Web を母集団とした超大規模コーパスの開発 —収集と組織化—

浅原　正幸（国立国語研究所コーパス開発センター）
今田　水穂（国立国語研究所コーパス開発センター）
保田　祥（国立国語研究所コーパス開発センター）
小西　光（国立国語研究所コーパス開発センター）
前川　喜久雄（国立国語研究所言語資源研究系・コーパス開発センター）

Early Results of Page Collection and Linguistics Annotation in Ultra Large Scale Web Corpus Construction

Masayuki ASAHARA (Center for Corpus Development, NINJAL)
Mizuho IMADA (Center for Corpus Development, NINJAL)
Sachi YASUDA (Center for Corpus Development, NINJAL)
Hikari KONISHI(Center for Corpus Development, NINJAL)
Kikuo MAEKAWA(Dept. of Corpus Studies, Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所では 2006～2010 年度の期間に 1 億語規模の書き言葉コーパス『現代日本語書き言葉均衡コーパス』（以下“BCCWJ”）（前川 2007, 前川・山崎 2008）を構築し，2011 年より一般公開した。BCCWJ は種々の母集団に沿った無作為抽出を実施することによって，高度な均衡性・代表性を備えた均衡コーパスとなっている。しかし，その規模は，現代のコーパス言語学の趨勢からすれば十分とはいいがたく，生起頻度の低い言語現象の被覆に問題がある。そのためより大規模な日本語コーパスの構築が望まれている。この問題を解消するため,国立国語研究所コーパス開発センターでは 2011 年度から 6 年の期間で，Web を母集団とした 100 億語規模の超大規模コーパスを構築する計画に着手した。

現在，6 年のプロジェクトの 3 年が過ぎ，収集と組織化について進捗している。またプロジェクトの初期において収集と組織化のための環境構築に工数をかけた。現時点の環境整備・収集・組織化についてはある程度目途がつきつつある。2012 年 10 月より本収集を開始し 2013 年 9 月まで 1 年間収集したデータについて組織化が進んでおり，基礎統計量が得られている。

本稿ではプロジェクトの進捗状況を過去 1 年間に収集されたデータの基礎統計量とともに示し，本言語資源を利用した言語研究の可能性について論じる。２節では収集の進捗状況について述べる。３節では組織化の進捗状況について述べる。４節では本言語資源とグーグルによる「Web 日本語 N グラム第 1 版」（1）との頻度上位 10 語の比較を定性的に分析する。５節で言語資源の大規模化による理論的・応用的研究の可能性について論じ，６節にまとめと今後の予定について述べる。

２．収集の進捗状況

本節では収集の経過と現在までに得られている統計量について示す。

2012 年 6 月に収集のための環境整備が完了し，環境整備に関して所内のセキュリティ委員会に諮り，クローラ運用の許諾を得た。許諾後すぐにクローラの試験運用 1 か月前よりクローラに関する情報提供･問い合わせ窓口としての Web ページ/メールアドレス/電話を設置した。1 か月の周知期間を経て，2012 年 7～8 月より試験収集を開始し，Heritrix（2）の各種パラメータを調整した。9 月に最終試験収集となる 1000 万 URL 規模の収集を行い，

その際に得られた収集速度の情報からクローラ運用方針を年間 4 回収集，1 回あたりの収集量は 1 億 URL 規模で 3 か月間に設定した。URL は 1 年間通して 4 回収集し，季節に偏らないように配慮する一方，URL の更新頻度およびリンク先の情報に基づき，収集する URL を 1 年毎に変更する方針にした。2012 年 10 月より 3 カ月ごとに収集を行い，2013 年 12 月末現在, 5 回目の収集の最中で次の 1 年間に収集すべき URL サンプルを決定する状況にある。

以下収集の統計量について示す。全体については 2012 年第 4 四半期（2012-4Q）～2013 年第 3 四半期（2013-3Q）についての統計量を，各論については 2012 年第 4 四半期のみの統計量を示す。

表 1 ：2012 年第 4 四半期から 2013 年第 3 四半期の収集ページ数

	2012-4Q	**2013-1Q**	**2013-2Q**	**2013-3Q**
ページ数　（1 期）	61,668,805	58,844,092	61,479,268	57,892,917
内容の重複なしページ数	45,933,605	42,932,982	45,111,527	42,192,931
4 期通しての統計				
総異なり URL 数（4 期）	64,539,233			
(内) 内容の重複ありページ数	27,604,915			
(内) 内容の重複なしページ数	36,934,706			

表 1 に収集したページ数の統計量を示す。1 億 URL を収集しても robots.txt の順守や各種 HTTP エラーにより，ページとして収集できたものが約六割にすぎない。重複検出は URL 毎に各ページのハッシュ値を計算し同一性を認定する。各期において内容の重複なし（異なり）ページ数は 4000 万強になる。4 期通しての総異なり URL 数は約 6400 万 URL と 1 億 URL に至らない。4 期中 2 期以上収集できたページ数の内，内容の重複があるページ数は約四割の 2700 万ページ，反対に内容の重複がないページ数は 3600 万ページになる。

