コーパスを利用した基本動詞ハンドブック作成
―コーパスブラウジングツール NINJAL-LWP の特長と機能―

Prashant Pardeshi（国立国語研究所言語対照系）
prashant@ninjal.ac.jp
赤瀬川 史朗（Lago 言語研究所）
lagoinst@gmail.com

1. 日本語学習者用基本動詞用法ハンドブック作成プロジェクト

コミュニケーションの基本単位となる文の骨格を決める重要な要素の一つが述語としての動詞である。日本語を外国語として学ぶ学習者にとって、日本語の運用能力を向上させるためには、使用頻度が高く多義的な基本動詞の体系的な学習が不可欠である。国立国語研究所では日本語研究の成果を日本語教育に応用する目的で、Pardeshi がリーダを務める共同研究プロジェクト「日本語学習者用基本動詞用例ハンドブックの作成」（以下、ハンドブックプロジェクト）を 2009 年 10 月から実施している[1]。

このハンドブック作成の最大の特色は、日本語母語話者の正用のコーパスと外国人日本語学習者の誤用のコーパスを利用した見出し語の執筆である。執筆担当者は、コーパスから得られる見出し語の使用頻度、表記（書字形）の分布、統語的・意味的な共起関係といった各種情報を客観的に分析しながら執筆を進めている。正用のコーパスには、『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）を使用し、Lago 言語研究所（赤瀬川）はその検索用システムとして、NINJAL-LagoWordProfiler（略称 NINJAL-LWP）を国立国語研究所と共同開発した[2]。また、誤用のコーパスには、寺村秀夫の『外国人学習者の日本語誤用例集』(1990)をデータベース化して[3]執筆に活用している。以下では、BCCWJ ブラウジングシステム NINJAL-LWP の特長と機能について紹介する。

2. NINJAL-LWP の特長

コーパスの歴史はそのままコーパス検索ツールの歴史でもある。汎用の検索ツールとしては、当初から現在に至るまで、KWIC コンコーダンスが使用されている。キーワードを中心に並べて前後の文脈を素早く直感的に確認できるこのツールは、John Firth が残した有名な言葉“You shall know a word by the company it keeps.”(Firth 1957)をまさに実践するためのツールといってよい。しかし、1990 年以降のコーパスの大規模化にともない、このようなコンコーダンスだけを頼りにした言語分析には限界が訪れた。その解決策として、2000 年頃にはコロケーションを文法関係によって分類して表示するレキシカルプロファイリングという新たなツールが生まれた。NINJAL-LWP もこのレキシカルプロファイリングの流れを汲むツールで、内容語（動詞、名詞、形容詞、副詞）の振る舞いを見出し語単位で網羅的に調査・観察できる点に最大の特長がある。

---

[1] http://www.ninjal.ac.jp/research/project/b/youhoujiten/

[2] 開発のベースになったのは、Lago 言語研究所が 2005 年から開発している LagoWordProfiler（LWP）である。

[3] 2011 年 12 月、国立国語研究所より『寺村誤用例集データベース』として一般公開された。
http://www.ninjal.ac.jp/teramuragoyoureishu/

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

NINJAL-LWP の開発で特に重視したのはユーザーインターフェースである。これまでコーパスは当然のように「検索する」ものだと考えられてきた。しかし、NINJAL-LWP では、コーパスを検索するという視点に加えて、コーパスを「ブラウズする」、さらに言えばコーパスを「読む」という視点に立って開発を進めている。本稿のタイトルに「コーパスブラウジングツール」という表現を用いたのもそのような意図がある。

NINJAL-LWP では、基本的にクリックという簡単な操作だけで、コーパスの森奥深くまで進んでいくことができる。操作性の向上を図るため、図 1 のように、文法パターン、コロケーション、用例のすべてを同一画面で確認できるインターフェースを採用している。そのため、思考を中断されずにコーパスを読み進めることができる。

図 1　NINJAL-LWP

また、日本語の表記の多様性に対応させるため、代表表記という考えを取り入れている。副詞の「やはり」を例に挙げると、「やはり」という一般的な表記のほかに、まれに「矢張り」という漢字表記も使われる。さらに、「やっぱり」、「やっぱし」、「やっぱ」などの変異形も存在する。NINJAL-LWP では、これらを「やはり」という一つの見出し語に集約することで、表記の違いに関係なく語の振る舞いを確認することができるようになっている。

