Building NINJAL Web Japanese Corpus : Use and Application

国語研プロジェクトレビュー
NINJAL Project Review　Vol.6 No.1 pp.1–10（June 2015）

〈プロジェクト紹介〉
超大規模コーパス構築プロジェクト

日本語 Web コーパスの構築―利活用―

Building NINJAL Web Japanese Corpus: Use and Application

浅原 正幸（ASAHARA Masayuki）

1. はじめに

1991 年にインターネット上で World Wide Web（以下 Web）が利用可能になって以来，電子化されたテキストが簡単に共有できるようになった。やがて，個人の情報交換媒体として成長し，独自の言語現象を育むようになる。Web に共有されているテキストは話し言葉的な表現や書き言葉的な表現を含むだけでなく，Web 特有の表現であるインターネットスラングを含む。近年ではインターネットスラングが現実世界でも利用されるようになり，話し言葉や書き言葉へ影響を与えるようになりつつある。

このような状況に鑑み，国立国語研究所コーパス開発センターでは 2011 年より Web を母集団とした超大規模コーパス（日本語 Web コーパス；NINJAL Web Japanese Corpus（仮称））を構築する計画に着手した。1 億語規模の『現代日本語書き言葉均衡コーパス』（Maekawa et al. 2014，以下 BCCWJ）に出現しない言語現象を被覆する 100 億語規模の日本語コーパスを時期横断的に収集することを目標とする。浅原ほか（2014）では，日本語 Web コーパスの収集と組織化について議論した。

2012 年第 4 四半期（2012-4Q）より本収集を開始した。表 1 に示すとおり，3 か月ごとに日本語のみで約 300 億語（IPADIC-2.7.0 による）の収集を進めている。表中 WARC ファイルは Web アーカイブを保存するファイル形式で，圧縮して約 1GB 前後のファイルサイズになる。

組織化においては，言語資源として活用できるように，日本語と他言語とを分ける日本語文抽出を行ったあと，形態素解析・係り受け解析を行う。また，Web 上の重複（コピペ）問題を緩和するため，文単位の異なりをとることにより，言語研究に必要な統計量を得る工夫が施されている（Asahara et al. 2014）。

以下では，日本語 Web コーパスの利活用について紹介する。2 節では現在開発を進めている検索系の設計について紹介する。3 節では活用事例としての「畳語」の枚挙について紹介する。

2. 検索系の設計

2013 年度より検索系の調達を開始し，3 か年で開発を進めている。2013 年度は BCCWJ 1

表 1　日本語 Web コーパスの収集

収集時期	2012-4Q	2013-1Q	2013-2Q	2013-3Q
収集 WARC ファイル数	814	870	910	905
URL 数	61,668,805	58,844,092	61,479,268	57,892,917
語数（IPADIC） （日本語文抽出なし）	64,714,650,129 647 億語	62,077,520,745 620 億語	63,414,252,638 634 億語	65,736,027,334 657 億語
語数（IPADIC） （日本語文抽出あり）	33,767,409,441 337 億語	32,651,138,004 326 億語	33,073,991,355 330 億語	30,923,912,566 309 億語
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

億語規模の検索系を開発した。2014 年度は 2012 年第 4 四半期に収集を行った日本語 Web コーパス（以下 NWJC）のうち約 10 億語で検索系の高速化を進めている。公開形態として次の 4 種類の検索系を提供することを予定している。

- 形態素 N-gram 検索
  形態素 N-gram データに対する検索系。
- 文字列検索
  単純な文字列検索。BCCWJ における『少納言』（巻末ツール一覧 6）相当，『ChaKi.NET』における String Search 相当の検索系。
- 形態素列検索
  形態素解析結果に対する形態論情報に基づく系列検索。BCCWJ における『中納言』相当（巻末ツール一覧 7），『ChaKi.NET』（巻末ツール一覧 8）における Tag Search 相当の検索系。
- 係り受け構造検索
  係り受け解析結果に対する，係り受け構造に基づく木構造検索。『ChaKi.NET』における Dependency Search 相当の検索系。

以下では各サービスの機能について紹介する。なお，いずれの検索系についてもバックエンドの開発を先行して進めており，ユーザーインターフェイスについては発展途上のものである。ユーザーインターフェイスについては 2015 年度に開発を進める予定である。

2.1　形態素 N-gram 検索

形態素 N-gram 検索についてはオープンソースソフトウェアの ssgnc（Search Systen for Giga-scale N-gram Corpus）を用いて構築する。ssgnc は以下の 4 つの機能を有している。

