1 億語コーパスで見えること、10 億語コーパスで見えること
–コロケーションについて–[1]

今井 新悟
筑波大学人文社会系

キーワード：コーパス　ウェブコーパス　コロケーション

１．はじめに

NLB（NINJAL-LWP for BCCWJ）は、国立国語研究所が構築した BCCWJ（現代書き言葉均衡コーパス）の検索ツールである。「国語研からは BCCWJ 用のオンラインコンコーダンサとして中納言と少納言が公開されていますが、NLB はコンコーダンサとは異なるレキシカルプロファイリングという手法を用いたコーパス検索ツールです。名詞や動詞などの内容語の共起関係や文法的振る舞いを網羅的に表示できるのが最大の特長です。」（http://nlb.ninjal.ac.jp/）これと同様の検索方法で利用できる NLT（NINJAL-LWP for Tsukuba Web Corpus）が近々公開予定である。BCCWJ は入念にサンプリングされたバランスの取れた約 1 億語のコーパスである。（現在公開中の NLB は現行 BCCWJ よりもデータサイズが小さい BCCWJ の領域公開データの 6 千 2 百万語を用いている。）筑波ウェブコーパス Tsukuba Web Corpus（TWC）はウェブ上からクローリンしてデータを集めた約 11 億語のコーパスである。ウェブ上からデータを収集する際の課題となるデータの偏りを修正するために、BCCWJ で得られた頻度情報を基に、BCCWJ の語分布に近づける工夫や、同一 URL からの重複データ収集を避ける工夫を施してある。

[1] ２節と３節は「筑波ウェブコーパス検索ツール NLT の開発」『第 3 回コーパス日本語学ワークショップ』（国立国語研究所）2013 年 2 月 28 日　からの抜粋。

２．動詞出現頻度の比較

NLB と TLB で抽出された動詞を頻度順にならべ、それぞれ上位１万語を抽出し、どちらか一方に現れない語を削除して、9001 語を得た。両者の頻度によるピアソンの相関係数は 0.989 であり、NLB での動詞の分布と TLB での動詞の分布は極めて相似している。筑波ウェブコーパス（TWC）のデータ収集はウェブのクローリングにより収集されるデータの偏りを克服するため、前述の通り、BCCWJ の語分布を模するという方略（および上記の各種方法）を使った。両者の動詞の相関を見ると、ウェブコーパスの弱点である偏りを克服するという課題は相当程度達成されたと言える。

順位を使ったスピアマンの順序相関は 0.887 である。順位で見てもマクロ的には両者は似ているといえるが、ミクロで見ると違いが現れる。順位の差が大きいものは、表１の通りである。

表１　筑波ウェブコーパス（TWC）と BCCWJ の動詞頻度順位の比較

動詞	TWC 順位	BCCWJ 順位	TWC 頻度	BCCWJ 頻度	順位差
答えする	2174	9493	4805	15	-7319
開講する	2546	8720	3726	20	-6174
退会する	3342	9323	2315	16	-5981
許諾する	3740	9696	1910	14	-5956
被曝する	2766	8583	3265	21	-5817
来場する	4184	9932	1538	13	-5748
選考する	4195	9932	1532	13	-5737
研修する	3396	8999	2257	18	-5603
支払いする	2113	7615	5004	30	-5502
祭りする	3747	8861	1904	19	-5114
フォーカスする	4081	9163	1626	17	-5082
リニューアルする	2863	7910	3046	27	-5047
マッチングする	4922	9932	1094	13	-5010
試行錯誤する	4630	9493	1256	15	-4863
拝読する	4141	8999	1576	18	-4858
目指せる	4641	9493	1249	15	-4852
出展する	3402	8215	2248	24	-4813
付帯する	3817	8583	1842	21	-4766
カスタマイズする	3351	8114	2307	25	-4763
正解する	4966	9696	1070	14	-4730
（中略）					
哀願する	9526	4825	190	85	4701
飛び退く	9782	5077	175	77	4705
血走る	9095	4364	220	103	4731
調味する	9542	4734	189	88	4808
舌打ちする	8173	3209	307	171	4964

上気する	9299	4331	205	104	4968
すすり泣く	9247	4259	209	107	4988
言いかける	8043	2940	322	195	5103
後ずさる	9095	3965	220	121	5130
しゃくる	8808	3620	242	143	5188
まさぐる	9359	4011	201	119	5348
にこりする	9552	4166	188	111	5386
微笑する	7997	2601	327	237	5396
くぐもる	9451	3896	195	125	5555
座り直す	9722	4126	179	113	5596
愛撫する	9396	3174	198	174	6222