3. NINJAL-LWP の機能

3.1 文法パターンの表示

NINJAL-LWP では、先に述べたレキシカルプロファイリングの手法を利用して、文法関係を複数のグループに分類し、そこから各パターンに該当するコロケーションを表示する仕組みになっている。図 2 は、「分ける」の文法パターンのうち、動詞に先行する「名詞＋助詞」のパターン、つまり、動詞項・付加詞のパターンを示している。動詞の文法パターンには、この他に、「名詞＋複合助詞」が先行するパターン、名詞が後続するパターン、複合動詞など、合計 9 つのグループがある。

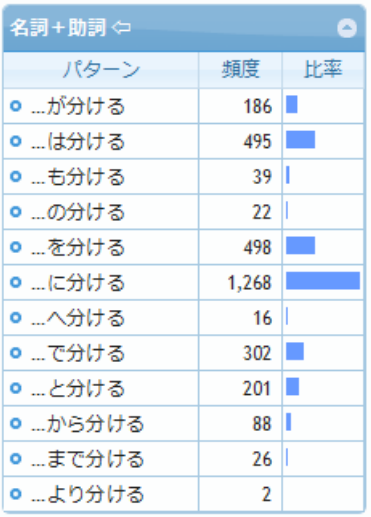

図 2　動詞項・付加詞のパターン

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

3.2 コロケーションの表示

先ほどの図２の文法パターンの［…を分ける］をクリックすると、ヲ格のコロケーションがその右にあるコロケーションパネルに表示される（図３）。クリックした直後は頻度順（画面上では FQ）になっている。

…を分ける 314件

コロケーション	FQ	LD	MI	N-S
血を分ける	38	7.81	9.90	-1.25
生死を分ける	9	8.48	11.77	-0.02
それを分ける	8	1.05	3.02	-0.25
命運を分ける	8	8.77	13.26	0.13
明暗を分ける	8	8.73	13.08	0.18
水を分ける	8	4.16	6.17	-0.07
株を分ける	7	5.89	8.05	0.13
勝敗を分ける	7	8.36	12.12	0.16
これを分ける	7	1.06	3.03	-0.14
人を分ける	5	-0.15	1.81	-0.09
道を分ける	4	3.35	5.37	0.09
血肉を分ける	4	7.93	13.68	-0.32
色を分ける	4	3.65	5.68	0.04
物を分ける	4	2.95	4.95	0.04
成否を分ける	4	7.69	11.85	0.09
地域を分ける	4	3.17	5.18	-0.11
人込みを分ける	4	7.52	11.20	0.09
パワーを分ける	4	6.18	8.61	0.07

Page 1 /4 100

図３　「分ける」のヲ格のコロケーション

NINJAL-LWP では、粗頻度のほかに、MI スコア（画面上では MI）、LogDice（画面上では LD）という２種類の統計値でも並べ替えができる。MI スコアの順に並べ替えると、頻度があまり高くなくても、慣用表現などの特徴的なコロケーションが上位に来やすい[4]。図３の「…を分ける」のコロケーションを MI スコアの順に並べ替えると、頻度では第３位の「命運を分ける」と「明暗を分ける」がそれぞれトップと第２位に来る。

高頻度の見出し語ではコロケーションの数が１万を超えることもある。そのような場合、MI スコアで並べ替えることで、頻度では上位に来ない重要なコロケーションを見逃すことなく観察できる。

この他、書籍サブコーパスの地の文と会話文の 100 万語当たりの頻度差（画面上では N-S）も表示されるため、書き言葉と話し言葉の使用頻度の違いをある程度推定することができる。

3.3 用例の表示

コロケーションパネルの各コロケーションをクリックすると、その右の用例パネルに用例とその出典が表示される。どのサブコーパスの用例かは、用例の前についた色つきの■で区別することができるようになっている。