● Unordered
トークンの出現順序を考慮しない AND 検索。トークン列の間に他のトークンが入ってもよい。クエリに入力したトークン数よりも長い N-gram についても返す。

● Ordered
トークンの出現順序を考慮する AND 検索。トークン列の間に他のトークンが入ってもよい。クエリに入力したトークン数よりも長い N-gram についても返す。

● Phrase
トークンの出現順序を考慮する AND 検索。トークン列の間に他のトークンが入らない。クエリに入力したトークン数よりも長い N-gram についても返す。

● Fixed
トークンの出現順序を考慮する AND 検索。トークン列の間に他のトークンが入らない。クエリに入力したトークン数と同じ長さの N-gram のみを返す。

図 1 に ssgnc の検索フォームを示す。左図では Fixed 条件での「猫＊＊」という 3-gram を，右図では Fixed 条件での「猫＊＊＊＊＊＊」という 7-gram をクエリとした例である。クエリ中 "*" はワイルドカードを表す。

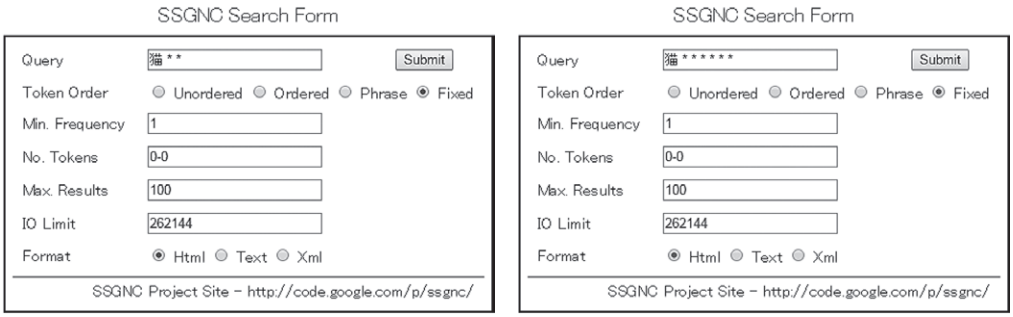

図 1　N-gram 検索系（左：Fixed 3-gram，右：Fixed 7-gram）

表 2 に NWJC 2012 年第 4 四半期収集データ（2012-4Q）を MeCab-0.98+mecab-ipadic-2.7.0-20070801 により形態素解析し，図 1 の条件で検索した際の結果（上位）を示す。N-gram データは文単位に単一化（uniq）したものとそうでないものの両方を準備する。表 2 の結果は文単位に単一化したものである。7-gram では句読点で区切られていないブログのタイトルなどが上位に来る傾向にある。N-gram 検索系については 2015 年度末の一般公開を目指す。

2.2　文字列検索

文字列検索では『少納言』相当の検索系を整備する。文字列検索は後に述べる形態素列検索・係り受け構造検索とともに Preferred Infrastructure 社に外部委託して開発を進めている。

表 2　N-gram 検索結果の例

3-gram “猫 * *”		7-gram “猫 * * * * * *”	
猫ちゃんの	13,500 件	猫をはじめとしたペット	867 件
猫は、	10,000 件	猫君♪】と選択した	722 件
猫を飼っ	8,580 件	猫画像／萌えアニマル画像／	637 件
猫さんの	8,500 件	猫とネコとふたつの本棚	451 件
猫ちゃんが	8,210 件	猫用品・ペット用品・ペット	281 件
猫の手	7,650 件	猫君♪|マイホーム購入前に	278 件
猫には	6,740 件	猫君♪|制度の問題点	271 件
猫たちの	6,670 件	猫・ペットと一緒に泊まれる	238 件
猫がい	6,610 件	猫を抱いて象と泳ぐ	214 件
猫のよう	6,590 件	猫風味なワっち　≪	212 件

文字列検索については同社の製品である Sedue に基づいて開発を依頼した。表 3 に 2013 年に実施した，BCCWJ 1 億語規模での機能検証結果を示す。現在のところ 1 億語規模で検索した際，500 件の結果を返答するのに最長 0.5 秒かかっており，100 億語に拡張しても返答件数を制約づけすれば実用上問題ないレベルであると考える。

