Vol.2014-DBS-159 No.15
Vol.2014-IFAT-115 No.15
2014/8/2

言語研究のためのWebコーパスの収集と組織化

浅原 正幸[1,a)]　今田 水穂[2]　保田 祥[1]　小西 光[1]　前川 喜久雄[1]

概要：国立国語研究所コーパス開発センターでは 2011 年より超大規模コーパスプロジェクトとして、Web を母集団とした 100 億語規模のコーパスの構築を進めている。構築にあたっては、工程を収集・組織化・利活用・保存の四つに分割して構築を進めている。本稿ではそのうち最初の二工程について言語研究に資する言語資源にするために行っている工夫について報告する。

キーワード：言語資源構築，Web アーカイブ，言語情報組織化

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)[1] を構築し，2011 年に一般公開した。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。

Web を母集団とした言語資源を構築するためには，何らかの方法で単リンクからなるハイパーテキスト空間から，標本空間として規定すべき範囲を決め，その範囲で得られる標本を収集することが必要になる。本研究では標本を「日本語」の Web テキストとする。ここで「日本語」とは，日本語母語話者が生産した誤りのない狭義の日本語ではなく，非母語話者や機械 (機械翻訳や自動要約など) が生産する日本語についても，日本語母語話者が受容する機会がある日本語であれば広義の日本語として標本に含める。Web テキストは，HTML ファイルや TXT ファイルだけでなく，PDF ファイル，Postscript ファイル，swf ファイル，Microsoft Word/Excel/PowerPoint ファイルなどにも埋め込まれている。本研究では，次に述べる組織化が容易な HTML ファイルと TXT ファイルのみを対象にする。

収集された HTML ファイルと TXT ファイルはそのままでは言語資源として扱いにくい。利用者にとって扱いやすい言語資源にするために組織化を行う。本研究は言語研究に資する組織化として，収集データの正規化・形態論から浅い統語論レベルの言語構造アノテーション・メタデータに相当するレジスタ分析を行う。

Web を母集団にした言語資源は後に示すように様々な言語研究を目的として多くの先行研究がある。それらの多くが収集・組織化まではなされているが，人文系の研究者にとって利活用しやすい利用者系までは提供されていないことが多い。また BCCWJ を除く過去の言語資源開発プロジェクトにおいては，言語資源構築までを目的として，利用者系まで整備しないことが多い。本研究ではプロジェクトの前半までに収集・組織化を軌道にのせ，プロジェクトの後半では人文系の研究者にとって利活用しやすい利用者系の整備を進めるとともに，組織化の精緻化にも努める。

人文情報学の分野では有形・無形の文化資源等を電子化して保存するデジタルアーカイブの研究が盛んに進められている。デジタルアーカイブの一分野として Web ページの保存を目的とする Web アーカイブの研究が各国の国立図書館が中心となり進められている。本研究では Web 上に生産される多様な言語現象の通時的な研究を将来可能にするために，Web コーパスの保存についても検討する。

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

[1] 人間文化研究機構 国立国語研究所
NINJAL, Tachikawa, Tokyo 190–8561, Japan
[2] 文部科学省
MEXT, Chiyoda, Tokyo 100–8959, Japan
[a)] masayu-a@ninjal.ac.jp

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における収集と組織化の概要とともに得られている統計について報告する。

本稿の構成は以下の通りである。まず 2 節では企業，研究機関，大学，官公庁，個人などで進められてきた関連研究について示す。3 節では，各種の既存技術を組み合わせて，いかにして超大規模コーパスを構築するか，収集・組織化の二つの観点について示す。4 節では，過去 1 年間に収集されたデータの基礎統計量について示し，最後に 5 節で現時点でのまとめと今後の展開について示す。

2. 関連研究

Web を母集団とした言語資源として，クローラを利用して検索エンジンを運営している企業や掲示板・Web サイトをホストしている企業により提供されている語彙表や n-gram 統計情報がある。本節では一般に入手利用可能な言語資源を中心に関連研究を示す。

