例えば「背広がかかっている」の「かかる」が実質的な動作の内容を表しているのに対し、「攻撃にかかる」の「かかる」は「攻撃」という名詞が表す動作の開始（始動相のアスペクト）を表しており、このような動作名詞と共起して文法的な意味（アスペクト、ヴォイス、ムード）を表す場合を機能動詞と呼ぶ。旧 JLPT1 級、2 級の名詞では、初級レベルの 3 級、4 級の名詞に比べ漢語名詞の占める割合が増え、抽象名詞や動作名詞の割合が多くなるため、このような機能動詞との共起も多くなると考えられる。

また漢語名詞は和語名詞に比べ、書き言葉で使用されることが多く、同様に書き言葉での使用が多い複合助辞と共起する場合も多い。複合助辞とは、「幾つかの語が複合して一まとまりの形で助辞的な機能を果たす表現」（松木 2011）[7]であり、特に助詞的な機能を持つものには「～を通して」「～に伴って」のように動詞を核とするものが多く見られる。

このようなことから、1 級 2 級名詞に多い漢語名詞と多くの共起が見られる機能動詞や複合助辞的用法を、動詞の実質的な意味を持つコロケーションから区別し、意味を示すことで、学習者が提示された多くのコロケーションの中から、必要とする意味のコロケーションを早く探し出せるように、動詞の分類とその表示を行った。具体的には、①村木（1991）の「機能動詞」、②複合助辞的意味、それ以外の③実質的な意味を持つ実質動詞の３つに分類し、セルには、機能的意味と複合助辞的意味に背景色を付与し、各意味を記して区別した。この表示についても、先述の名詞「開発」を検索した結果、図２で確認することができる。

2.4　ふりがなの表示

ふりがなについては、名詞や動詞、修飾語にマウスカーソルを重ねるとポップアップで表示される。

2.5　例文の表示

「名詞＋動詞」の「例文訳」の列に示されている「英中韓」などの文字にマウスカーソルを置くと、図３のように例文と「英中韓」の各文字に対応した訳がポップアップウィンドウで表示される。「名詞＋動詞」では、例文約 2,500 文を付与し、訳語も随時付与している。また例文作成には時間がかかるため、例文が未作成のコロケーションにはコロケーションの訳を順次付与している。

また「修飾語＋名詞」を選択した場合にマウスカーソルを「例文訳」列のセルに置くと、図４のように「修飾語＋名詞」の後続語の共起頻度の上位３語が示される。

図３　「名詞＋動詞」選択時の「例文訳」表示

図４　「修飾語＋名詞」選択時の「例文訳」表示

2.6　サブコーパスの表示

「コロケーションの強さ」を示す横棒グラフ（黄色）にマウスカーソルを重ねると、図 5 のようにそのコロケーションの BCCWJ のサブコーパス別の相対頻度が示され、学習者はそれぞれのコロケーションが具体的にどのような文献で使用されているかを知ることができる。サブコーパスは大きさが異なるため、このように相対頻度を計算し、グラフで視覚化している。

図５　サブコーパスの表示

3.　試行の結果

3.1　実施方法

本システム「かりん」について、2016 年 6 月 2 日に大分大学で、6 月 6 日に山口大学で試行を行った。対象者は 40 名、日本語中上級～上級の学習者で両大学合わせて中国語母語話者 17 名、韓国語母語話者 12 名、英語母語話者 5 名、タイ語 3 名、その他の母語話者 3 名（ロシア語、ドイツ語、ウクライナ語の英語以外のヨーロッパ言語母語話者各 1 名）である。調査は授業時間に任意の名詞を 3～5 語各自で決めてもらい、「修飾語＋名詞」、「名詞＋動詞」のコロケーションを検索し、その感想をシートに記入するという形で実施した。

3.2　結果と考察

調査では、(1)「かりん」は役に立ったか、(2)画面（見やすさ・美しさ）、(3)操作性（使いやすさ・わかりやすさ）、(4)例文・翻訳（長さ・わかりやすさ）について 5 段階評価で回答、さらにその理由を自由に記述してもらった。以下、これら 4 つの質問に対する 5 段階評価を図 6 に示し、設問ごとの理由（自由記述）の一部を記す。またコメント末の（　）内は学習者の母語を表し、(他) は英語以外のヨーロッパ言語を表す。