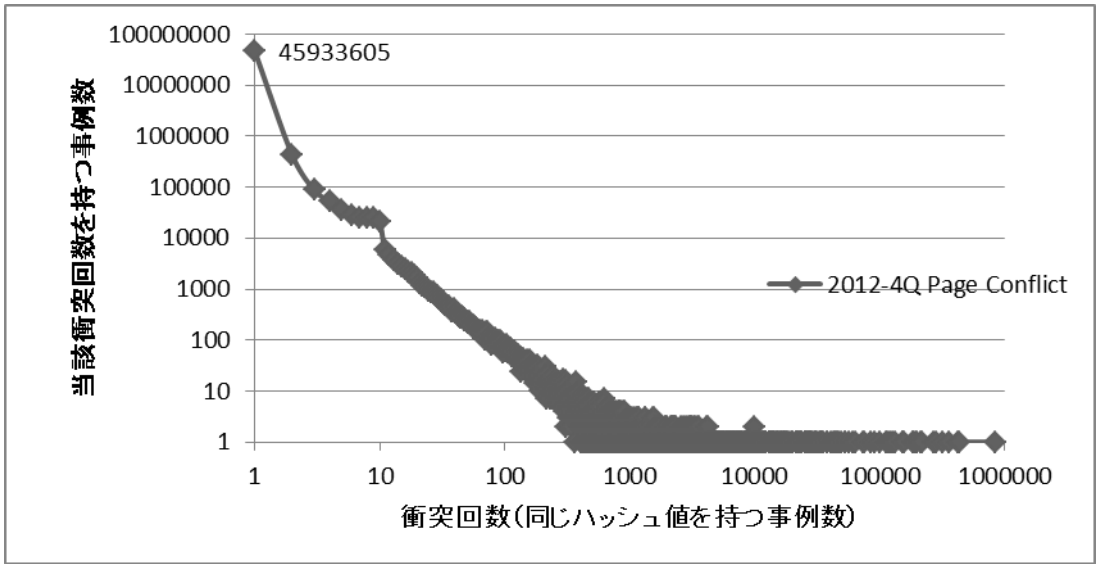

図 1 ：2012 年第 4 四半期収集ページの重複

図 1 に 2012 年第 4 四半期収集 Web ページの重複検出結果について示す。同じハッシュ値を持つ URL が複数存在することを「衝突」と呼ぶ。グラフ横軸は同じハッシュ値を持つ URL 数を示し，「衝突回数」と呼ぶ。グラフ縦軸は横軸の「衝突回数」を持つ衝突事例数（URL 数ではなくハッシュ値の異なり数で計算）を示す。グラフは両軸とも対数で表示している。グラフ中の左上の点が表の「内容の重複なしページ数」（他の URL と内容が重複しないページ数）に相当する。衝突回数 10 以下のものは同一内容の異なる URL 表示もしくはいわゆるコピーサイトであると考える。それ以上の衝突については，robots.txt や，「ソフト 404」

と呼ばれる当該 URL はサーバ上にないということを 404 HTTP ステータスコードでは返さず 200 HTTP ステータスコードで返し当該ページがないことを示すコンテンツを返すページである。

表２：2012 年第 4 四半期から 2013 年第 3 四半期の収集リンク数

	2012-4Q	**2013-1Q**	**2013-2Q**	**2013-3Q**
リンク先（のべ）	6,905,805,383	6,610,763,700	7,064,611,259	7,222,958,033
	69 億 URL	66 億 URL	70 億 URL	72 億 URL
リンク先（異なり）	892,135,930	843,166,672	865,694,816	855,684,918
	8.9 億 URL	8.4 億 URL	8.6 億 URL	8.5 億 URL
4 期通しての統計				
リンク先（異なり）	1,642,699,579			
	16 億 URL			

表２に 2012 年第 4 四半期（2012-4Q）～2013 年第 3 四半期（2013-3Q）の収集リンク数を示す。おおよそ 6000 万 URL の収集に対し，のべ 70 億前後，異なり 9 億弱の URL が収集できている。4 期を通した集計によるリンク先数が異なり 16 億 URL であることから 1 年間通して同じ URL を 4 期収集することにより 1 期のみクロールするのみ比べてリンクが約 1.8 倍（8.5 億～8.9 億→16 億 URL）に成長していることがわかる。