グーグルは「Web 日本語 N グラム第 1 版」[2] として，元データ 2550 億語/200 億文規模の語彙表・n-gram データを作成し，一般公開した。バイドゥ株式会社 [3] は 2000〜2010 年にかけてのブログや掲示板のデータ 1000 万文を対象に，月毎のコーパス母集団を元に作成した「Baidu ブログ・掲示板時間軸コーパス」の語彙表・n-gram 統計情報を公開した。また，同時期にバイドゥ株式会社 [4] はモバイル検索向けに収集した Web データを元に作成した「Baidu 絵文字入りモバイルウェブコーパス」の語彙表・n-gram 統計情報も公開した。楽天技術研究所は 2010 年より「楽天データセット」としてレビューデータなどを公開している [5]。ヤフー株式会社は「Yahoo!知恵袋」コーパス (2004 年 4 月〜2009 年 4 月)[6] を公開している。

研究機関等においては，情報通信研究機構 (NICT)・京都大学などがそれぞれクローラを用いて Web アーカイブを構築し，整形したデータを一般公開している。例えば，NICT は検索エンジン基盤 TSUBAKI[7] を構築し，約 345GB(非圧縮) 規模の日本語係り受けデータベース [8] を公開した。京都大学は Web データ 16 億文を用いて自動構築した格フレームを公開した [9]。これら二つのデータは形態素解析のみならず係り受け解析や格解析までの処理が行われている。官公庁においては，国立国会図書館 (NDL) が官公庁自治体の Web サイトや冊子体から電子版に移行した雑誌の保存を目的として，インターネット資料収集保存事業 [10] を 2006 年より本格事業化している。NDL の Web アーカイブでは保存が主目的であり，同一 URL を複数回収集し，経年変化を確認できるようなユーザインターフェイスが提供されている。矢田 [11] は形態素解析用辞書 IPADIC の見出し語の Yahoo! Web API による検索結果を収集することで約 396GB 規模 (非圧縮) のテキストアーカイブを作成し公開している。筑波大学は矢田と同様の手法で 11 億語規模のコーパスを構築している [12]。また「Corpus Factory」[13] というプロジェクトにおいて，10^{10}(100 億語) の TenTen という新しい Web コーパス群の開発が始まり，その一つとして日本語 Web コーパス JpTenTen が 2011 年に作成された [14]。上記以外にも著作権の関係で公開されてはいないが各機関で Web アーカイブ・Web コーパスの構築を行っている研究報告が多々ある。

Web を母集団とした言語資源が各機関で構築されつつあるが，同様に構築するためのツールも整備されている。そのいくつかはオープンソースソフトウェアとして公開されている。次節では公開されているツールなどを利用しながら，いかにして超大規模コーパスを構築するか，コーパスの収集と組織化について解説する。

3. 超大規模コーパスの収集と組織化

本節では既存の技術を用いていかにして超大規模コーパスを構築するか，また，自然言語コーパスとしての価値を高めるためにどのような工夫を行うか，さらに，様々な先行研究とどのようにして差別化するかについてくわしく説明する。具体的には大きく分けて収集・組織化・利活用・保存の四つの工程に分割して実装を進める。

収集　超大規模コーパスを構築するための Web テキストの収集は Web クローラを用いることによる。約 1 億 URL を三か月ごとに収集し，一つの URL に対し，複数の版を取得する。

組織化　超大規模コーパスを言語研究に利用可能にするためものである。一般的な Web コーパスで用いられている正規化技術・形態素解析だけでなく，係り受け解析・格解析・レジスタ推定を行い，言語コーパスとしての利用価値を高める。

利活用　組織化されたデータから，言語研究に必要な語彙表/n-gram データを整備する。100 億語規模のテキストから特定の形態論・統語論的パタンの事例を効率的に検索するアプリケーションを構築する。

保存　言語の経年変化を観察するための資料として，収集したコーパスは Web アーカイブとして永続保存する。収集時期を時間軸とした組織化を行う。

以下，四つの工程のうち**収集**と**組織化**の工程の各論について解説する。

3.1 収集技術

Web テキストの収集手法はクローラの運用 (Remote harvesting)，コンテンツ会社からの提供 (Database archiving)，検索エンジン/ソーシャルネットワーキングサービス会社が提供する Web API(Transactional archiving) の利用などがある。2 節に述べた先行研究においては，グーグル・バイドゥ・NICT・京都大学・NDL・JpTenTen'11

が Remote harvesting による収集にあたり，楽天技研・ヤフーが Database archiving にあたり，矢田・筑波大学が Transactional archiving にあたる。

本研究では継続的に収集を行うために Remote harvesting を行う。これは，まず本研究所が自前で超大規模の Web コンテンツをホストしているわけではないために Database archiving が実質的に不可能であること，Web API に基づく Transactional archiving は継続的に収集が続けられるか否かが他機関に依存するだけでなく Web の実態がつかみづらいことという消去法的な理由があげられる。自前でクローラを運用することは工数的にも資金的にも多大な負担が強いられる一方，収集の継続性が確保できるだけでなく，収集範囲をある程度制御することが可能になる。本研究ではバルク収集が可能なクローラ Heritrix を運用する。Heritrix クローラは，wayback machine と呼ばれる Web アーカイブの構築実績を持つ米国 Internet Archive が中心となり開発しているクローラソフトウェアである。各国国立図書館が Web アーカイブを構築するために利用しており，日本では国立国会図書館がインターネット資料収集保存事業において利用している。アーカイブの保存形式は，後述する Web アーカイブの標準化ファイル形式である WARC 形式が選択できる。

各国国立図書館で運用するクローラは画像ファイル・音声ファイル・動画ファイルも含めたバルク収集ができることが重要である。しかしながら本研究においてはテキストデータの収集が主な目的であるために，.html ファイル・.txt ファイルに限定して収集する。

約 1 億 URL をシード URL リストとして，年に 4 回のペースで定点観測的に Web テキストとリンク - 被リンク構造の収集を行う。収集対象は基本的に「日本語」の Web ページとする。ここで「日本語」であることを生産者側の観点から規定するのではなく受容者側の観点から規定する。生産者が日本語非母語話者であろうとスパムサイト (splog) であろうと機械翻訳結果であろうと，日本語母語話者が誤用を含めて日本語として認識できるものについて収集を行い組織化し保存する。外国語で書かれた文は正規化により排除し，外国語で書かれたページはレジスタ分析などにより定期収集対象から除外する。

2012 年 7 月に 100 万 URL 規模の第一次収集テスト，2012 年 8〜9 月に 1000 万 URL 規模の第二次収集テストを繰り返し行い，クローラの設定を検討した結果，週次の収集量を 1000 万 URL 程度とし，3 か月ごとに 1 億 URL 規模の収集を行うことにした。2012 年第 4 四半期 (2012-4Q) から本収集 (第一期) を開始した。第一期から第四期の 1 年間のクロール結果の URL の更新頻度推定などを行い，第六期以降は更新されない URL を収集範囲から外したうえで，新しい URL を収集範囲として含め，収集範囲の拡充を行う。収集範囲の拡充においては，代表性・均衡性ではなく網羅性を重要視する。

クローラの運用においては，robots.txt およびメタタグなどのロボット排除プロトコルを確認し，サイト運営者側のクローラプログラムへの指示を順守する。さら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/電話を設置している。

3.2 組織化技術

Web テキストは収集しただけではそのままコーパスとして用をなさない。以下では，HTML タグ排除や文字コードの統制などの正規化，言語解析としての形態素解析，係り受け解析，格解析・述語項構造解析，コーパスとしての母集団を規定するための基礎情報となるレジスタ推定について説明する。

3.2.1 正規化技術

収集した Web テキストは，HTML タグを含んでいるだけでなく，文字コードが多様である。さらに言語コーパスとして扱うためには，一般的に分析に利用される単位である文境界の認定が必要になる。この HTML タグの排除・文字コードの統制・文境界の認定を Web テキストの正規化と呼ぶ。Web データの正規化については，2 節に示した先行研究の中で，グーグル「Web 日本語 N グラム第 1 版」が採用している手法が事実上の標準となっており，これに準じた正規化が行える「日本語ウェブコーパス用ツールキット」(nwc-toolkit) が公開されている。文字コードの統制においてはまず UTF-8 に変換したうえで正規化形式 C(NFC) を施す。文分割においては句点・感嘆符・疑問符で分割するほか，文の文字数で選別 (6 文字以上 1023 文字以下) を行い，日本語の文字の割合 (ひらがなが 5%以上・日本語の文字が文全体の 70%以上) で明らかに日本語でない文を選別する。

Web テキストの正規化の問題のほかに，異なる URL で全く同じ Web ページであるかいなか，同じ URL に対する異なる収集時期の版であるかいなかを検出する技術を重複性・同一性検出と呼ぶ。重複性・同一性検出は Web ページのハッシュ値比較により行うことが一般的であるが，本研究でも同様の重複性・同一性検出を行う。

3.2.2 形態素解析

収集し，正規化を行った Web テキストに対して，形態素解析を行う。形態素解析を行うことにより，明示的に分かち書きされない日本語に対して，分析する単位としての形態素境界を与える。まず先行研究でよく用いられている MeCab のデフォルトの辞書 IPADIC は，これに基づく統語分析以上の解析器が良く整備されている。一方，形態素解析辞書 UniDic が採用している国語研短単位は，斉一性があり，形態論的な言語分析を行うには適した単位である。日本語教育などの分野で行われるコロケーション分析で

は，国語研短単位では粒度が細かく，より長い単位である国語研長単位で言語分析を行う傾向にある。一方，UniDic が採用している可能性による品詞体系では必要な情報が可能性の名のもとに未定義となり利用できない。このため，係り受けなどの統語分析を行う研究者は益岡・田窪文法に基づく品詞体系 [15] とその品詞に基づいた文節単位を利用する傾向にある。さらに，辞書に登録されない Web 上に新たに生産される形態素 (ネット新語・スラング) を中心に分析する研究者もいる。

このような多様な利用者を想定して，本研究では形態素解析手法として，MeCab/IPADIC による形態素解析，MeCab/UniDic による国語研短単位解析，JUMAN による益岡・田窪品詞体系に基づく解析，ベイズ階層言語モデルによる教師なし形態素解析 [16] の四つを同時に利用する。

3.2.3 係り受け解析・格解析・述語項構造解析

今回作成するコーパスには係り受け解析・格解析・述語項構造解析を行う。係り受け解析手法として，京都大学テキストコーパスの基準に基づいて学習した IPADIC に基づく CaboCha と益岡・田窪品詞体系形態論情報に基づく KNP の二つを利用し，係り受け木を作成する。さらに前者には述語項構造解析器 ChaPAS により NAIST テキストコーパスに基づいた述語項構造を付与する。後者は KNP が京都大学テキストコーパスに付与されている格構造相当の情報を出力する。

3.2.4 レジスタ分析

言語学の観点からすると，Web コーパスの信頼性を下げる大きな要因のひとつは，収集されたテキストがどのような目的で書かれているかというレジスタ情報の欠落である。そのため本コーパスでは，収集された Web ページのレジスタ推定を実施する。収集の時点では，シード URL からリンク構造をたどることによりクロールするため，自然言語コーパスとして均衡性・代表性を持たせた母集団を規定することが困難である。分散を大きくするようなクローラ運用ポリシーにより網羅性を重視したうえで，あらかじめ文書分類的な手法を用いて適切な部分サンプル集合をレジスタとして規定することにより，この問題を緩和する。

4. 収集と組織化の経過と統計

4.1 収集

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 7〜8 月より試験収集を開始し，Heritrix 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，2014 年 4 月 (第六期) に次の 1 年間に収集すべき URL サンプルを決定した。

表 1 収集したページ数の統計量

	ページ数	(内) 重複無
2012-4Q	61,668,805	45,933,605
2013-1Q	58,844,092	42,932,982
2013-2Q	61,479,268	45,111,527
2013-3Q	57,892,917	42,192,931
4 期 (異なり)	64,539,233	*36,934,706

* 4 期 (異なり) の重複無は 4 期を通して重複がなかったもの。

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 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，内容の重複がないページ数は約 3700 万ページになる。

図 1 2012 年第 4 四半期収集ページの重複

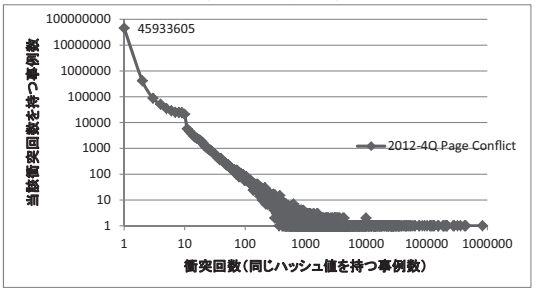

図 1 に 2012 年第 4 四半期収集 Web ページの重複検出結果について示す。同じハッシュ値を持つ URL が複数存在することを「衝突」と呼ぶ。グラフ横軸は同じハッシュ値を持つ URL 数を示し，「衝突回数」と呼ぶ。グラフ縦軸は横軸の「衝突回数」を持つ衝突事例数 (URL 数ではなくハッシュ値の異なり数で計算) を示す。グラフは両軸とも対数で表示している。グラフ中の左上の点が表 2 の「内容の重複なしページ数」(他の URL と内容が重複しないページ数) に相当する。衝突回数 10 以下のものは同一内容の異なる URL 表示もしくはいわゆるコピーサイトであると考える。それ以上の衝突については，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

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	4 期
リンク数 (のべ)	6,905,805,383	6,610,763,700	7,064,611,259	7,222,958,033	27,804,138,375
リンク数 (異なり)	892,135,930	843,166,672	865,694,816	855,684,918	1,642,699,579

ンツを返すページである。

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。約 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→ 16.4 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

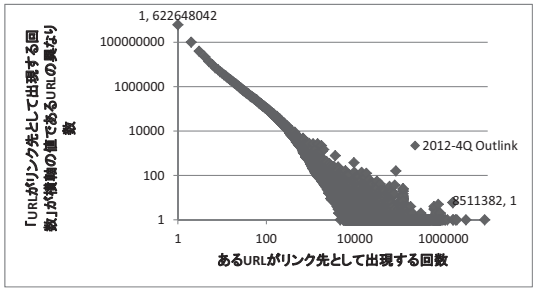

図 2 に 2012 年第 4 四半期の収集リンク先統計を示す。グラフ横軸はある URL がリンク先として出現する回数 (被リンク数) で，グラフ縦軸が横軸相当の被リンク数を持つ URL の異なり数である。グラフの両軸は対数である。グラフ左上が 1 回しかリンクされていない URL 数で約 6 億件である。グラフ右下が最も多い被リンク数約 850 万を持つページで 1 件ある。これは有名ブログサイトのトップページであり，ブログの個々のページからリンクされている。

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。

4.2　組織化

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。8-core のワークステーション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。

図 3　2012 年第 4 四半期の収集文の重複

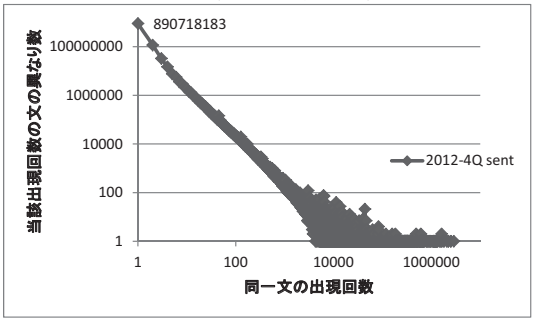

図 3 に 2012 年第 4 四半期の収集文の重複を示す。横軸が同一文の出現回数で縦軸が当該出現回数の文の異なり数を表す。両軸とも対数で表現する。10 億文のうち約 9 割の 8.9 億文が 1 回しか出現しない文である。以下，ページの重複も含めて同一文が異なりで 1 億文規模存在する。これらは定型的な表現やリンクの見出し語であることが多く，一番多く出現した文は 2,885,654 回出現する「職業とキャリア」(Yahoo!知恵袋のカテゴリ名) であった。

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 4 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

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	62,077,520,745	63,414,252,638	65,736,027,334
形態素数 (文抽出あり)	33,767,409,441	32,651,138,004	33,073,991,355	30,923,912,566
文数 (文抽出あり; のべ)	2,678,315,774	2,600,122,908	2,659,617,620	2,478,309,312
文数 (文抽出あり; 異なり)	1,097,011,506	1,048,772,913	1,063,649,324	1,007,771,383

表 4 本言語資源とグーグル「Web 日本語 N グラム」との比較

	本言語資源 (2012-4Q)	本言語資源 (2012-4Q)	グーグル [2]
足切り	頻度 3 以上	頻度 3 以上	頻度 20 以上
総形態素数	180 億形態素	337 億形態素	2,550 億形態素
文処理	異なり	のべ	のべ
総文数	10 億文	26 億文	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

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。

5. 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であるが，これについても早急に環境を整えたい。

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。

参考文献

[1] Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y.: “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation* 48 (2), 345-371, (2014).
[2] 工藤拓・賀沢秀人: 『Web 日本語 N グラム第 1 版』, 言語資源協会発行 (2007).
[3] バイドゥ株式会社: 『Baidu ブログ・掲示板時間軸コーパス』(2010).
[4] バイドゥ株式会社: 『Baidu 絵文字入りモバイルウェブコーパス』(2010).
[5] 楽天技術研究所: 『楽天データセット』(2010).
[6] ヤフー株式会社:『Yahoo!知恵袋データ (第 2 版)』(2011).
[7] Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C., and Kurohashi, S.: “TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access”, *Proceedings of Third International Joint Conferencce on Natural Language Processing (IJCNLP)*, (2008).
[8] 情報通信研究機構: 『日本語係り受けデータベース Version1.1』, (2011).
[9] 京都大学大学院情報学研究科黒橋研究室: 『京都大学格フレーム (Ver 1.0)』(2008).
[10] 国立国会図書館: 『インターネット資料収集保存事業』.
[11] 矢田晋:『日本語ウェブコーパス 2010(NWC2010)』(2010).
[12] 今井新悟・赤瀬川史朗・プラシャントパルデシ: 『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, 299-206, (2013).
[13] Kilgarriff, A., Reddy, S., Pomikálek, J., and Pvs, A.: “A Corpus Factory for Many Languages”, *Proceedings of the Seventh International Conference on Language resources and Evaluation (LREC’10)*, 904-910, (2010).
[14] Pomikálek, J., and Suchomel, V.: “Efficient Web Crawling for Large Text Corpora”, *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, 39-43, (2012).
[15] 益岡隆志・田窪幸則: 『基礎日本語文法』改訂版, くろしお出版 (1992).
[16] 持橋大地・山田武士・上田修功: 『ベイズ階層言語モデルによる教師なし形態素解析』, 情報処理学会研究報告: 2009-NL-190, 49, (2009).