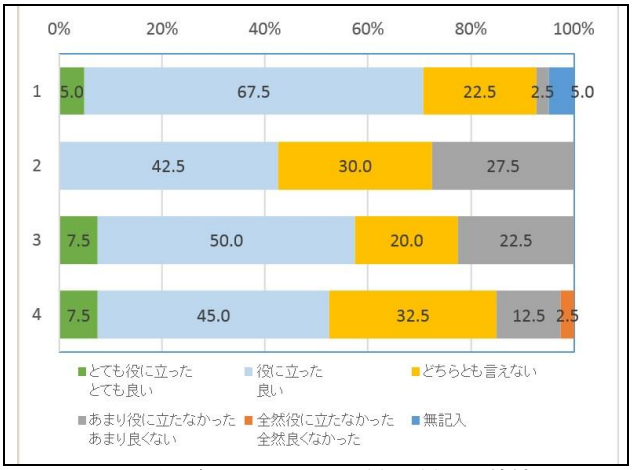

図６　試行後アンケート設問(1)～(4)の回答結果

(1)「かりん」の有用性

まず「『かりん』は役に立ったか」という設問に対し、今回の調査では実際の必要に応じた検索ではなかったが、「とても役に立った」5.0%、「役に立った」67.5%の回答を得た。その理由として「作文を書くときに、すぐに探せるので便利」(中)、「論文やレポートを書く時、ある単語とある単語の組み合わせが正しいかわからない場合に役に立つと思いました」(韓) など検索システム「かりん」の目的に適った理由や「電子辞書になかった言葉の組み合わせは”Karin″で検索できました」(他)、「言葉の自然な使い方を学ぶためにとても良い教材に思える」(英) など本検索システムの意義を認めるコメントも見られた。

一方で、「N3－N4 もあれば、もっと役に立ったと思う」(他)、「例文があまりない。よく意味が分からない」(タイ)、「入力したコロケーションの翻訳がない。翻訳されたコロケーションの意味にちょっと不自然で変な感じがある」(中) など、システムの充実度の低さから評価が下がっていると思われる回答もあった。

(2) 画面（美しさ・わかりやすさ）

今回の 4 つの設問の中では「画面（美しさ・わかりやすさ)」は一番評価が低く、「大変良い」は回答がなく「良い」が 42.5%であり、「あまり良くなかった」は４設問中最も多く 27.5%であった。その理由は「見やすいが白くてあまりきれいではない」(中)、「かわいくない」(中・韓)、「見た目は良くないが、機能が使えれば問題はない」(英) など画面の美しさ・魅力という点で多くの否定的なコメントが見られた。また「デザインが簡単で見やすい」(他) などシンプルでわかりやすいという評価も散見された一方で、「あまりはっきりしていない」(他)「どこから始めてよいのかよくわからない」(英) など、操作のわかりにくさについての指摘も複数あった。

また「言葉と動詞・修飾語の文字の色が異なっていて見やすい」(韓) など文字の見やすさについての言及もあったものの、「フォントが読みにくい」(タイ)、「表が読みにくい」(英)、「列の中でテキストがとても近いのでわかりにくい」(英) など文の読みにくさについて、中国語・韓国語以外を母語とする学習者からのコメントがいくつか見られ、様々な母語の学習者からのコメントを収集する重要性が改めて明らかになった。

(3) 操作性（使いやすさ・わかりやすさ）

「操作性」についても、「大変良い」「良い」はそれぞれ 7.5%、50.0%であったが、「あまり良くない」も 22.5%あった。その理由は多岐に渡っており、評価された点としては、「画面の変換が新しいデータを読み込んでくるのではなくて、すぐに飛んでいくことがいい（時間がかからない)」(韓)、「システムにどんな言葉が入っているか、頭文字のリストを使って分かる」(他) などがあった。

また全般的なわかりやすさについては様々で、「使いやすい」(中)、「２～３回トライしたら、誰でも使えると思います」(韓) などの意見がある一方で、「慣れるまで時間かかると思います」(中)、「日本人に聞きながらでなければ理解するのは難しい」(他) などの意見も複数見られた。

個々の機能や情報については「いくつかのボタンや用語はあまりはっきりしていない。もしダイス係数などがしっかり定義されていたら、もっと良い」（英）などダイス係数や学習指標値の意味に関する質問が複数見られた。こうした用語の説明は、すでに記載されているものもあるが、学習者にわかりやすい提示方法に修正する必要がある。