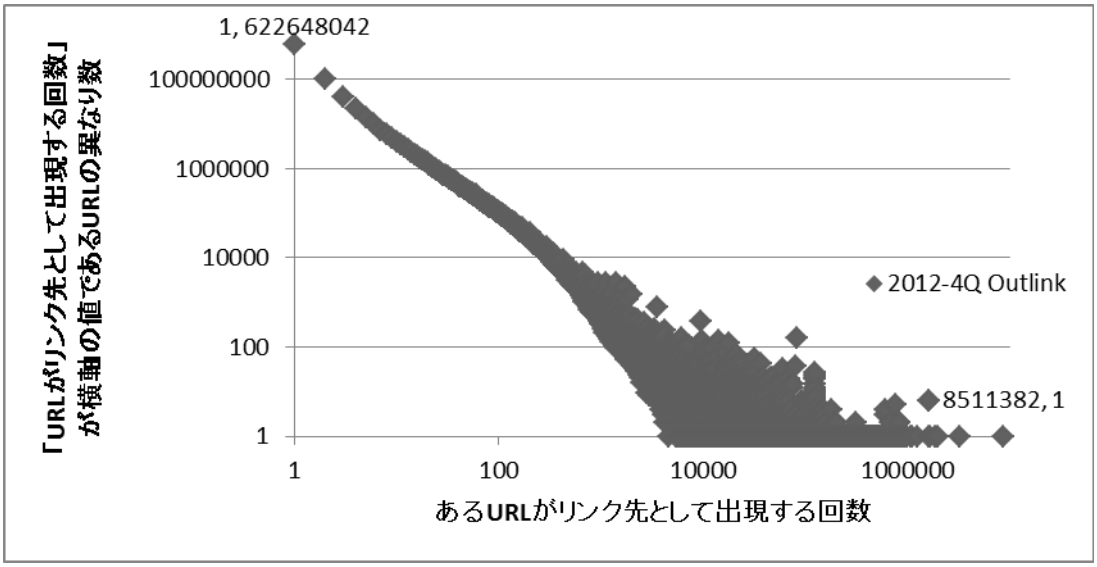

図２：2012 年第 4 四半期収集リンク先の統計

図２に 2012 年第 4 四半期の収集リンク先統計を示す。グラフ横軸はある URL がリンク先として出現する回数（被リンク数）で，グラフ縦軸が横軸相当の被リンク数を持つ URL の異なり数である。グラフの両軸は対数である。グラフ左上が 1 回しかリンクされていない URL 数で約 6 億件である。グラフ右下が最も多い被リンク数約 850 万を持つページで 1 件ある。これは有名ブログサイトのトップページであり，ブログの個々のページからリンクされている。

これらの統計情報から 2014 年以降の収集対象 URL を決定する。収集対象 URL は，4 期にクロールした同一 URL のうち内容が毎回変わっていた URL と，収集した Web ページのリンク先 URL の二種類を想定している。現在までに収集した Web ページに対するレジスタ分析は進んでいないが，レジスタ分析が進み次第，レジスタ分析結果の分散を見ながら収集対象 URL を決定したい。

３．組織化の進捗状況

本節では組織化の経過と現在までに得られている統計量について示す。

現在までのところ，正規化・形態素解析・係り受け解析までが部分的に進捗している。正規化においては nwc-toolkit による文字コード統制・文抽出を行った。利用するライブラリの関係でサーバ群ではなく安価な 8-core のワークステーション上で正規化作業を行った。8-core の CPU の並列処理により約 1 週間で 1 年分の収集データの正規化作業が完了する。正規化されたデータは MeCab/IPADIC，MeCab/UniDic，JUMAN により形態素解析を行う。32-core の計算サーバ上の並列処理により，それぞれ 1 日弱で 1 年分の収集データの形態素解析作業が完了する。さらに IPA 品詞体系に基づく CaboCha と益岡・田窪品詞体系に基づく KNP により係り受け解析を行う。32-core の計算サーバ上の並列処理により 1〜2 週間で係り受け解析作業が完了する。

表３に組織化したデータの基礎統計量を示す。Heritrix は収集 Web ページを圧縮 1GB サイズの WARC データに分割して出力する。展開すると約 3 倍程度になるため，表中の収集 WARC ファイル数に 3GB をかけた値が収集 Web ページ容量（メタデータを含む）と概算することができる。URL 数は前節の収集における URL 数である。正規化処理は nwc-toolkit による。正規化処理の際に文抽出なしに形態素解析（MeCab/IPADIC）を行うと各期のべ約 620〜647 億形態素になる。文抽出を行うと形態素数は各期約 300 億強になることから大体半分の形態素が日本語の文中の形態素ではないとして排除されている。抽出された文数はのべ数で各期 25 億文前後，文単位の同一性を認定すると文の異なり数は各期 10 億文になる。

表３：組織化したデータの基礎統計量

	2012-4Q	2013-1Q	2013-2Q	2013-3Q
収集 WARC ファイル	814	870	910	905
URL 数	61,668,805	58,844,092	61,479,268	57,892,917
形態素数（文抽出なし）	64,714,650,129 647 億形態素	62,077,520,745 620 億形態素	63,414,252,638 634 億形態素	65,736,027,334 647 億形態素
形態素数（文抽出あり）	33,767,409,441 337 億形態素	32,651,138,004 326 億形態素	33,073,991,355 330 億形態素	30,923,912,566 309 億形態素
文数（のべ数）	2,678,315,774 26 億文	2,600,122,908 26 億文	2,659,617,620 26 億文	2,478,309,312 24 億文
文数（異なり数）	1,097,011,506 10 億文	1,048,772,913 10 億文	1,063,649,324 10 億文	1,007,771,383 10 億文

図３に 2012 年第 4 四半期の収集文の重複を示す。横軸が同一文の出現回数で縦軸が当該出現回数の文の異なり数を表す。両軸とも対数で表現する。10 億文のうち約 9 割の 8.9 億文が 1 回しか出現しない文である。以下，ページの重複も含めて同一文が異なりで 1 億文規模存在する。これらは定型的な表現やリンクの見出し語であることが多く，一番多く出現した文は 2,885,654 回出現する「職業とキャリア」（Yahoo!知恵袋のカテゴリ名）であった。ここで文抽出処理の妥当性について検証する。検証データとして BCCWJ に nwc-toolkit を適用して文数とバイト数がどの程度変化するかについて示す。

表５に nwc-toolkit のうち文抽出を行う nwc-toolkit-text-filter を BCCWJ に対して適用した場合のジャンル毎の文数とバイト数の変化を示す。表中左の文数はそれぞれの文境界基準により認定しており，OC（Yahoo!知恵袋）のように句点・感嘆符・疑問符が文境界として利用されない場合が多いジャンルについては nwc-toolkit により文数が増える場合があり，単純にどの程度削減されるかは評価できない。表中右のバイト数が純粋にデータとして削

減されたテキスト量を表す。バイト数に関して変化率が低い（多く削除されている）ジャンルは OP（広報紙）と PM（雑誌）であった。これらに対し，何が削除されているのかを確認したところ，人口などの統計情報・イベントの日時・窓口やお店の営業時間・連絡先（電話番号・住所）などが削除されていることがわかった。情報抽出の分野においては重要な情報ではあるが，言語研究においては削除しても多くの研究で影響は少ないと考える。

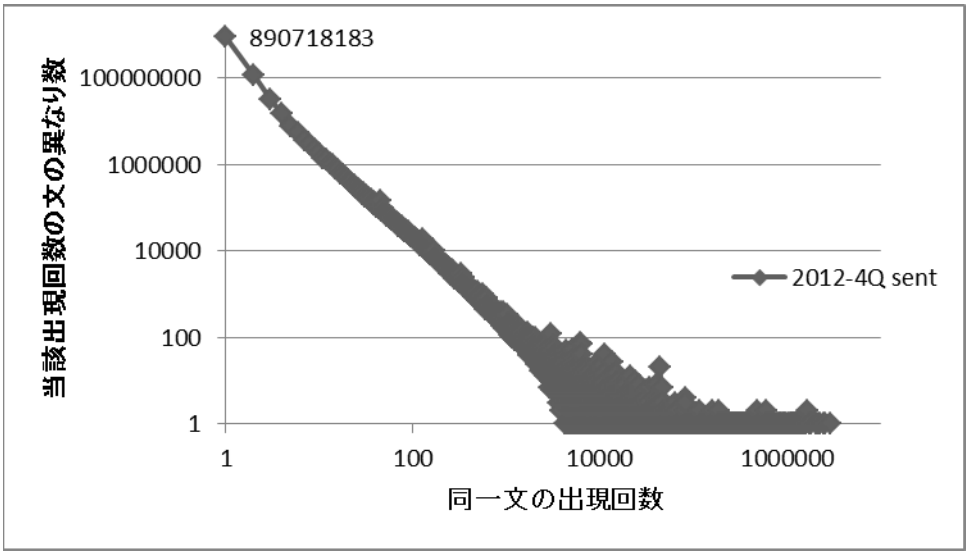

図３：2012 年第 4 四半期収集文の同一性

表４：BCCWJ の nwc-toolkit-text-filter の適用による変化

ジャンル名	文数			バイト数		
	処理前	処理後	変化率	処理前	処理後	変化率
LB	451,273	394,782	0.87	186,908	176,792	0.95
OB	222,437	203,467	0.91	23,236	22,242	0.96
OC	42,506	45,082	1.06	369,004	368,572	1.00
OL	38,827	34,761	0.90	6,004	5,864	0.98
OM	140,422	116,295	0.83	26,948	26,052	0.97
OP	257,796	142,660	0.55	22,372	14,636	0.65
OT	64,100	47,680	0.74	5,992	5,340	0.89
OV	18,977	17,807	0.94	1,656	1,632	0.99
OW	146,402	126,140	0.86	28,072	24,876	0.89
OY	117,816	76,205	0.65	224,064	213,480	0.95
PB	392,301	308,827	0.79	176,812	163,476	0.92
PM	301,399	227,089	0.75	30,248	25,812	0.85
PN	80,563	65,055	0.81	10,164	9,152	0.90
合計	2,274,819	1,805,850	0.79	1,111,480	1,057,926	0.95