筑波ウェブコーパスの方が BCCWJ より相対的に順位が高いもののうち、「答えする」「支払いする」などはそれぞれ「お答えする」「お支払いする」の形で使われているものである。この表では割愛したが、同様に筑波ウェブコーパスの方が BCCWJ より相対的に順位が高いものの中には、「(お) 届けする」「(お) 預かりする」のように相手を想定した敬体での使用が多い。ウェブ上では顧客相手の情報が多いことの反映であろう。また、「フォーカスする」「リニューアルする」「マッチングする」「カスタマイズする」等のカタカナ語も目立つ。また、「被曝する」のように時事的な話題を反映したと思われるものが入っている。一方で、BCCWJ の方が筑波ウェブコーパスよりも相対的に順位が高いものには、小説など文学作品においての人物の動作描写に使われそうな語が並んでいる。

３．コロケーション

３．１「～が走る」

「走る」のガ格に共起する名詞について、BCCWJ に基づく NLB から頻度 2 以上の共起語を取り出し、120 語を得た。それら 120 語の筑波ウェブコーパスにおける同様の共起頻度を求め、両者の相関を取ると、0.758 となった。このことから、両者の相関はある程度高いものの、収集されているコロケーションにはある程度違いがあることが予想される。なお、筑波ウェブコーパスの 5 億 8 千万語のパイロット版と今回の筑波ウェブコーパスの 11 億語版での相関は 0.995 であったことから、「～が走る」のコロケーションについては約 5 億語で相当程度安定して収集できることが示唆される。ただし、「～が走る」では、頻度が高いことから比較的安定して収集できたものであり、頻度の低いコロケーションでは、11 億語版であっても安定しないということもありうる。

筑波ウェブコーパスの「走る」のガ格に共起する名詞で頻度 20 以上のものは 103 語であった。そこから、代名詞、「もの」、「こと」など実質語的意味が希薄な語を除き、意味でカテゴリ化した。例えば、「車」「電車」「自転車」などを「乗り物」というカテゴリにした。表 2 にカテゴリ内の頻度計が 70 頻度以上となったものを示す。なお、右に添えられている数字はそれぞれの出現頻度出現頻度である。

表２　筑波ウェブコーパスにおける「～が走る」の共起語

順位	カテゴリ	共起語例
(1)	乗り物 3284	車 889、電車 513、バス 413、列車 343、自転車 181
(2)	人・動物 1896	人 251、馬 197、私 137、自分 121

(3)	痛み 1078	痛み 565、激痛 513
(4)	経路 616	道路 178、鉄道 128、道 109
(5)	動揺・衝撃 473	衝撃 275、激震 81、戦慄 49、動揺 48
(6)	感覚 261	～感 81、悪寒 59、痺れ 38、寒気 36
(7)	緊張 248	緊張 248
(8)	線 292	線 236（路線名も含む)、ライン 28、筋 28
(9)	光 212	光 77、閃光 68、稲妻 67
(10)	電気 205	電気 134、電流 71
(11)	溝・亀裂 180	亀裂 102、断層 51、溝 27
(12)	地形 102	～系 45、山脈 32
(13)	虫唾 86	
(14)	線状器官 77	神経 44、血管 33

コロケーションの頻度情報とそのカテゴリ化はコーパス準拠（corpus-based）の辞書編纂に有用である。表 2 の順番に辞書の語義を並べることに特に違和感はなく、ほぼ直観に合っていると言えよう。語義とその配列順序を決めてから例文を探すあるいは作例するという従来の方法とは逆に、コーパスのコロケーションから意味のカテゴリ化を行い、語義を決めるという方法の可能性を示唆している。ただし、「走る」の中心義は「人・動物が足を速く動かして移動する」であり、「乗り物が速く移動する」は意味拡張であろうから、後者の方が圧倒的に頻度が高いものの、辞書編纂においてはコーパス駆動（corpus-driven）ではなく、コーパス準拠（Corpus-based）が望ましい。

さて、コーパスのコロケーション頻度の有用性を確認したが、この頻度がある程度高くないと、コロケーションの情報が不安定になり、有用性が損なわれる可能性があるので注意が必要である。筑波ウェブコーパスでは共起語の出現頻度上位 51 語（頻度 48 以上の語）に限ってみても各カテゴリの順位は 5 番目までは表 2 と変わらない。