(4) 例文・翻訳（長さ・わかりやすさ）

「例文・翻訳」については、「大変良い」「良い」がそれぞれ7.5%、45.0%であり、有用性に次いで評価が高く、学習者からも様々なコメントがあった。

「例は簡単に理解するのに十分な短さだった」（英）、「例文は簡単でコロケーションの意味が分かりやすい」（他）などの好評価が多く見られたものの、最も多かったのは「もっと例文、翻訳が欲しいです」（中）、「私の見方が悪いのか、英語が見られない」（英）など例文や翻訳の不足であり、これらは随時充実していく必要がある。

また翻訳について中国語や韓国語でその妥当性に関するコメントが複数見られた。現在翻訳は各言語の上級レベルの留学生を中心に依頼しているが、細かい訳の違いなどは、筆者らが確認することは不可能であり、留学生複数人で確認するなどの慎重な作業を行っていく必要がある。

(5) その他

「名詞＋動詞」のコロケーション検索において付与した、機能動詞や複合助辞的用法の動詞の分類については、「一番役に立つ」（中）、「意味がピンとこない動詞が早く理解できました」（韓）などの好評価も複数得られたが、他方「もう少し詳しい説明があるともっと良いツールになる」（英）などもあり、すべての日本語学習者に理解しやすいツールになるためには改善が必要と考えられる。

また、JLPTや学習指標値の表示については、「とても役に立った」（中）、「良かった」（中、英）、「もっと目立つように」（中）など両数値の表示に好意的な記述が複数見られたが、「あまり意味がない」（中）、「私はJLPTの順番で言葉を覚えるわけではない」（他）といった記述も見られた。

同様にふりがなについても「とても有用だと思う」（中）、「詳しい画面にもあったら便利だと思います」（中）など、好意的な記述が複数見られた一方で、「自分で名詞を記入するように、選択でなしに」（中）、「見にくくて目が回る」（タイ）など、マウスカーソルを置くと常に表示されてしまう今の表示方法については検討の余地があると考えられる。

さらに今回の調査では、検索可能な名詞の数が少ないことについて多くの指摘があり、名詞の数を増やすことは重要と考えられる。しかしその際、先述の「N3ーN4もあれば、もっと役に立ったと思う」（英）や「私はJLPTの順番で言葉を覚えるわけではない」（他）などの指摘を考えると、JLPT受験者や上級レベルの学習者の数がアジア圏ほど多くない欧米圏の学習者にとっても使いやすく有用であるためには、名詞の追加基準について、JLPTのみでなく一般的な名詞の使用頻度や重要度などを考慮することや、語彙レベルを広げることも検討していく必要があると考えられる。

4. おわりに ―今後の課題―

以上、日本語学習者を対象に開発したコロケーション検索サイト「かりん」の説明と、試行の結果について述べた。

今回の試行により、本検索システムの有用性については、初めてアクセスした学習者からも多くの理解が得られたことは確認できたが、他方、画面の魅力、操作性、データの充実など多くの課題があることがわかった。また全般的な翻訳の妥当性や、特にヨーロッパ言語母語話者など中国語・韓国語以外の母語話者にとっての字の見やすさなど、学習者の直接の意見の聴取と確認が不可欠な課題もあることが明らかになった。

今後も、新規利用者の利便性や利用継続を促す魅力などを兼ね備えた有用なサイトを目指してさらなる改善を進めていきたい。

謝辞

本研究は科研費（基盤研究(C)25370591）の助成を受けたものである。

参考文献

(1) 国立国語研究所・Lago言語研究所：http://nlb.ninjal.ac.jp/
(2) 筑波大学・Lago言語研究所：”Webコーパス”, http://nlt.tsukuba.lagoinst.info/
(3) 日本語学習支援システム 日本語共起語検索システム「なつめ」https://hinoki-project.org/natsume/
(4) 国立国語研究所コーパス開発センター http://pj.ninjal.ac.jp/corpus_center/bccwj/（2011）
(5) 徳弘康代：”日本語教育における中上級漢字語彙教育の研究”, 早稲田大学大学院日本語教育研究科 博士論文, http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/5428/1/Honbun-4252.pdf,（2006）
(6) 村木新次郎:”日本語動詞の諸相”, ひつじ書房（1991）
(7) 松木正恵：”接続関係を表示する複合辞的表現―名詞性接続成分のタイプから見た連体複文構文と連用複文構文の接点―”,「複文構文の意味の研究」ワークショップ（2012年12月18日、国立国語研究所）https://www.ninjal.ac.jp/event/specialists/project-meeting/files/20111217-043re/20111218_matsuki.pdf#search='%E8%A4%87%E5%90%88%E5%8A%A9%E8%BE%9E'