表 3　文字列検索機能の性能評価（BCCWJ 1 億語規模に対する返答時間）

検索文字列	ヒット件数	10 件返答時間	500 件返答時間
分	328,415 件	0.053 秒	0.447 秒
すなわち	9,473 件	0.034 秒	0.418 秒
アーティスト	1,114 件	0.018 秒	0.397 秒
カバン	483 件	0.017 秒	0.354 秒
上と下	79 件	0.013 秒	0.062 秒
突きあた	41 件	0.012 秒	0.035 秒
うかびあがっ	22 件	0.012 秒	0.022 秒
しぶい	19 件	0.011 秒	0.018 秒
しんきくさい	2 件	0.005 秒	0.005 秒
無季	0 件	0.004 秒	0.004 秒

2014 年度は 10 億語規模の検証を引き続き進めている。2015 年度は 100 億語規模の検証を進めるとともに，ユーザーインターフェイスの設計に着手し，2015 年度末の一般公開を目指す。

2.3　形態素列検索

形態素列検索は，形態素解析結果に対する検索環境である。機能としては BCCWJ における『中納言』相当，『ChaKi.NET』における Tag Search 相当のことが可能である。形態素解析結果に含まれる，さまざまな形態論情報に基づく形態素連接をクエリとした検索が可能とな

る。

2013 年度に実施した BCCWJ 1 億語規模で行った開発・検証では返答に 10〜20 秒かかることがわかった。2014 年度は NWJC 2012 年第 4 四半期収集データ（2012-4Q）を MeCab-0.98+mecab-ipadic-2.7.0-20070801 により形態素解析したもの（約 10 億語規模）での検証と高速化を進めている。図 2 に形態素列検索の現状を示す。図中上部に「猫 / 名詞」と「*/ 助詞 - 格助詞」の 2-gram クエリを表す。10 億語規模データに対して当該クエリは 21,264 件がヒットし，そのうちの 50 件を表示するための返答時間は 1.54 秒かかることがわかる。

2015 年度は 100 億語規模の検証を進める。返答時間はデータ量に対して線形であることがわかっており，単純計算で 10 億語規模データにおける返答時間の 10 倍かかることが予測される。オフサイト（Web 上，登録制）で利用可能なサービスとして 10 億語規模のデータによる検索系の一般公開を行うとともに，オンサイト（国語研来訪者限定）で利用可能なサービスとして 100 億語規模のデータを準備する予定である。

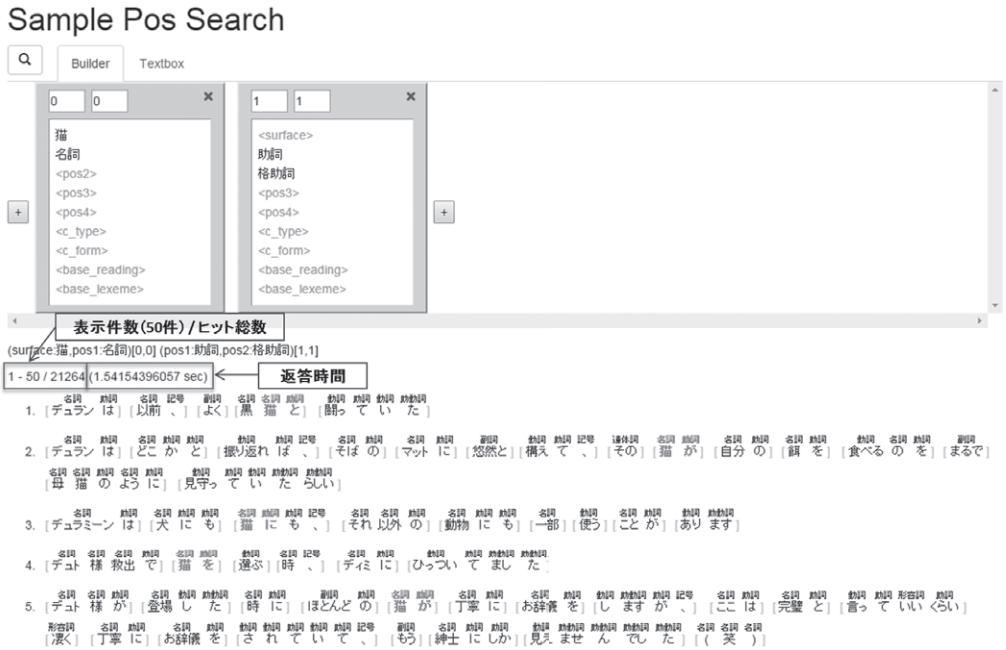

図 2　形態素列検索のユーザーインターフェイス

2.4　係り受け構造検索

係り受け構造検索は，係り受け解析結果に対する検索環境である。『ChaKi.NET』における Dependency Search 相当のことが可能である。形態素解析・係り受け解析結果に含まれる，さまざまな形態論情報・文節境界・係り受け関係に基づくクエリを用いた検索が可能となる。