(1) 乗り物 2992、(2) 人・動物 1307、(3) 痛み 1078、(4) 経路 529、(5) 動揺・衝撃 453 (6) 緊張 248、(7) 線 236、 (8) 光 212、(9) 電気 205、(10) 感覚 140、(11) 溝・亀裂 102、(12) 地形 51

一方、BCCWJ では、頻度上位 50 語（頻度 5 以上の語）で見ると、カテゴリの頻度順は相当変化し、表 2 と等しいのは順位 1 位の「乗り物」だけになり、コロケーションの情報がやや不安定になっている。コーパス駆動（Corpus-driven）ではなく、コーパス準拠（Corpus-based）だとしても、コロケーション情報は安定して得られる方がよい。

(1) 乗り物 151、(2) 痛み 112、(3) 人・動物 96、(4) 光 73 、(5) 感覚 66、(6) 動揺・衝撃 60、(7) 経路 35、(8) 緊張 32、(9) 溝・亀裂 26、(10) 線 25、(11) 電気 15、(12) 地形 10、(13)予感 8

BCCWJ でも頻度 2 以上を採用すると共起語として出現する語数は 120 語となり、以下のような順番・頻度となる。これにより筑波ウェブコーパスのカテゴリ頻度順に近づく。それでも上位 3 位までは同じになるが、それ以下の順位は異なる。

(1) 乗り物 174、(2) 人・動物 148、(3) 痛み 112、(4) 光 77、(5) 感覚 74、(6) 動揺・衝撃 68、(7) 溝・亀裂 44、(8) 経路 39、(9) 緊張 32、(10) 線 29、(11) 電気 21、(12) 地形 16

以上、「～が走る」のコロケーションの場合は筑波ウェブコーパスではコロケーション情報を安定して取り出せるが、BCCWJ の場合はコロケーションの頻度についてはやや不安定になる嫌いがある。BCCWJ においても、出現頻度が 2 までと低いものまで観察の範囲を広げることによって、安定性をある程度向上させられることを見たが、一方、出現頻度が 2 というのは少なすぎて、ノイズ（誤り、個人的な癖など）の影響が高まる懸念も生じる。

３．２「～を駆ける」

前節では、「～が走る」という比較的頻度が高い例を見たが、本節では比較的頻度が低い「～を駆ける」を見てみる。BCCWJ では共起語のうち、頻度 3 以上のもので、(「なか」、「ウマ」、「間」、「上」のように共起語の分析に適さないものを除くと、道 10、廊下 4、階段 3、戦場 3、山 3、夜道 3、前 3 の 7 語のみである。頻度 2 のものはノイズ（誤り、個人的な癖など）が影響する可能性が高いので、対象外とするが、例えこれらを含めてもあと 13 語増えるのみであり、コロケーションを意味でカテゴリ化して示すことは難しい。一方、筑波ウェブコーパスでは頻度 3 以上の語は 50 ほどあり、以下のようなカテゴリ化が可能である。ただし、頻度が「～が走る」に比べる大分少ないので、カテゴリの頻度で順序を見るには適さないだろう。

表３　筑波ウェブコーパスにおける「～を駆ける」の共起語

カテゴリ	共起語例
空 92	空 61、大空 11、宇宙 8、天空 7、夜空 5
野山 59	草原 16、野 12、野山 11、山 7、森 6
経路 47	道 15、廊下 13、路 11、階段 8
戦場 31	戦場 31
世界 27	世界 27
地 22	地 11、大地 8
区域 21	街・町 11、庭 4、コート 3
前 16	先頭 6、先 5、前 5
時 16	時代 10、時 6
海 11	海 8、海上 3

４．「に」と「へ」

４．１．意味

日本語教育の初級学習項目である「に」と「へ」については、「に」は着点、「へ」は方向を表すと教えるのが一般的だろう。確かに、着点を表す「着く」では「へ」よりも「に」の方が自然な感じがする。

(1a) 東京に着いた。
(1b) 東京へ着いた。

しかし、「へ着く」も非文ではない。さらに、「行く」の場合、(2a)は動作主が東京という着点に到達したことを含意し、(2b)は移動の方向を表して着点への到達を含意しないというような違いは、筆者には全く感じられない。

(2a) 東京に行った。
(2b) 東京へ行った。

もし、到着が含意されないなら、「着かなかった」を後続させてもそれが文法的になるはずであるが、(3b)は(3a)と同程度に不自然であることから、両者に到達性の含意の違いを認めることはできない。

(3a) *東京に行ったが着かなかった。
(3b) *東京へ行ったが着かなかった。

(4a)(4b)の「向かった」は到着が必然的に含意されないため、「着かなかった」の後続が可である。しかし、「に」と「へ」が交替可能であるから、単純に「に」は着点、「へ」は方向を表すとは言えない。

(4a) 東京へ向かったが着かなかった。
(4b) 東京に向かったが着かなかった。

結局、日本語教育の現場では、「に」と「へ」の意味的な違いに言及することはあっても、両者の違いはほとんどなく、交替可能であると教えるのが一般的であろう[2]。以下では、「に」「へ」と共起するコロケーションの出現頻度に注目してその使用実態を明らかにし、日本語教育の現場での説明の方法についての提案を行う。

４．２．データと考察

筑波ウェブコーパスを使って、以下のように共起語を数える。表４は「行く」の例である。「行く」には各種活用形も含む。「～に行く」の「～」の部分に共起する語で頻度の高いものの上位５位は、「(地域)、病院、学校、店、家」である。(地域) は地名のことであり、「東京、大阪、名古屋」などをすべて含む。「～へ行く」のは上位から「(地域)、病院、学校、海外」の４位までと同頻度の「店」と「家」である。「海外」は「～に行く」では上位５位に入っていないが、「～へ行く」では、上位５位以内であるので、「海外」も採用する。このように、原則として「～に行く」「～へ行く」との共起出現頻度がそれぞれ上位５位のものを採った。また、「遊びに行く」のような目的を表すもの、「ところに行く」などの具体性を欠く名詞などは採らなかった。また、「大学に行く」のような比喩的な用法が多いものも採らなかった。

表４　「～に行く」と「～へ行く」

	に	へ	「に」使用率
(地域)	21762	7624	0.741
病院	8301	3161	0.724
学校	6889	2058	0.770
店	3690	464	0.888
家	2688	464	0.853
海外	2135	557	0.793
計	**45465**	**14328**	**0.760**

このようにして21語についてコロケーション頻度を数え、「に」の出現率を割り出したものを資料として末尾に載せる。なお、形態素解析を自動で行なっているために、頻度には

[2] 「の」の後続の可否による違い、つまり、「にの」は非文法的であるということや、移動の目的には「に」は使えるが「へ」は使わない、などの文法的な制約は本稿では扱わない。

形態素解析が誤っている語もカウントされている。(あるいは本来カウントされるべきものが抜け落ちている。) しかし、そのような解析誤りは大勢に影響を与えるほどではなく、また、比較対象（ここでは「～に」と「～へ」）の両者に同程度に起こるため、比較自体に与える影響は少ないという仮定の下で以下の考察を行う。

21 の動詞は、概ねその動詞の意味カテゴリーに基づく種類を勘案して選定した。各動詞を「に」の使用率（表４の例では、計の 0.760）の高いものから降順に並べたのが以下の表である。

表５　「に」の使用率と動詞の種類

動詞	「に」使用率	動詞の種類
面する	1.000	接合
会う	0.999	対面
触れる	0.999	接触
ぶつかる	0.998	接触
掛ける	0.996	接触
接する	0.995	接合
座る	0.994	接触
あげる	0.991	所有移動
衝突する	0.990	接触
入れる	0.970	着点移動
着く	0.951	着点移動
入る	0.947	着点(位置)移動
差し上げる	0.936	所有移動
到着する	0.936	着点移動
提出する	0.919	所有移動
来る	0.911	位置移動
登る	0.901	位置(着点)移動
行く	0.760	位置移動
走る	0.666	様態性移動
歩く	0.594	様態性移動
向かう	0.544	方向移動

この表から、概ね、動詞の種類による以下のような「に」使用率の階層性が読み取れる。

接合／接触／対面＞所有移動／着点移動＞位置移動＞様態性移動＞方向移動

「面する」と「接する」は接合動詞と名付ける。両動詞には物理的移動が伴わないことが多い。接触動詞と名付けた「掛ける、ぶつかる、触れる[3]、座る、衝突する」は物理的な

[3] 「触れる」よりも「触る」の方がより基本的な語であるが、「触る」の例は受け身での使用が多く、NLT の検索では、受け身の動作主「に（よって)」と「に触る」の「に」が区別されずにカウントされてしまい、本稿で扱うデータと分析に与える影響が大きいので、

移動を伴うが、それがプロファイルされることはなく、背景化し、言語的にも「～を通って」「～を」などの経路や、「～から」のように始点を伴うことがない。対面動詞と名付けた「会う」も、物理的には主体の移動を伴ったとしても、移動の経路や始点は言語化しない。これら 3 種類の動詞に共通しているのは、移動がないか、あるいは背景化してしまっているものである。このように移動がないか背景化している場合、「～に」が表すのは、移動の終点ではなく、静的な接点である。これは存在位置を表す「机の上に本がある」の「に」の意味・用法と通じるものがある。

次に「に」の使用率が高い動詞の種類として所有移動動詞と着点移動動詞がある。所有移動動詞として「あげる、差し上げる、提出する」を代表させ調べた。所有移動動詞の場合、「～から～に／へ」のように始点が言語的に表示できる。これは先に見た接合／接触／対面動詞とは異なり、移動の意味が少し含まれることを示唆する。ただし、経路を表す「～を通って」「～を」が共起できないことから、移動の意味はまだ弱い。

「あげる」と「差し上げる」の両方を取り上げたのは、郭・林(2012)が「へ」は心理的距離の遠い相手や目上の相手など文の敬意が高まると使われやすいと主張していることを再検証するためである。郭・林(2012)は

(5)「佐藤さんは田中さん（へ・に）プレゼントをあげた。」(64 人・574 人)
(6)「生徒が先生（へ・に）花束を差し出した。」(353 人・285 人)

という「へ」と「に」の選択肢問題を日本語母語話者に与え、どちらを使うかを問うた。括弧内はそれぞれを選択した人数である。(5)では「に」の回答が 90.0%であるのに対し、(6)では「へ」の回答が 55.3%と過半数を超えたことから、目上の相手に対して「へ」を使いやすいという主張を裏付けるものとしている。しかし、本稿の筑波ウェブコーパスのデータでは敬意が含まれない「あげる」と敬意が含まれる「差し上げる」における「へ・に」の使用率には顕著な違いが認められなかった。もし、郭・林(2012)が主張するように「心理的距離」が「へ」の使用に関連するのであれば、それは比喩的な拡張である可能性が高く、「へ」に「物理的距離」の意味があるとすれば整合性が高まるであろうが、今のところそのような意味を「へ」に認める論証はないようだ。郭・林(2012)のデータと本稿のデータとの食い違いの原因は不明であり、本稿では、心理的距離・敬意により「に」と「へ」の使用に差が生じるという主張は保留しておきたい。

着点移動動詞は「着く」「到着する」「入れる」であり、着点に焦点がある動詞である。経路の「～を通って」「～と通して」は許されるだろうが、経路の意味での「～を」とは共起しない。「入る」と「登る」は、着点移動動詞と次に見る位置移動動詞の中間にあるものと考える。これらは着点動詞と同様、着点に焦点がある動詞でありながら、他方、次に見る位置移動動詞と同様に「～を」で経路を表すことができる。ただし、「入る」と「登る」とではその程度に差が感じられ、「入る」は着点動詞に近く、「登る」は位置移動動詞に近く感じられる。実際、「を」と「に」の共起数を見ると、「を入る」が 8107、「に入る」が 365172、「を登る」が 13610、「に登る」が 14883 であり、「入る」では「を」の現れが相対的に非常に少なく、「登る」では「を」と「に」が拮抗していた。よって、ここでは、「入る」は着点移動動詞として、「登る」は位置移動動詞として扱う。

着点移動動詞の「に」の使用率は所有移動動詞と同じ程度で、９割以上である。着点移動動詞の「着く」と「到着する」は和語と漢語の対であるが、「に」の使用率に顕著な差はない。「入る」と「入れる」は自他の対になっているが、これも両者の「に」使用率に顕著な差はなかった。

「触る」は調査の対象としなかった。

位置移動動詞は「行く」「来る」「登る」である。位置移動動詞は「～を」で経路を表しうる。これは移動の意味が上で見た他の種類の動詞よりも強くなっていることを示唆する。位置移動動詞の「に」の使用率は７～９割程度に下がる。

「走る、歩く」は様態性移動動詞と名付けた。これらは移動の様態を表す動詞である。なお、「走る、歩く」のデータには「て行く／て来る」が後続するものを含む。様態性移動動詞は位置移動動詞と同じく、「～を」で経路を表しうる。これらの「に」の使用率はさらに下がって６割程度である。「走る、歩く」の共起語に「方面」「方向」や方位の「東、西、南、北」が多いことが「へ」の使用率を押し上げている可能性があるので、方向性を持たない名詞について確認すると、以下のようになった。

表６　「に／へ歩く」と「に／へ走る」と非方向性名詞の共起

歩く（て行く・来る）	に	へ	「に」使用率	*走る(って行く・来る)*			
（地域）	27	66	0.290	（地域）	158	87	0.645
町・街・村	28	37	0.431	町・街・村	47	26	0.644
駅	16	38	0.296	家	30	34	0.469
家	12	10	0.545	駅	12	10	0.545
公園	9	8	0.529	公園	6	5	0.545
計	**92**	**159**	**0.367**	**計**	**253**	**162**	**0.610**

「走る」については「に」の使用率にさほどの変化はなく、約６割である。「歩く」については、「に」の使用率が上がることはなく、むしろ下がっている。ただし、頻度が低いため、使用率が不安定になっていることが考えられる。以上の観察から、少なくとも、「歩く」「走る」の様態性動詞については、「に」の使用率が位置移動動詞よりも低いということは検証された。

最後に方向移動動詞の「向かう」について見る。これと「に」の共起率はこれまでに見た動詞の中で最も低く、約５割である。「へ」の使用率が高まるのは、「へ」が方向を表すとされている従来の説明とも合致している。ただし、本稿では、「に」を着点、「へ」を方向という従来の２分法によるのではなく、動詞の種類の持つ「移動性」に注目して、その強弱の連続性に動詞の種類を位置づけたい。

以上をまとめて図示すると以下のようになる。この図には「～にある／いる」などの存在の「に」を左端に加えて、上で見た動詞の延長線上にこの用法を位置づける。左に位置する動詞ほど静的つまり存在の意味が強く、右に位置するほど動的、つまり移動の意味が強くなる。静的な動詞には「に」が、動的な動詞には「へ」が共起しやすくなる。「に」の意味は左端では存在点を、接合／接触／対面動詞では接点を、その他の動詞では着点を表す。「へ」は方向を表す。ただし、いずれの動詞の場合も「へ」は「に」に置き換えが可能であるが（注２参照）、その逆は必ずしも正しくなく、存在動詞では不可、接合／接触／対面動詞ではほぼ不可、所有移動／着点移動動詞では可能であるが頻度は低い。

静的（存在）　　動的（移動）

存在＞接合／接触／対面＞所有移動／着点移動＞位置移動＞様態性移動＞方向移動

に　　へ
存在点　接点　着点　方向

図２　動詞種類の階層

これらをまとめると、日本語教育の現場では以下のような説明が妥当だろう。

> 移動という動きがなければ「に」を使う。移動があっても、その着点に注目しているときは「に」が使いやすく、着点よりも動きに注目しているときは「へ」も使いやすくなるが、いずれの場合でも「に」で表してよい。

これを以下のようにイメージスキーマで表すこともできよう。

「に」のイメージスキーマ１　　「へ」のイメージスキーマ

「に」のイメージスキーマ２

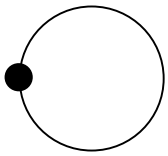

図３　「に」と「へ」のイメージスキーマ

「に」のイメージスキーマ１では、●が接している点をプロファイルしていることを示す。矢印が点線になっているのは、例え物理的な移動はあってもそれが背景化されていることを示す。そしてそれが完全になくなれば、存在動詞や接合／接触動詞のようになり、それはイメージスキーマ２で示される。両者が連続していることがイメージでも捉えやすいだろう。一方「へ」のイメージスキーマは方向・移動を実線の矢印で示して、それがプロファイルされていることを表しており、着点が必ずしも含意されなくてよいことを矢印が円から離れていることで表している。

５．まとめ

本稿では、まず、BCCWJ と筑波ウェブコーパスによってコロケーション情報の抽出を行い、両コーパスからの結果を比較した。コロケーション情報の抽出には、10 億語程度の規模が望ましく、コーパスサイズが小さいと、特にコロケーションのカテゴリー化に難があることが分かった。また、今回、「走る」と「駆ける」で比較したところ、頻度の高い「走る」ではコロケーション情報も安定して取り出せたのに対し、頻度の低い「駆ける」では、不安定になり、より大きなコーパスの必要性が示唆された。

次に、「に」と「へ」を取り上げ、内省では整理しにくい微妙な意味の違いを、コーパスのコロケーションの頻度に基づいて分析した。「に」と「へ」を伴う動詞にそのカテゴリカルな意味の種類により階層性が認められることを、共起語の頻度によって示した。さらに、日本語教育への応用として、「に」と「へ」の説明方法への提言を行った。

以上、これまでは作例と限られた使用例と内省とによって進められてきた研究に対して、大規模コーパスを使ったアプローチの例を示した。大規模コーパスとその検索ツールを使うことによって、短時間に膨大なデータが手に入り、研究・分析の効率がアップし、また、その圧倒的な量により、内省や作例では気付き難いことが明らかになる。ただし、大規模

コーパスは形態素分析などの処理が自動化されているので、誤分析は免れないということには常に注意しておかなくてはならない。このようなノイズあるいはゴミといわれるデータが紛れ込んでいることを認識して分析・利用することが求められる。

謝辞

筑波ウェブコーパスの構築および NLT（NINJAL-LWP for Tsukuba Web Corpus）の開発には、教育関係共同利用拠点「筑波大学留学生センター　日本語・日本事情遠隔教育拠点」の予算の一部が充てられています。NLT は同上拠点事業としてウェブ上で公開予定です。NLT の基盤となった NLB（NINJAL-LWP for BCCWJ）は、協同研究として、筑波大学留学生センターが国立国語研究所および Lago 言語研究所から使用許可を得て使用しています。

文献

Baroni, M. and Bernardini, S. (2004) *BootCaT: Bootstrapping corpora and terms from the web.* Proceedings of LREC 2004, Lisbon: ELDA. pp.1313-1316.
(http://www.cs.utah.edu/nlp/readinglist/BaroniB04.pdf よりダウンロード可能)

Fletcher,W.H.(2007) *Toward cleaner Web corpora: recognizing and repairing problems with hybrid online documents*. Corpus Linguistics 2007, Birmingham pp.27-30.
(http://webascorpus.org/CL07BhamWHFletcher.pdf よりダウンロード可能)

Hundt, Marianne, Nadja Nesselhauf and Carolin Biewer (Eds.) (2007) *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Sharoff S. 2006. *Creating general-purpose corpora using automated search engine queries.* In Marco Baroni and Silvia Bernardini (Eds), *WaCky! Working papers on the Web as Corpus*, Gedit, Bologna.

今井新悟・赤瀬川史朗 (2012)『日本語ウェブコーパスと BCCWJ コーパスの比較と日本語教育への応用』、2012 年日本語教育国際研究大会パネルセッション「日本語につながるコーパス研究―現状と今後の展望―」、日本語教育国際研究大会名古屋 2012 予稿集第 2 分冊、p.65.

郭潔・林伸一(2012)「格助詞『に』と『へ』の使い分けについてーアンケート調査の分析を基にー」『山口国文』35，70-84.

田野村忠温(2009)「コーパスからのコロケーション情報抽出ー分析手法の検討とコロケーション辞典項目の試作ー」『阪大日本語研究』21, 21-41.

プラシャント・パルデシ、赤瀬川史朗 (2012)『レキシカルプロファイリング手法を用いた BCCWJ 検索ツール NINJAL-LWP とその研究事例』、日本言語学会第 144 回大会ワークショップ「コーパス基盤の日本語研究の新地平」、日本言語学会第 144 回予稿集、pp.364-369.

関連 URL

NLB（NINJAL-LWP for BCCWJ）　http://nlb.ninjal.ac.jp/

NLT（NINJAL-LWP for Tsukuba Web Corpus）　http://corpus.tsukuba.ac.jp　（2013 年 4 月公開予定）

Sketch Engine　https://the.sketchengine.co.uk/

BootCaT　http://bootcat.sslmit.unibo.it/

国研コーパス開発センター 超大規模コーパス http://www.ninjal.ac.jp/corpus_center/ulc/

資料

	に	へ	「に」使用率		に	へ	「に」使用率
行く				*来る*			
（地域）	21762	7624	0.741	（地域）	13363	1398	0.905
病院	8301	3161	0.724	家	2577	170	0.938
学校	6889	2058	0.770	学校	1158	141	0.891
店	3690	464	0.888	店	1117	41	0.965
家	2688	464	0.853	病院	991	100	0.908
海外	2135	557	0.793	町	295	49	0.858
計	**45465**	**14328**	**0.760**	**計**	**19501**	**1899**	**0.911**
向かう				*登る*			
（地域）	4301	3870	0.526	山	2334	211	0.917
方面	1034	765	0.575	上	583	130	0.818
駅	872	576	0.602	木	564	1	0.998
家	417	201	0.675	丘	179	23	0.886
西	406	371	0.523	山頂	126	44	0.741
北	382	434	0.468	屋根	116	22	0.841
計	**7412**	**6217**	**0.544**	**計**	**3902**	**431**	**0.901**
歩く（て行く・来る）				*走る(って行く・来る)*			
方向	441	212	0.675	方向	482	120	0.801
方面	218	122	0.641	（地域）	158	87	0.645
南	76	63	0.547	方面	141	97	0.592
東	70	77	0.476	南	64	45	0.587
北	65	81	0.445	西	55	71	0.437
西	55	77	0.417	東	45	53	0.459
計	**925**	**632**	**0.594**	**計**	**945**	**473**	**0.666**
着く				*到着する*			
（地域）	3073	219	0.933	（地域）	2173	150	0.935
駅	1543	38	0.976	駅	931	42	0.957
家	973	17	0.983	空港	377	36	0.913
山頂	518	15	0.972	現場	283	10	0.966
空港	499	27	0.949	現地	216	12	0.947
病院	310	32	0.906	会場	160	24	0.870
港	302	22	0.932	病院	157	20	0.887
計	**7218**	**370**	**0.951**	**計**	**4297**	**294**	**0.936**
入る				*入れる*			
中・内	22593	1488	0.938	なか・内	10994	446	0.961

目	6603	2	1.000	袋	3660	50	0.987
風呂	5409	68	0.988	口	2916	38	0.987
（地域）	4674	445	0.913	箱	1858	58	0.970
部屋	3602	275	0.929	容器	1622	2	0.999
山	1200	139	0.896	冷蔵庫	1142	60	0.950
町	832	111	0.882	カート	517	41	0.927
計	**44913**	**2528**	**0.947**	**計**	**22709**	**695**	**0.970**
ぶつかる				***衝突する***			
壁	1526	0	1.000	地球	96	1	0.990
人	155	0	1.000	車	69	0	1.000
車	145	1	0.993	壁	50	0	1.000
物	135	0	1.000	船	40	0	1.000
岩	69	0	1.000	月	29	2	0.935
山	52	2	0.963	氷山	28	0	1.000
戸	11	2	0.846	**計**	**312**	**3**	**0.990**
計	**2093**	**5**	**0.998**				
会う				***触れる***			
人	3136	2	0.999	空気	918	0	1.000
～さん	2179	5	0.998	先	587	0	1.000
（人名）	1566	3	0.998	肌	518	2	0.996
先生	983	1	0.999	手	504	0	1.000
彼	734	0	1.000	体	427	0	1.000
彼女	409	0	1.000	**計**	**2954**	**2**	**0.999**
計	**9007**	**11**	**0.999**				
座る				***掛ける***			
イス	3566	15	0.996	壁	258	1	0.996
席	2568	24	0.991	肩	186	1	0.995
隣	1416	4	0.997	上	135	1	0.993
ベンチ	661	3	0.995	首	130	0	1.000
ソファー	509	3	0.994	ハンガー	76	0	1.000
計	**8720**	**49**	**0.994**	**計**	**785**	**3**	**0.996**
面する				***接する***			
道路	886	1	0.999	道路	485	1	0.998
海	761	0	1.000	地面	86	0	1.000
通り	703	0	1.000	通り	76	0	1.000
湾	209	0	1.000	海	57	1	0.983
広場	166	0	1.000	壁	46	0	1.000
計	**2725**	**1**	**1.000**	道	27	1	0.964
				水面	7	1	0.875
				計	**784**	**4**	**0.995**

あげる				*差し上げる*			
人	560	3	0.995	お客様	105	23	0.820
子供	157	1	0.994	人	81	1	0.988
～さん	138	1	0.993	～さん	73	0	1.000
あなた	118	1	0.992	全員	60	0	1.000
赤ちゃん	107	0	1.000	あなた	58	1	0.983
自分	77	3	0.963	皆様（さま）	41	5	0.891
同僚	7	2	0.778	先生	19	0	1.000
計	**1164**	**11**	**0.991**	**計**	**437**	**30**	**0.936**
提出する							
国会	1640	44	0.974				
市長	1281	16	0.988				
大臣	668	14	0.979				
窓口	602	163	0.787				
知事	509	10	0.981				
事務所	284	105	0.730				
学校	99	65	0.604				
税務署	394	64	0.860				
計	**5477**	**481**	**0.919**				

＊「目へ入る」の２例は形態素分析の誤りで、「２週目に入る」という例なので本来はカウントされるべきではない。同様の形態素分析誤りが「木へ登る」の１例、「人へあげる」の３例中１例が含まれる。

＊「～へ衝突する」「～へ会う」「～へ差し上げる」は頻度が極端に低かったためそれぞれ「～に衝突する」「～に会う」「～に差し上げる」の頻度上位６語を採った。