現在収集しているコーパスデータについて語彙表の作成や n-gram データを進めている。unix コマンド sort, uniq, wc などに基づく単純な集計手法では sort アルゴリズムや途中で出力される中間ファイルにより時間計算量・空間計算量が想定よりも大きく時間がかかることがわかった。現在様々なソフトウェアを組み合わせて効率化をはかっており，2012 年第 4 四半期の n-gram 統計のみ得つつある。

４．本言語資源とグーグル「Web 日本語 N グラム」との比較

最後に 2012 年第 4 四半期収集データとグーグル「Web 日本語 N グラム」との比較を行う。

表５：2012 年第 4 四半期収集データとグーグル「Web 日本語 N グラム」との比較

	本言語資源(2012-4Q) MeCab/IPADIC 頻度 3 以上（文抽出有） 文単位での重複性排除有	本言語資源(2012-4Q) MeCab/IPADIC 頻度 3 以上（文抽出有） 文単位での重複性排除無	「Web 日本語 N グラム」 MeCab/IPADIC 頻度 20 以上 [工藤・賀沢 2007]
総形態素数 （のべ）	18,075,529,620 180 億形態素	33,767,409,441 337 億形態素	255,198,240,937 2550 億形態素
総文数	1,097,011,506 （異なり） 10 億文	2,678,315,774 （のべ） 26 億文	20,036,793,177 200 億文
1-gram	0.039 億	0.050 億	0.025 億
2-gram	0.47 億	0.85 億	0.80 億
3-gram	1.6 億	4.4 億	3.9 億
4-gram	2.1 億	8.7 億	7.0 億
5-gram	1.7 億	10.3 億	7.7 億
6-gram	1.2 億	9.7 億	6.8 億
7-gram	0.84 億	8.5 億	5.7 億

表 5 に 2012 年第 4 四半期収集データと 2007 年公開のグーグル「Web 日本語 N グラム」の比較結果を示す。「Web 日本語 N グラム」では頻度 20 以上の n-gram データのみを配布している。本研究ではできるだけ低頻度のデータを被覆するべく適切な頻度を検討中であるが，速報値として頻度 3 以上の概算値を報告する。総文数においては「Web 日本語 N グラム」の規模の 20 分の 1 から 10 分の 1 くらいの規模である。しかし低頻度のものも組織化することにより n-gram データの被覆では遜色ないレベルにできると考える。

表 6 ・表 7 は 2012 年第 4 四半期収集データとグーグル「Web 日本語 N グラム」の頻度順位上位 10 件を比較するものである。頻度順位において，文境界メタ記号や句読点・記号類を含む n-gram は排除している。これらより各 n-gram の定性的な分析を行う。

まず本言語資源と「Web 日本語 N グラム」とで標本の量を比較すると 10 倍（文単位での重複性排除無）～20 倍（文単位での重複性排除有）くらい「Web 日本語 N グラム」の方が規模が大きい。単純に収集規模が小さい本言語資源において文単位の重複性排除無で n-gram を取ると 3-gram や 4-gram のレベルで複数のページで出現する定型的な表現が頻度順位上位にきてしまう。2-gram のレベルでも「利用規約」という定型的な表現が確認された。一方，文単位での重複性排除有の本言語資源では，規模が小さくても「Web 日本語 N グラム」と同様に機能語中心の n-gram が得られる。このように文単位での重複性排除を行うことにより Web 特有の定型的な表現の頻度の偏りをなくすことができ，定型的な表現の影響が少なくなる，より規模の大きな言語資源と同じ傾向のデータが得られていることがわかる。

5-gram 以上になると「Web 日本語 N グラム」においても定型的な表現の影響が避けられず，多くみられるようになる。一方，文単位での重複性を排除したものについては，文の一部を変えるような定型的な表現（「タグ が 付け られ た 質問」「に関する ウェブ 上 の 情報 を 探す」などは見られるが，基本的には機能表現中心のコロケーションが得られていることがわかる。

特殊な事例として一文中に同じ文字が連続している表現がある。「ああ ああ ああ ああ

ああ ああ ああ」や「え え え え え え え」などは，一文中に同じ文字が連続している表現の形態素解析結果である。例えば「え」が p 回出現するような一文に対して，p-6 回の「え え え え え え え」7-gram が枚挙される。一文中に複数の n-gram が出現する際にどのように数え上げるかは議論の余地がある。これは構造をもつデータに対する問い合わせにおいて起きる常に問題で，現在は多くのコンコーダンサにおいて正規化は行っていない。

表６：2012 年第 4 四半期収集データとグーグル「Web 日本語 N グラム」との比較
頻度順位上位 10 件（1-gram～4-gram）

	順位	1-gram	2-gram	3-gram	4-gram
本言語資源	1	の	し て	て い ます	し て い ます
MeCab/IPADIC	2	に	まし た	て い た	て い まし た
	3	て	て い	し て い	さ れ て いる
2012-4Q	4	が	て いる	し て いる	し て い た
文単位での重複性排除有	5	は	し た	と 思い ます	さ れ て い
	6	を	で は	さ れ て	た の です が
	7	た	に は	に なっ て	て き まし た
	8	で	さ れ	の です が	れ て い ます
	9	と	ませ ん	し まし た	は あり ませ ん
	10	し	い ます	さ れ た	に なり まし た
本言語資源	1	の	まし た	記事 へ の	記事 へ の トラック
MeCab/IPADIC	2	に	でしょ う	お願い し ます	専用 ページ を 表示
	3	を	行っ て	Q & A	利用 する こと が
2012-4Q	4	は	思っ て	続き を 読む	機能 を 利用 する
文単位での重複性排除無	5	て	情報 を	マーク へ 投稿	おすすめ の 知恵 ノート
	6	が	利用 規約	専用 ページ を	正確 性 の 保証
	7	た	おすすめ の	機能 を 利用	お客様 自身 の 責任
	8	で	記事 へ	済み の 質問	回答 を 指示 する
	9	と	追加 する	おすすめ の 知恵	便利 に 新規 取得
	10	し	場合 は	エンターテインメント と 趣味	はてな ブック マーク へ
「Web 日本語 N グラム」	1	の	し て	て い ます	し て い ます
MeCab/IPADIC	2	に	まし た	し て い	さ れ て いる
	3	を	て い	て い た	さ れ て い
[工藤・賀沢 2007]	4	は	て いる	し て いる	は あり ませ ん
	5	て	し た	さ れ て	れ て い ます
	6	が	ませ ん	に なっ て	て い まし た
	7	た	さ れ	し まし た	に なり まし た
	8	で	に は	さ れ た	し て おり ます
	9	と	で は	れ て いる	て き まし た
	10	し	い ます	あり ま せん	し て い た

例えば，「N1 が N2 が V1 する N3 を V2」といった文に対して「『N が』が先行する『V』」といった問い合わせを行う際に何件出現すると言えるだろうか？多くのコンコーダンサは <N1 が,V1>, <N1 が,V2>,<N2 が,V1>,<N2 が V2>の全組み合わせを出力し 4 件と出力するだろう。

このように n-gram を含めて 2 語以上の組み合わせを行う際には，重複して枚挙される現象に注意する必要がある。さらに，係り受け構造など構造を持ったデータに対する部分構造の枚挙に際しては，同様な事例が多く出現する。このような部分構造を共有する事例に

ついて，n-gram データとして正規化することは一切行わないが，利用者においては統計処理を行う場合に必要に応じて対処する必要がある。

本 n-gram データについては近いうちにデータそのものを公開するとともに，簡単な検索環境を Web API つきで対外的に公開し利用可能なようにする。2014 年度以降進めていく利用者系に関する意見などをいただければ幸いである。

表 7 ：2012 年第 4 四半期収集データとグーグル「Web 日本語 N グラム」との比較
頻度順位上位 10 件（5-gram～7-gram）

	順位	5-gram	6-gram	7-gram
本言語資源	1	さ れ て い ます	で は な い でしょ う か	の で は な い でしょ う か
MeCab/IPADIC	2	で は あり ませ ん	て い た の です が	の タグ が 付け られ た 質問
	3	と 思っ て い ます	の で は な い でしょ う	で は な い か と 思い ます
2012-4Q	4	し て い まし た	の で は な い か と	に関する ウェブ 上 の 情報 を 探す
文単位での重複性排除有	5	で は な い で しょう	に 行っ て き まし た	ああ ああ ああ ああ ああ ああ ああ
	6	の で は な い か	よう な 気 が し ます	の で は な い か と 思い
	7	は な い でしょ う か	タグ が 付け られ た 質問	し て いた の です が
	8	に なっ て い ます	の タグ が 付け られ た	思っ て い た の です が
	9	て い まし た が	さ せ て いただき まし た	え え え え え え え え
	10	て い た の です	たい と 思っ て い ます	と 思っ て い た の です
本言語資源	1	記事 へ の トラック バック	機能 を 利用 する こと が	機能 を 利用 する こと が でき
MeCab/IPADIC	2	機能 を 利用 する こと	利用 する こと が でき ませ	利用 する こと が でき ませ ん
	3	利用 する こと が でき	正確 性 を 保証 し て	正確 性 を 保証 し て おり
2012-4Q	4	正確 性 を 保証 し	お客様 自身 の 責任 と 判断	お客様 自身 の 責任 と 判断 で
文単位での重複性排除無	5	お客様 自身 の 責任 と	すべて の 機能 を 利用 する	すべて の 機能 を 利用 する こと
	6	はてな ブック マーク へ 投稿	知恵袋 の すべて の 機能 を	知恵袋 の すべて の 機能 を 利用
	7	更新 情報 が 届き ます	おすすめ の 解決 済み の 質問	ニックネーム の My 知恵袋 で 確認 でき
	8	おすすめ の 解決 済み の	記事 へ の トラック バック URL	質問 年月 や 画像 の 有無 を
	9	すべて の 機能 を 利用	ニックネーム の My 知恵袋 で 確認	質問 や 知恵 ノート は 選択 さ
	10	質問 年月 や 画像 の	する こと が でき ませ ん	以上 更新 が ない ブログ に 表示
「Web 日本語 N グラム」	1	さ れ て い ます	無料 で お 届け し ます	料 無料 で お 届け し ます
MeCab/IPADIC	2	で は あり ませ ん	料 無料 で お 届け し	配送 料 無料 で お 届け し
	3	で お 届け し ます	配送 料 無料 で お 届け	国内 配送 料 無料 で お 届け
[工藤・賀沢 2007]	4	無料 で お 届け し	国内 配送 料 無料 で お	以上 国内 配送 料 無料 で お
	5	1500 円 以上 国内 配送	円 以上 国内 配送 料 無料	円 以上 国内 配送 料 無料 で
	6	料 無料 で お 届け	以上 国内 配送 料 無料 で	1500 円 以上 国内 配送 料 無料
	7	配送 料 無料 で お	1500 円 以上 国内 配送 料	は インライン フレーム を 使用 し て
	8	国内 配送 料 無料 で	を 使用 し て い ます	フレーム を 使用 し て い ます
	9	以上 国内 配送 料 無料	インライン フレーム を 使用 し て	インライン フレーム を 使用 し て い
	10	円 以上 国内 配送 料	この 記事 へ の トラック バック	部分 は インライン フレーム を 使用 し

５．本言語資源を利用した理論的・応用的研究の可能性

本節では本言語資源を利用した理論的・応用的研究の展望について述べる。

まず語彙研究のために Web を母集団とするコーパスに対してどのような統計処理をすればよいかについて述べる。コーパスに基づく語彙研究は大きく分けて，生コーパスから得られる統計情報のみから語彙の性質を明らかにする教師なし統計学習に基づく手法と，何らかのアノテーションを付与してその情報を未知データに再現するような教師あり統計学習に基づく手法の二つが考えられる。

前者の教師なし統計学習においては，近年ベイズ統計学が注目されトピックモデルに基づく手法が盛んに利用されている。コーパスを用いてトピックモデルを構築する際に「文書」という単位と「単語」という単位が重要になる。本言語資源においては「文書」が個々の Web ページ，「単語」が形態素解析により認定された形態素が対応する単位になる。文書内の単語の共起などに基づき潜在変数を推定する。単語の与え方によっては，トピックを文書に遍在するレジスタとみなしレジスタ分析にも利用することができる。ここで本研究では「文」という単位を重要視する。Web の場合，コピーサイトや複数のページで共有する部分ページなどがあり，単純に「文書×単語関係」のみにより表現するとこれらの影響を受けてしまう。３節に示した統計量のように約 1 割の 1 億文については複数の文書に共有されている。そこで文書と単語の間に文という単位を仮定し，「文書×文関係」「文×単語関係」と階層的にモデル化することにより，コピーサイトなどの影響を前者の「文書×文関係」で吸収しながらレジスタをモデル化することが可能になると考える。さらに Web の場合には「文書×文書間関係」としてリンク―被リンク関係が規定されている。単リンクが張られる関係，もしくは同一文書にリンクを張る関係など，レジスタに影響を与える関係がいろいろ考えられる。この「文書×文書間関係」「文書×文関係」「文×単語関係」の多層の共起関係を，現実的に処理可能な計算量でどのようにモデル化するかを検討する。

後者の教師あり統計学習においては，古くは二値分類器が中心に用いられたが，最近では，順位・極性・系列などといった構造学習が自然言語処理の分野で広く用いられている。また特徴量としても木構造や有向非循環グラフなどといったものの部分構造の重複などを効率的に枚挙しながら距離（類似度）を設定して識別する学習手法が多く提案されている。本言語資源では形態論情報だけでなく，係り受け構造や述語項構造などといった情報を付与する。これにより言語研究の分野における構造の利用（特徴量としての利用と推定する対象としての利用）が促進されればと考えている。

利用者系において，３節で通常の n-gram データと文単位での重複を排除した n-gram の二つについての統計量を提示した。n-gram 検索ツールについては，この双方を提供して利用者による評価を待ちたい。また KWIC 検索においては後者の文単位での重複を排除したコーパスに対する検索ツールの開発のみを検討している。３節や４節に示すような様々な情報を提供しながら文単位での重複性の影響について議論していきたい。

形態論研究においては，Web 上の多様な語彙を観察することが可能になるだけでなく，保存する予定であるデータを通時的に観察することで経年変化を分析することが可能になる。Web の世界においては多様な形態論変化が生産され，すぐに廃れていく傾向にある。これらを適切にとらえることができるような環境を構築できればと考えている。

コーパス日本語学において，統語論研究に必要な情報が現状のコーパスから得られているとは言い難い状況にある。まず BCCWJ では品詞情報や係り受け構造に基づく問い合わせをおこなったとしても真に分析したい言語現象が数例しかないということが起きうる。例えばガ格が二つ以上出現する文（複文）の従属節境界の左端のあいまい性（pre-head attachment; Kamide and Mitchel 1999）を研究する際に，BCCWJ に様々な構造を問い合わせてみると「N ガ N ヲ VP N ガ」という語順が 1 例，「N ガ ADV N ガ」という語順が 3 例，「N ガ N ニ N ガ」という語順が 1 例しか出現しないことがわかる。規模を 100 倍にすることでこういった言語現象を発見する可能性は高くなる。また人文系の研究者側の問題として，多くの研究が語彙情報を単純なベクトル表示にする bag-of-words 的な共起以上の情報を使いこなせていないという問題がある。自動解析ながらもあらかじめ係り受け構造等を付与し，使いやすい利用者系を構築することで，語順や部分木構造などを考慮したコーパス分析手法が根付くような土壌を整備していきたい。

意味論研究においても，頻度情報によらない研究にまで踏み込むことが重要であると考える。この分野においては，共起に基づく研究は盛んに行われている一方，アノテーションを行い，さらに構造学習を導入してアノテーションを未知データに復元することで何かを明らかにするような水準に達していないと考える。ここで重要なのは，解明しようとし

ている研究課題が本質的にコーパスを用いる手法が適しているかどうかを見極める必要があるということである。コーパスを用いる手法が何も証拠を提供できない問題は，言語研究において多々あると考える。そういった問題に対しても，コーパスが仮説を立てるための手がかりを研究者に与えることは可能な場合が多いが，手がかりのみをもって証拠とすることのないように注意されたい。

６．おわりに

本論文では国立国語研究所コーパス開発センターで進めている Web を母集団とした超大規模コーパス開発プロジェクトの進捗について報告した。6 年間のプロジェクトのうち半分の 3 年が経ち，単純に 100 億語規模のコーパスを構築するだけでなく，継続的に同規模のスナップショット的なコーパスを 1 年間に複数回構築可能である環境が構築されつつある。収集においては Heritrix クローラを利用して年間 4 回のクロールを行い，組織化においては正規化・形態素解析・係り受け解析まで進捗している。組織化における残る課題は述語項構造解析とレジスタ分析と語彙表構築であるが，これについても早急に環境を整えたい。得られた基礎統計量から 3 か月単位で目標の 100 億語規模のコーパスが構築できていることがわかった。今後より一層の効率化を進めたい。

残りの期間で人文系の研究者が柔軟に利用可能な利用者系と保存環境の構築を行う。語彙調査や用例検索に留まらない，自然言語処理で培われた構造に基づく問い合わせ環境を構築したい。これにより高い被覆性を持ちながらも柔軟なコーパス調査を可能にし，統語論・意味論研究を前に進める研究環境を提供できると考える。

一方で，本言語資源に基づく調査で解明できない問題が何であるのかを示すが重要だと考える。規模を大きくすることだけでは解明できない問題についても示していきたい。

謝　辞

本研究は国語研「超大規模コーパス構築プロジェクト」によるものです。

文　献

Kamide, Yuki and D. C. Mitchell （1999） Incremental Pre-head Attachment in Japanese Parsing, *Language and Cognitive Processes,* 14(5/6), 631-662.
工藤拓・賀沢秀人（2007）『Web 日本語 N グラム第 1 版』，言語資源協会発行 (http://www.gsk.or.jp/catalog/GSK2007-C/).
前川喜久雄（2007）「コーパス日本語学の可能性–大規模均衡コーパスがもたらすもの–」日本語科学, 22, 13-28.
前川喜久雄・山崎誠（2008）「『現代日本語書き言葉均衡コーパス』」国文学解釈と鑑賞, 932（74 巻 1 号）, 15-25.
矢田晋（2010）.『日本語ウェブコーパス 2010（NWC 2010）』, (http://s-yata.jp/corpus/).

関連 URL

（1） Web 日本語 N グラム第 1 版：(http://www.gsk.or.jp/catalog/GSK2007-C/).
（2） クローラ Heritrix-3.1.1：(http://webarchive.jira.com/wiki/display/Heritrix/Heritrix).
（3） 日本語ウェブコーパス用ツールキット：(http://code.google.com/p/nwc-toolkit/).
（4） 形態素解析器 MeCab-0.996：(http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html).