形態素列検索と同様に，2013 年度に実施した BCCWJ 1 億語規模で行った開発・検証では

返答に 10〜20 秒かかることがわかった。2014 年度は NWJC 2012 年第 4 四半期収集データ（2012-4Q）を MeCab-0.98+mecab-ipadic-2.7.0-20070801 により形態素解析し，CaboCha-0.67 により係り受け解析したもの（約 10 億語規模）での検証と高速化を進めている。

図 3 に係り受け構造検索の現状を示す。図中上部左側に「猫 / 名詞」と「は / 助詞 - 係助詞」の 2-gram を含む文節（文節 ID 0）と，図中上部右側に「だ」を含む文節（文節 ID 1）を含み，2 つの文節間に係り受け関係があるような部分木クエリを示す。10 億語規模データに対して当該クエリは 196 件ヒットし，そのうちの 50 件を表示するための返答時間が 2.64 秒かかることがわかる。

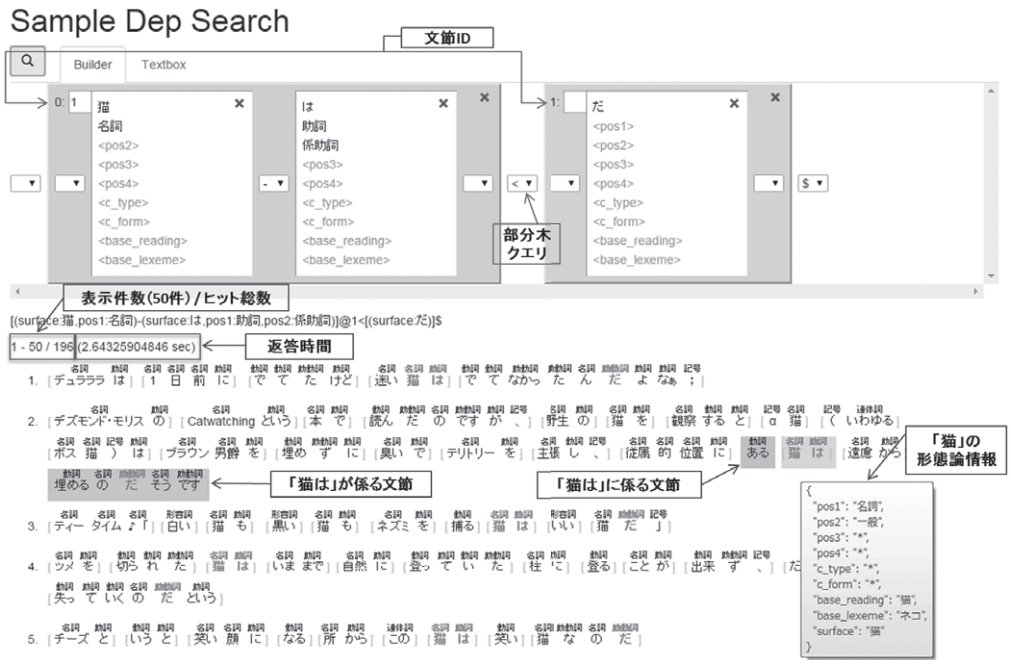

図 3　係り受け構造検索のユーザーインターフェイス

現在作成しているユーザーインターフェイスでは，簡易的な形態論情報・係り受け構造表示機能を備えている。図中下部の用例中の文節（[猫 は]）にマウスカーソルを合わせると，ポップアップで形態論情報が表示されるとともに，その文節に係る文節（[ある]）とその文節が係る文節（[埋める の だ そう です]）がハイライトされる。

形態素列検索と同様，2015 年度は 100 億語規模の検証を進める。返答時間はデータ量に対して線形であることがわかっており，単純計算で 10 億語規模データにおける返答時間の 10 倍かかることが予測される。オフサイト（Web 上，登録制）で利用可能なサービスとして 10 億語規模のデータによる検索系の一般公開を行うとともに，オンサイト（国語研来訪者限定）で利用可能なサービスとして 100 億語規模のデータを準備する予定である。

3. Web データに対する系列パターンマイニング（畳語の枚挙）

前節では一般公開に向けて準備している検索系について紹介した。各検索系は 10 億語もしくは 100 億語規模であっても現実的な時間で動作するものにするために，いわゆる正規表現相当の検索は整備していない。さらに著作権の問題から Web データそのものを対外的に頒布することも困難である。しかしながら国語研所内で特定のパターンの文字列を枚挙して，語彙表として配布することは可能である。以下では，前節で示した検索系では展開できないデータの例として「畳語」の枚挙について述べる。

ここで「畳語」とは，部分形態素・形態素・形態素連接などの単位を反復して作られた合成語のことをいう。今回展開する畳語は形態素境界を無視して連続する文字列とし，文字列が反復していれば句単位・節単位の畳語についても展開することとする。また，連濁や踊り字にも対応することとする。展開する畳語の仕様を以下に示す。

- 正規表現 /（.?*）¥1/ にマッチする AA，ABAB，ABCABC のようなパターンを枚挙する。なお，¥1 はグループ化した括弧内要素の後方参照とする。
- 連濁対応　Unicode 正規化において濁点を含む文字を語幹分解したものに対して正規表現 /（.）（.*）¥1゛¥2/ にマッチする AA゛，ABA゛B，ABCA゛BC のようなパターンを枚挙する。
- 踊り字対応　後方参照要素がグループ化した括弧内要素の文字数と同じ数の踊り字（々，ゝ，ゞ）である場合にもマッチする。例えば A々，Aゝ，Aゞ，AB々々のようなパターンを枚挙する。

これらの畳語パターンのみを枚挙するために系列パターンマイニングアルゴリズム PrefixSpan（Pei 2001）を改変した『prefixspan-rel』を用いる。同アルゴリズムを BCCWJ と 2012 年 NWJC 第 4 四半期収集データ（2012-4Q）の日本語文抽出結果（約 100 億語相当）に対して適用する。処理の都合上 BCCWJ については任意の長さの畳語を，NWJC については長さ 12 文字までの畳語を展開した。表 4 に処理時間と枚挙された畳語の件数を示す。主記憶 512GB，CPU AMD Opteron 6140（2.6GHz 8 core）× 4 CPU の機材で 30 並列で実行すると前者については約 5 分で，後者については約 2 日で枚挙が完了する。

表 4　畳語の枚挙

コーパス	枚挙する畳語長	母集団規模	処理時間	異なり	のべ
BCCWJ	無制限	約 1 億語	約 5 分	452,574 件	139,431,772 件
NWJC（2012-4Q）	12 文字以下	約 100 億語	約 2 日	26,928 件	1,622,656 件

表 5 に NWJC で頻度が 3 であった畳語の例を示す。

表 5　NWJC から枚挙された畳語の例（頻度 3）

ゃんほむらちゃんほむらち	ラインのラインの
安全安全安全安全	粉→卵→パン粉→卵→パン
不幸が訪れ、不幸が訪れ、	勝ちました、勝ちました、
ほーれ、ほーれ、	わにわにわにわにわにわに
ハンデハンデ	観た、観た、

形態素解析結果を用いていないため，必ずしも反復の単位が形態素境界と一致しているわけではない。さらに禁則処理などを全く行っていないために捨て仮名や約物などが先頭に来る場合もある。これらを後処理で取り除く必要があるだろう。

上の例では，一般公開のシステムで処理することが困難な畳語の枚挙について示した。このようなデータが必要な場合はパターンに基づいて国語研内サーバで枚挙したうえで語彙リストとして提供する予定である。

4.　おわりに

本稿ではコーパス開発センターの超大規模コーパス構築プロジェクトで開発している日本語 Web コーパスの利活用について紹介した。現在開発を進めている検索系の設計と開発の進捗を報告するとともに，活用事例としての畳語の枚挙について紹介した。

検索系の設計においては，2014 年度までに 10 億語規模での一般向けのサービスが可能であることを確認した。2015 年度は法的問題について検討するとともに，100 億語規模への規模拡張を進める。

利活用の一形態として，系列パターンマイニングアルゴリズムによる畳語の枚挙について紹介した。検索系で展開できない特定パターンの枚挙については，国語研内でパターン枚挙を行うことにより利用できるようにしていきたい。

最後に重要な問題として公開に際しての権利関係の問題がある。こちらについては著作権関連法の動向を見ながら，フェアユースに基づく公開ができるよう引き続き努力していきたい。

●付記●
・超大規模コーパスシステム開発メンバーは以下のとおり（50 音順）。
浅原正幸，今田水穂，加藤祥，小西光，前川喜久雄
・検索系の開発には，株式会社 Preferred Infrastructure のみなさんに大変お世話になりました。この場を借りて感謝申し上げます。また，本研究に用いているさまざまなオープンソースソフトウェアや言語資源を公開されている方々に感謝申し上げます。

●参照文献●

浅原正幸・今田水穂・保田祥・小西光・前川喜久雄（2014）「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―」『国立国語研究所論集』7: 1–26.

Asahara, Masayuki, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi（2014）Archiving and analysing techniques of the ultra-large-scale web-based corpus project of NINJAL, Japan. *Alexandria* 25（1–2）: 129–148.（DOI 10.7227/ALX.0024）

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den（2014）Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation* 48（2）: 345–371.（DOI 10.1007/s10579-013-9261-0）

Pei, Jian, Jiawei Han, Behzad Mortazavi-asl, Helen Pinto, Qiming Chen, Umeshwar Dayal and Mei-chun Hsu（2001）PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. *Proceedings of 2001 International Conference on Data Engineering*（*ICDE'01*）, 215–224.

●ツール●（すべて 2015 年 3 月 1 日確認）

1. Masayuki Asahara 『prefixspan-rel』1.3 http://prefixspan-rel.sourceforge.jp/
2. Taku Kudo 『CaboCha』0.67 （最新版は 0.69） https://code.google.com/p/cabocha/
3. Taku Kudo 『MeCab』0.98 （最新版は 0.996）
   http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
4. Preferred Infrastructure 『統合検索プラットフォーム Sedue』 https://preferred.jp/product/sedue/
5. Susumu Yata 『SSGNC: Search System for Giga-scale N-gram Corpus』 Version 0.4.6
   https://code.google.com/p/ssgnc/
6. 国立国語研究所コーパス開発センター『少納言』 http://www.kotonoha.gr.jp/shonagon/
7. 国立国語研究所コーパス開発センター『中納言』 1.1.0 https://chunagon.ninjal.ac.jp/
8. 奈良先端科学技術大学院大学自然言語処理学研究室・総和技研『ChaKi.NET』 2.08 Revision 496
   http://sourceforge.jp/projects/chaki/releases/

《要旨》 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構築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 4 つに分割して実装を進めている。2012 年第 4 四半期より 3 か月ごとに 1 億 URL のクロールを繰り返し実施している。本稿では構築されたコーパスデータの基礎統計量を示し，本コーパスを用いて，どのような理論的・応用的研究が可能になると考えられるかを論じる。

Abstract: In 2011, the National Institute for Japanese Language and Linguistics launched a corpus compilation project with the aim of constructing a ten-billion-word Web corpus. The project was split into the following four sub-projects: page collection, linguistic annotation, release, and preservation. In the page collection stage, crawling began during the fourth quarter of 2012. We crawled 100 million URLs every three months as fixed-point observations. This paper presents the basic statistics of the crawled data and discusses possible theoretical and practical implications of these language resources.

浅原 正幸（あさはら・まさゆき）

国立国語研究所言語資源研究系・コーパス開発センター准教授。博士（工学）（奈良先端科学技術大学院大学）。奈良先端科学技術大学院大学助手・助教，国立国語研究所コーパス開発センター特任准教授を経て，2014 年 10 月より現職。
主な著書・論文：Archiving and analysing techniques of the ultra-large-scale web-based corpus project of NINJAL, Japan（*Alexandria* 25(1–2), 2014），BCCWJ-TimeBank: Temporal and event information annotation of Japanese text（*International Journal of Computational Linguistics and Chinese Language Processing* 19(3), 2014）.
社会活動：言語処理学会会誌編集委員，情報処理学会自然言語処理研究会運営委員.

コーパス開発センタープロジェクト「超大規模コーパス構築プロジェクト」
プロジェクトリーダー　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 教授）

プロジェクトの概要

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構築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。1 億語規模の『現代日本語書き言葉均衡コーパス』に出現しない言語現象を収集するとともに，次々と出現する Web 特有の新語の変遷が分析可能になるよう，3 か月ごとに 1 億 URL を収集する。各期の日本語テキストコーパスの規模は 300 億語規模になる。収集したコーパスは形態論情報・係り受け構造を自動付与し，人文系の研究者が活用できる利用者系の構築を目指す。