筑波ウェブコーパス検索ツール NLT の開発

今井 新悟（筑波大学）
赤瀬川 史朗（Lago 言語研究所）
プラシャント・パルデシ（国立国語研究所）

Development of NLT: the Search Tool for Tsukuba Web Corpus

Shingo Imai (Tsukuba University)
Shiro Akasegawa (Lago Institute of Language)
Prashant Pardeshi (National Institute for Japanese Language and Linguistics)

１．はじめに

本稿では 2013 年に一般公開を予定している NLT（NINJAL-LWP for Tsukuba Web Corpus）の開発とそのシステムの特長について述べる。NLT は 2012 年 6 月に公開した NLB（NINJAL-LWP for BCCWJ）と同一のシステム NINJAL-LWP で動作するレキシカルプロファイリング型のコーパス検索ツールである。検索対象となる筑波ウェブコーパス（TWC）は、ウェブ上から収集した約 11 億語の日本語のテキストデータである。以下では、TWC の構築と検索システムへの実装について述べた上で、公開前のシステムから得られた動詞頻度と動詞（「走る」と「駆ける」）と名詞のコロケーションの結果を NLB と比較し、その有用性と可能性について探りたい。

２．筑波ウェブコーパスの構築の目的と規模

２．１　構築の目的

一般に、コーパス基盤の言語研究においては、研究対象となる言語現象を複数のコーパスで比較して観察することで研究の信頼性や客観性を高めることができる。2011 年に完全公開された『日本語書き言葉均衡コーパス』（以下、BCCWJ）は日本語初の均衡コーパスで、その規模は約 1 億語である。2012 年 6 月にはこの BCCWJ 向けのレキシカルプロファイリング型のコーパス検索ツール NLB（NINJAL-LWP for BCCWJ）が一般公開された[1]。筑波ウェブコーパス（以下、TWC）の開発の最大の目的は NLB と同じ検索システムを利用して BCCWJ と比較できるウェブコーパスを構築することにある。同一のインターフェースを利用することで BCCWJ と TWC の比較が容易になるため、均衡コーパスの「質」とウェブコーパスの「量」の双方のメリットを言語研究や日本語教育に生かすことが期待できる。

２．２　コーパスの規模

均衡コーパスは、厳密な統計的手法に基づいてデータが採取されることから、コーパスの規模に関しては常に時間的・資金的制約がつきまとう。それに対して、ウェブ上のテキストを収集して構築するウェブコーパス[2]には事実上そのような制限はない。つまり、コーパスの規模は限りなく大きくできる。1 億語のウェブコーパスと 10 億語のウェブコーパスを比べれば、10 億語のコーパスのほうがより多くの有用な言語情報を含むと考えてよい。イギリスのコーパス統合ツールサイト Sketch Engine では、10 億語規模から TenTen と呼ばれる 100 億語規模の各国語のウェブコーパスが検索できる。国立国語研究所でも 100 億語を超える超大規模コーパスを開発中である。日本語においても、ウェブコーパスがふつうに活用される時代がすぐそこまで到来している。

TWC については、2012 年夏に 5 億 8 千万語のパイロット版を制作し、2013 年に公開する

[1] URL は文末参考 URL を参照。本稿では、現行の BCCWJ よりもデータサイズがやや小さい BCCWJ の領域公開データ（2009 年版）の 6 千 2 百万語を採録した NLBVer1.10 を用いる。

[2] 英語では、Web As Corpus（WaC）という言い方がよくされる。

一般公開版では 11 億語まで拡張する予定である。10 億語規模にした理由としては、BCCWJ の 10 倍強の規模で比較しやすい大きさであること、英語コーパスを利用した辞書制作のこれまでの経験から見て、10 億語が中型辞書の見出し語の用例を十分に採取できる一つの目安となること[3]、比較的短期間で構築できることなどが挙げられる。

３．筑波ウェブコーパスの構築の過程

３．１　収集方法

ウェブ上からのテキストの収集については、検索エンジンの API を利用して、ウェブページの URL を収集した後、その URL のデータを収集する一般的な手法に従った。具体的な手順については、ウェブコーパス構築ツール BootCaT を参考にしてプログラムを作成した。

【シードおよびタプルの生成】検索エンジンのクエリパラメータに与えるタプルを構成するシードには、NLB の開発過程で作成した BCCWJ（2009 年の領域公開データの一部、約 6 千 2 百万語）の頻度リストを利用した。品詞ごとに分かれた頻度リストのうち、内容語である名詞、動詞、形容詞、副詞のリストをマージして、上位 500 語をシードとして選んだ。ただし、名詞のうち、数詞、固有名詞は排除し、また、動詞、形容詞については活用形も含めた。この 500 語のシードから無作為に 3 語を選び出し、計 50 万組のタプルを作成した。以下にタプルの例を示す。

```
駄目 皆 構造
条件 とても 様々
法律 (答える OR 答え OR 答えよ OR 答えれ OR 答えろ OR 答えりゃ OR 答えん) 人々
```

【検索エンジン API による URL の収集】URL の収集には、Yahoo!ウェブ検索 API を利用した。1 タプル当たりで収集する URL 数は 10 ページとし、2012 年 1 月初旬から下旬にかけて計 500 万 URL を収集した。重複した URL を削除した URL 総数は約 3 割減の約 350 万件になった。

【HTML ページの収集】URL データを 5 万件ごとに分割した上で、3 台の端末を利用して 2 週間をかけて HTML ページを収集した。

３．２　コーパスデータの抽出

【テキストの抽出】次に収集した HTML ファイルからテキストを抽出する作業を行った。具体的には、HTML タグの削除、文字コードの統一（utf8）、日本語以外の言語で書かれたテキストの削除[4]を行った。

【不適正なページの排除】ウェブ上のテキストの収集の目的は日本語の用例を採取することにあるので、単に項目やリンクを列挙しただけのページ、広告と思われる内容の多いページ、センテンス境界の判定が難しいページは、あらかじめコーパスデータの対象から外した。

【センテンスの抽出】レキシカルプロファイリングツール NINJAL-LWP では、センテンス単位にした用例の中にどのようなコロケーションが含まれるかを文法パターン別に抽出する。そのため、コーパスデータはあらかじめセンテンス単位に分割しておく必要がある。一つ前の作業でセンテンス境界の判定が難しいページを排除したのもこの理由による。

【用例データの抽出】センテンス単位のデータのなかには、見出しに相当するものや、メニュー項目に相当するものが含まれる。センテンス中にどの程度名詞が含まれるか、セン

[3] 英語と日本語を比べた場合、同じ語数では英語のほうが情報量が多い。そのため、10 億語の英語と日本語では英語のほうが情報量が多くなる。その意味では 10 億語という数字はあくまでも目安に過ぎない。

[4] Perl モジュール Encode::Guess を利用した。

テンス中に動詞は現れるか、「クリック」や「ログイン」などのウェブページで多用される表現が用いられているかなどの複数の観点から、用例としての適正度を数値化し、用例としてふさわしいデータを抽出した。図 1 は、適正率を示したウェブページのテキストの例である。網がけになったセンテンスは適正率が高く、用例データとしてふさわしいと判断されたものである。さらに、同一ページで同じセンテンスが現れた場合も、最初の 1 件のみを用例として採取し重複を避ける工夫をした。

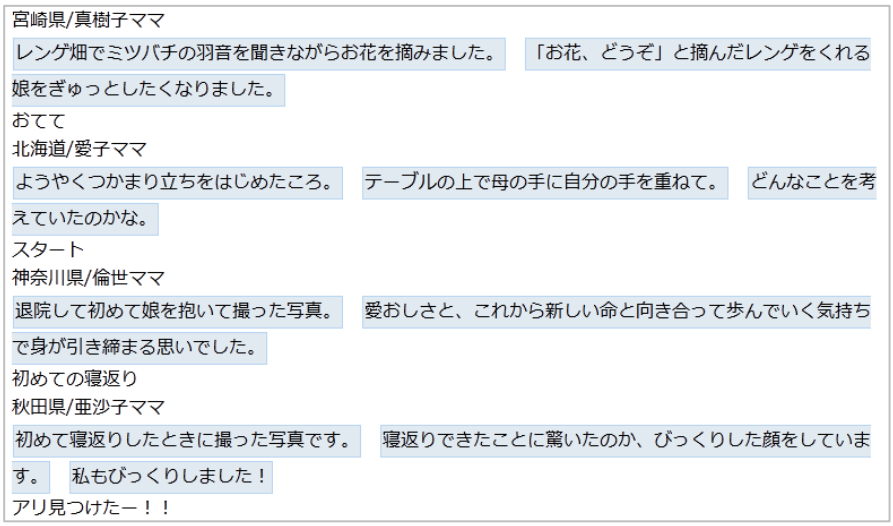

図 1 用例としての適正率

【重複する用例データの削除】一つ前の作業で、同一ページでは同じ用例が複数回採取されないようにしたが､6 億語弱のパイロット版 NLT を開発して実際に運用してみたところ、同一サイトで同一の用例が頻出することが確認された。そのため、URL の情報をもとに同一サイト[5]での同じ用例は一度だけ採取するように改良し、最終的に語数にして 11 億 3781 万語、用例数にして 4672 万 7 千例の筑波ウェブコーパスが完成した。

4．NINJAL-LWP への実装

図 2 NLT の見出し語ウィンドウ

[5] 正確には同一の FQDN（完全修飾ドメイン名）。

NINJAL-LWP は日本語コーパスの汎用的な検索システムである。2012 年の BCCWJ への実装（NLB）に続き、今回の TWC は 2 例目になる。図 2 は、動詞「走る」の見出し語画面である。NLB と同一のインターフェースなので、画面を左右に並べれば BCCWJ との比較が簡単にできる。

５．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。両者の頻度を対数変換してピアソンの相関係数を求めると 0.923 であり、NLB での動詞の分布と NLT での動詞の分布は極めて相似している。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。

スピアマンの順位相関は 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　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

TWC 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 TWC 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 TWC 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。

６．コロケーション

６．１　「～が走る」

「走る」のガ格に共起する名詞について、BCCWJ（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 TWC（NLT）における同様の共起頻度を求め、両者の順位相関は 0.585 となった。このことから、両者の相関はある程度あるものの、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、NLT の 5 億 8 千万語のパイロット版と今回の NLT の 11 億語版での順位相関は 0.973 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。

TWC 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である。

表２　TWC 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 など
(2)	人・動物 1896	人 251、馬 197、私 137、自分 121、～たち 112 など
(3)	痛み 1078	痛み 565、激痛 513
(4)	経路 616	道路 178、鉄道 128、道 109、～号線 66、線路 48 など
(5)	動揺・衝撃 473	衝撃 275、激震 81、戦慄 49、動揺 48、電撃 20
(6)	感覚 261	～感 81、悪寒 59、痺れ 38、寒気 36、感覚 25、震え 20
(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、～帯 25
(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。TWC 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 TWC 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 TWC 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。

６．２　「〜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、TWC 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　TWC における「〜を駆ける」の共起語

カテゴリ	共起語
空 95	空 61、大空 11、宇宙 8、天空 7、夜空 5、銀河 3
野山 59	草原 16、野 12、野山 11、山 7、森 6、原野 4、荒野 3
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、大陸 3
区域 21	街 8、庭 4、コート 3、町 3、街なか 3
前・後 21	先頭 6、先 5、前 5、後ろ 5
時 16	時代 10、時 6
海 11	海 8、海原 3

７．まとめ

本稿では筑波ウェブコーパス構築に当たり、BCCWJ の「均衡性」に近づけ、ウェブコーパスの弱点であるデータの偏りを回避する方略を提案した。また、NLB（NINJAL-LWP for BCCWJ）と同じレキシカルプロファイリング型のコーパス検索ツール NLT（NINJAL-LWP for Tsukuba Web Corpus）を使ってデータの抽出を行い、双方を比較した。動詞の頻度の比較では非常に高い相関が得られ、個々の動詞には両者の特徴が現れて違いが見られるところがあるものの、概ね両者の動詞の分布が近似していることが実証できた。また、コロケーションについては、データサイズの大きい筑波ウェブコーパスの方が安定的にコロケーション情報を抽出できることを示した。ただし、共起語出現語頻度が本稿で扱ったものより高いコロケーションの場合には､BCCWJ のサイズでも十分な情報が得られるであろうし、一方、共起語出現語頻度が本稿で扱ったものより低いものの場合には、筑波ウェブコーパスのサイズでもなお不十分ということも当然予想される。このような、より稀なコロケーション及びその他の稀なデータについてはさらに大きなサイズのコーパスが要求される。より大きなサイズのコーパスの構築においては、現実的に考えてウェブコーパスとならざるを得ないだろうから、今後の大規模コーパスの構築には本稿での知見が貢献できるとこ

ろも多いと思われる。

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エ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.

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.

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.

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.

関連 URL

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