また、辞書執筆で欠かせない作例を考慮して、用例はセンテンスの短い順に表示して、執筆者の便宜を図っている。前後の文脈を確認したいときは，出典部分をクリックすると

血を分ける
NR: 0.53　SP: 1.78　OM: 0　OC: 0.37　OY: 0.82

やはり、血を分けた御前様はちがったもの。
（安西篤子著『麒麟の母』，1989, 9 文学）

それは血をわけた姪を見つめるものではなかった。
（高遠砂夜著『リルファーレの冠』，2004, 9 文学）

遊び相手はすべて身内であり、血をわけたわが同胞であった。
（矢川澄子著『わたしのメルヘン散歩』，1987, 9 文学）

姓は違っても血を分けた孫がいるだけで十分だと思いますが。
（Yahoo!知恵袋, 2005, 子育てと学校）

まず第一に、菅野菜月には血を分けた肉親がひとりもいなかった。
（柴田よしき著『少女達がいた街』，1997, 9 文学）

血を分けた息子に拒否されるとは、考えてもいなかった愚かな私。
（辻仁成著『ニュートンの林檎』，1999, 9 文学）

つうか血を分けた兄弟ならマナがノアでも不思議じゃない気もしますが。
（Yahoo!ブログ, 2008, Yahoo!サービス）

Page 1 /1 100　38件中　1 - 38を表示

図４　「血を分ける」の用例

[4] MI スコアでは低頻度のコロケーションが強調されて高い値を示す傾向があるため、NINJAL-LWP では低頻度の低いコロケーションを排除して表示する機能を備えている。

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

文脈を示したダイアログが開く。

3.4 見出し語比較機能

語の振る舞いについて考察するときの重要な調査の一つにシノニムの比較がある。NINJAL-LWP では、シノニムなどの２語の見出し語の違いを調べるための比較機能を備えている。図５は、「起こる」と「発生する」のガ格の名詞を比較したものである。シノニムはネイティブでも簡単にその違いを説明することができない場合が少なくない。NINJAL-LWP では、コロケーションや用例の比較を通じて、シノニムの違いを明確に捉えることができる。

…が～

コロケーション	起こる	MI差	LD差	発生する
拍手が起こる	31	9.77	7.63	0
何が起こる	354	6.15	7.31	0
運動が起こる	59	6.52	7.17	0
反応が起こる	35	7.05	7.12	0
革命が起こる	34	7.01	7.08	0
奇跡が起こる	18	8.74	6.82	0
どよめきが起こる	16	11.55	6.8	0
何事が起こる	19	8.05	6.79	0
騒ぎが起こる	16	8.02	6.58	0
炎症が起こる	14	9.05	6.52	0

…が～

コロケーション	起こる	MI差	LD差	発生する
徴が発生する	0	-13.54	-7.95	15
結露が発生する	0	-12.17	-7.81	14
効力が発生する	0	-9.21	-7.6	17
癌が発生する	0	-9.43	-7.46	14
霧が発生する	0	-8.67	-7.16	13
ガスが発生する	0	-7.6	-6.78	15
事案が発生する	0	-8.97	-6.73	8
雑草が発生する	0	-8.96	-6.58	7
水素が発生する	0	-8.55	-6.49	7
義務が発生する	0	-7.05	-6.48	16

図５　「起こる」と「発生する」のガ格の比較

4. まとめ

以上、BCCWJ ブラウジングシステム NINJAL-LWP の特長と機能を述べてきた。NINJAL-LWP は、ハンドブックプロジェクトの一つの成果として，2012 年５月初旬、国立国語研究所からの一般公開が予定されている。今後は、辞書執筆という限定された用途だけでなく、広く日本語研究や教育に資するツールとして発展させていくつもりである。また同時に、それぞれの研究分野に特化させたバージョンの開発にも力を注ぎたいと考えている。

参考文献

Firth, J. (1957) “A Synopsis of Linguistic Theory 1930-1955” In Studies in Linguistic Analysis, Philological Society, Oxford; reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Harlow: Longman.

Kilgarriff A. and Rundell, M. (2002) “Lexical Profiling Software and its lexicographic applications: a case study” In Proceeding of EURALEX 2002, Copenhagen. pp. 807-818.

前川喜久雄 (2008) 「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, 4(1), pp.82-95.（招待論文）

プラシャント・パルデシ，赤瀬川史朗（2011）「BCCWJ を活用した基本動詞ハンドブック作成―コーパスブラウジングシステム NINJAL-LWP の特長と機能―」，特定領域研究「日本語コーパス」『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.205-216.

寺村秀夫 (1990) 『外国人学習者の日本語誤用例集』（科研費報告資料）

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved