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

著者	浅原 正幸, 今田 水穂, 保田 祥, 小西 光, 前川 喜久雄
雑誌名	国立国語研究所論集
号	7
ページ	1-26
発行年	2014-05
URL	http://doi.org/10.15084/00000522

国立国語研究所論集 (NINJAL Research Papers) 7: 1–26 (2014)
ISSN: 2186-134X print/2186-1358 online

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[a]　今田　水穂[b]　保田　祥[b]　小西　光[c]　前川喜久雄[d]

[a]国立国語研究所 コーパス開発センター
[b]国立国語研究所 コーパス開発センター プロジェクト研究員
[c]国立国語研究所 コーパス開発センター 非常勤研究員
[d]国立国語研究所 言語資源研究系／コーパス開発センター

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実装を進めている。本論文ではそのうち最初の2工程について報告する。収集に関しては，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。また組織化に関しては，2013年第3四半期までの約1年間に収集されたWebページの文抽出・形態素解析・係り受け解析を実施した。これらの作業に生じた問題とその解決法を示した後，2013年末において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*。

キーワード：コーパス構築，Webアーカイブ，言語情報組織化，言語解析

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言い難く，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。本稿では，超大規模コーパスをどのようにして構築するか，どのような情報を付与するか，どのような検索環境を提供するのかなど，設計について概説する。

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

* 本研究の内容は平成25年12月3日開催第102回NINJALサロンでの発表をもととしている。

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。組織化とは図書館情報学における専門用語で情報資源に対し目録や分類を行い，研究に資する情報資源の表現を体系化する活動のことを言う。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。この言語資源を利用してどのような理論的・応用的研究が可能だろうか？

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，一般に利用可能な Web を母集団とした言語資源について示す。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・利活用・永続保存の四つの観点から概要設計について示す。4 節では超大規模コーパスを開発するために必要な環境整備について論じ，その環境の制約の中で現在どのように収集と組織化が実施されているかについて述べる。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，5 節では今回構築する言語資源によりどのような理論的・応用的研究が可能かについて述べる。最後に 6 節で現時点でのまとめと今後の展開について示す。また，付録として本稿の末尾に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 語の比較を行う。また，稿末に関連 URL の一覧を示す。

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見る。

グーグルは『Web 日本語 N グラム第 1 版』（工藤・賀沢 2007）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（2010a）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，月ごと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（2010b）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（楽天技術研究所 2010）。ヤフー株式会社は『Yahoo! 知恵袋データ（第 2 版）』（2004 年 4 月〜 2009 年 4 月）（ヤフー株式会社 2007, ヤフー株式会社 2011）を公開している。

研究機関などにおいては，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（Shinzato et al. 2008）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（情報通信研究機構 2011）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（河原・黒橋 2006, 京都大学大学院情報学研究科黒橋研究室 2008）。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（国立国会図書館；関根 2010）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。様々な技術の集積により，検索エンジンを運営している企業やコンテンツを保持している企業だけでなく，個人でも Web スケールの言語資源を構築することが可能になっている。矢田（2010）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（今井ほか 2013）。また「Corpus Factory」（Kilgarriff et al. 2010）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（Pomikálek and Suchomel 2012）。

表 1 に一般に公開されている Web を母集団とした言語資源のまとめを示す。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，コーパスの概要設計について解説する。

表 1　関連研究：Web を母集団とした言語資源

一般企業	
グーグル	『Web 日本語 N グラム第 1 版』 元データ 2550 億語 /200 億文規模の語彙表・n-gram データ。検索エンジン用収集日本語データについての 2007 年 7 月時点のスナップショット。
バイドゥ	『Baidu ブログ・掲示板時間軸コーパス』 ブログや掲示板データを対象にした語彙表・n-gram データ。2000 〜 2010 年 7 月にかけてのブログや掲示板のデータ計 1000 万文。月ごとに母集団を設定。
バイドゥ	『Baidu 絵文字入りモバイルウェブコーパス』 2010 年 6 月までにモバイル検索向けに収集したデータを元に作成された語彙表・n-gram 統計情報。
楽天技研	『楽天データセット』（2010 年初回公開。以下は 2012 年 8 月公開版の情報） 楽天市場のレビュー（1660 万レビュー），楽天トラベルのレビュー（465 万評価・レビュー），楽天レシピのレシピ情報（44 万レシピ）ほか。
ヤフー	『Yahoo! 知恵袋データ』（以下は「第 2 版」の情報） 2004 年 4 月〜 2009 年 4 月の QA 記事。質問数 2600 万，回答数 7300 万。
研究機関など	
NICT	『日本語係り受けデータベース Version 1.1』 6 億ページ（約 430 億文規模）の係り受け関係 4.8 億対。収集時期 2007 年 5 月 19 日〜 11 月 13 日。
京都大学	『京都大学格フレーム Version 1.0』 約 16 億文規模のテキストから自動構築した約 4 万用言の格フレーム。
NDL	国立国会図書館『インターネット資料収集保存事業』 国・自治体・法人・機構・大学などのサイトと電子雑誌の保存事業。
矢田	『日本語 Web コーパス 2010』 2010 年に IPADIC-2.7.0 の見出し語をシードとし Yahoo! Web API から Web ページ, HTML アーカイブ（1 億ページ, 非圧縮 3.25TB），テキストアーカイブ（非圧縮 396GB），n-gram コーパス（文字，形態素単位）を配布。
筑波大学	『筑波大学 Web コーパス』 2011 年に構築。［矢田 2010］と同様の手法で収集して構築した約 11 億語のコーパス。
Corpus Factory	『JpTenTen'11』 2011 年にクロールした 100 億語規模のコーパス。

3. 超大規模コーパスの概要設計

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 3 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。

図 1 に超大規模コーパスの概要設計について示す。四つの工程の細分類として必要な要素技術を示し，さらにその細分類として具体的にどのようなツールや技術を導入するのかについて示している。以下四つの工程の各論について解説する。

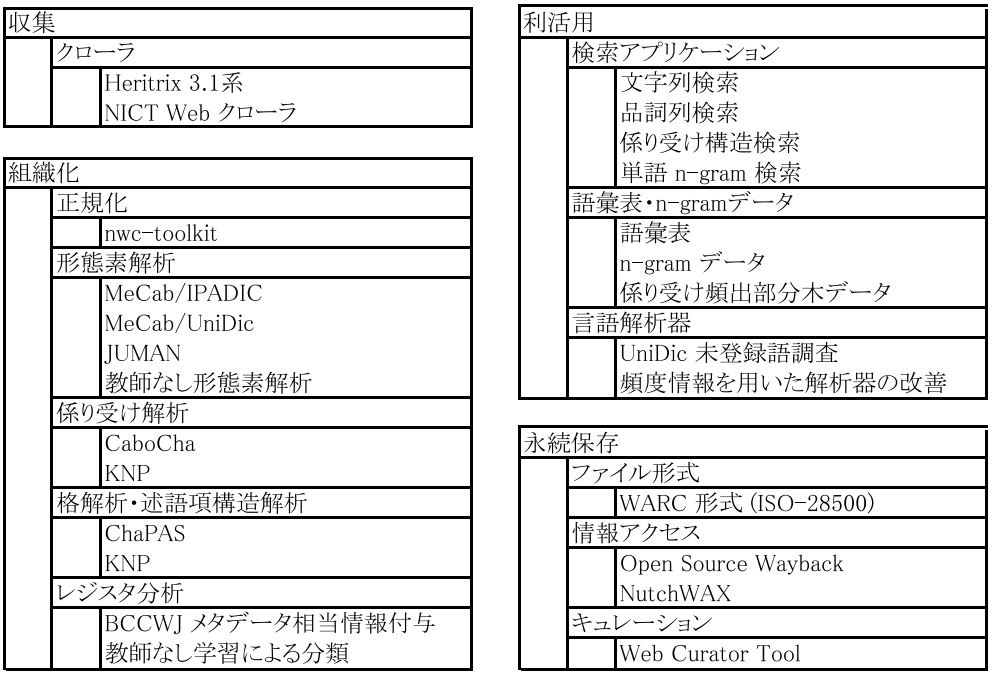

図 1　超大規模コーパスの概要設計

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・Corpus Factory 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。その理由としては，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いな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 [(2)] を運用する。クローラ Heritrix 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。生産者が日本語母語話者であろうと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，2013 年 12 月末現在，本収集（第五期）を行っている。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。また，あとに詳述する環境整備において国立国語研究所が接続する SINET（学術情報ネットワーク）を利用しないでクロール可能なネットワーク環境を構築し，研究所のネットワークに負荷をかけないようにしている。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。

また，情報通信研究機構（NICT）によりオープンソースソフトウェア NICT Web クローラ [(3)] が 2013 年夏に公開された。技術調査が進み次第導入を行い，クローラソフトウェアの多重化を

行いたい。

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）[(4)] 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：正準等価性により分解され，再度合成することによる文字の正規化：「か」「゛」→「が」）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 [(5)] 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，形態素解析用辞書 UniDic [(6)] 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。例えば品詞「名詞 - 副詞可能」は，係り受けの認定に重要な『名詞であるか副詞であるか』ということを形態素解析器において判定することは行わない。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（益岡・田窪 1992）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 [(7)] 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（持橋・山田・上田 2009）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

海外の Treebank においては，GB 理論・X バー理論などに基づく句構造木のコーパスを構築することが多い。一方，日本語においては，深い統語構造について日本語学・言語学双方に共有可能なレベルにない。しかしながら言語処理の分野で係り受け解析器が実用のレベルに達しつつある。またコーパスコンコーダンサにおいては，句構造木レベルの問い合わせを実装すると XPath/XQuery 相当の処理が必要になる一方，非終端記号を要しない係り受け木レベルの問い合わせの場合には関係データベース相当の実装でかなりの部分木構造を被覆する。また，日本語学者・言語学者双方が求める統語構造を係り受け木に投射することで高速な問い合わせが可能になる。さらに格構造・述語項構造を付与することにより必要な統語構造が可視化できる。

そこで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 [(8)] の基準に基づいて学習した IPADIC に基づく CaboCha [(9)] と益岡・田窪品詞体系形態論情報に基づく KNP [(10)] 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 [(11)] により NAIST テキストコーパス [(12)] 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一つは，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。

具体的には，外国語・スパムサイト（splog）・機械翻訳や機械生成されたテキストで非規範的なものを認定するための分類（（半）教師あり機械学習），BCCWJ に付与された各種メタデータ・ファイル単位アノテーションを推定するための分類（（半）教師あり機械学習），クラスタリングに基づく分類（教師なし機械学習）などを検討している。教師あり機械学習は，多クラスのトラ

ンスダクティブ SVM [13] による境界事例分析と，ブースティング法 BACT [14] による有効特徴量分析を行い，クラスタリングによる分類については得られたクラスタに対して言語学（文体論）的な見地からの分析を行う。教師なし機械学習においては，文書集合をどのような特徴量空間に写像するか（持橋・菊井・北 2005）の検討を行う。

3.3 利活用技術

組織化されたコーパスとして利活用していくうえで必要な環境整備について，検索アプリケーションと語彙表・n-gram 頻度情報について説明する。また利活用の事例として想定している言語解析技術への利用についても述べる。構築したコーパスを計算機の扱いが不得手な研究者が利用可能にするために，高速な検索アプリケーションを提供する。レジスタに基づいた絞込（ファセットナビゲーション）を可能にする高速文字列検索機能，コーパス検索アプリケーション「中納言」[15] のような品詞情報に基づいた検索機能，コーパスアノテーション支援環境「ChaKi.NET」[16] の Dependency Search のような係り受けの部分木構造に基づいた検索機能を，100 億語規模で現実的に動作する機能に絞って提供する予定である。

3 か月おきにクロールするデータに対して組織化を行ったうえで，語彙表（形態素 1-gram 頻度情報）・文字列上の n-gram 頻度・形態素列上の n-gram 頻度情報・文節係り受け木上の部分木頻度情報を収集時期ごとに区切られたサンプル単位で取得する予定である。なお，この語彙表・n-gram 頻度情報を得る母集団は，分散を大きく尖度を小さくするように収集を行うが，歪度については制御しないために代表性は担保されない。n-gram データの構築には FREQT [17] を利用する。また，別処理により，HTML タグの頻度情報・リンク - 被リンク関係・同一コンテンツ関係など Web テキスト特有の情報を取得し保持する。可能であれば，レジスタ推定時にこれらの情報を活用する。

得られたコーパスを用いた言語解析技術の向上手法について検討を行う。形態素解析においては，教師なし形態素解析技術や未知語処理技術により得られた UniDic 未登録語について，人手で形態論情報を付与することにより辞書の拡充を行う。他の言語解析器については，教師なし機械学習に基づく手法の n-gram 頻度情報や部分木頻度情報を用いた各種言語解析技術の性能改善手法について検討を行う。

3.4 保存技術

収集したデータは言語の経年変化を分析するための基礎データとするために永続保存する。

IIPC（International Internet Preservation Consortium）[18] における各国国立図書館の活動動向を見ながら，保存のための組織化を行う。具体的には Heritrix で収集されたデータは，Web アーカイブの保存形式の国際標準 WARC 形式 [19] で保存する。WARC ファイルは Internet Archive が公開している Wayback Machine [20] と同じ機能を持つハーベストソフトウェア Open Source Wayback [21] と，情報検索システム NutchWAX（Nutch Web Archive eXtension）[22] により組織化し，Web アーカイブとしての情報アクセスを可能にする。また，選択的な Web クロールを可能にするための

キュレーションツール WCT（Web Curator Tool）[23] の技術調査を行う。日本におけるコーパス言語学は，表層的な情報を用いた統計的手法に基づく分析に偏重しがちだが，用例・事例分析に基づくキュレーション分析に回帰すべく，キュレーションを効率的に行う環境を検討する。最後に，長期保存可能な記憶媒体を内外に確保し，収集し組織化したデータの保存に努める。

4. 環境整備と収集と組織化—現況と課題

本節では環境整備と収集と組織化の現況について概説する。必要に応じて現在得られている統計情報を示し，明らかになりつつある課題について示す。

4.1 環境整備

本節では Web スケールのコーパスの構築における環境整備について議論する。本研究を進めるためには機材を導入するだけでなくネットワークを構築する必要がある。国立国語研究所は Web 関連企業のようなデータセンターや理工系の大学のようなスーパーコンピュータや PC クラスタを保有するわけではなく，新たに本プロジェクトに必要な環境を整備する必要がある。

環境整備の前提条件として，予算の制約・コンピュータ室の制約・管理者の制約・既存ネットワークとの整合性などがある。まず予算の制約上，単年度に大規模の機材調達を行うことができない。そこで，2011 年度・2012 年度に収集と組織化に資する機材と環境を調達し，2012 ～ 2015 年度の 4 か年で利活用と保存に資する機材を調達することとした。コンピュータ室の制約として，電源の制約・空調の制約・空間の制約・耐荷重の制約などがある。電源の制約はコンピュータ室に引き込んでいる電源容量の制約で現況では NEMA L6-30P 2 系統が利用可能である。これに対応する無停電電源装置を 2011 年度に 3600V 規模のものを 1 台，2012 年度に 5000V 規模のものを 1 台調達した。空調の制約は 2011 年度の空調工事により，先に示した電源容量範囲のサーバの熱容量に応じた空調が導入されている。空間の制約においては 2011 年度に 42U ラック 1 本，2012 年度に同ラック 1 本を調達した。しかしながら，コンピュータ室の床の耐荷重がボトルネックとなり，ラックにサーバを半分以上搭載できないことがわかったため，2012 年のラック増設時に床の耐荷重を高める工事を行った。また本プロジェクトには専任の計算機管理者がいないために研究者が自分で管理可能な機材構成にする必要がある。さらに既存のネットワークは Web ページの収集用途に構成されているものではなく，新たに別系統のネットワークを構成する必要がある。

最初に，このような前提条件のもと，2011 年 10 月～ 2012 年 6 月に実施した収集・組織化向けの環境整備について詳しく述べる。

まずネットワーク構成について述べる。ネットワーク構成は国語研の既存のネットワークとは分離された二つのセグメントを新たに構成する。二つのセグメントとも，国語研の既存のセグメントからアクセス可能だがこれらのセグメントから国語研の既存のセグメントにアクセスできないような論理構成になっている。一つはサーバ群を接続するセグメントであり，もう一つのセグメントは外来研究者がサーバ群にあるデータを利用するためのセグメントである。サーバ群を接続するセグメントは物理的にはコンピュータ室と研究者が在室する作業室の 2 か所にわたる。あ

とに述べるように基盤的な機能を担うラックマウントサーバがコンピュータ室に，小規模な実験を行う計算機が作業室に設置されている。外来研究者用のセグメントは作業室の特定の机の上にあるネットワーク機器に持ち込み機材を接続することで利用可能な設定になっている。新たに構成した二つのセグメントは国語研の既存のセグメントが接続する SINET（学術情報ネットワーク）ではなく，別途導入する B フレッツ回線から外部ネットワークに接続し，一般のプロバイダを介して Web ページの収集を行う。家庭用の B フレッツ回線ではあるが近隣に一般家庭がないために末端帯域を外部と共有しない。本ネットワークを構成するために既存のネットワーク機器の設定を変更するとともに，外部回線に接続するためのルーター 2 台（1 台が運用機材で，1 台が補器）とサーバ群を接続するギガビットスイッチ 2 台を調達した。

図 2 に構成したネットワークの物理構成図，図 3 に論理構成図を示す。

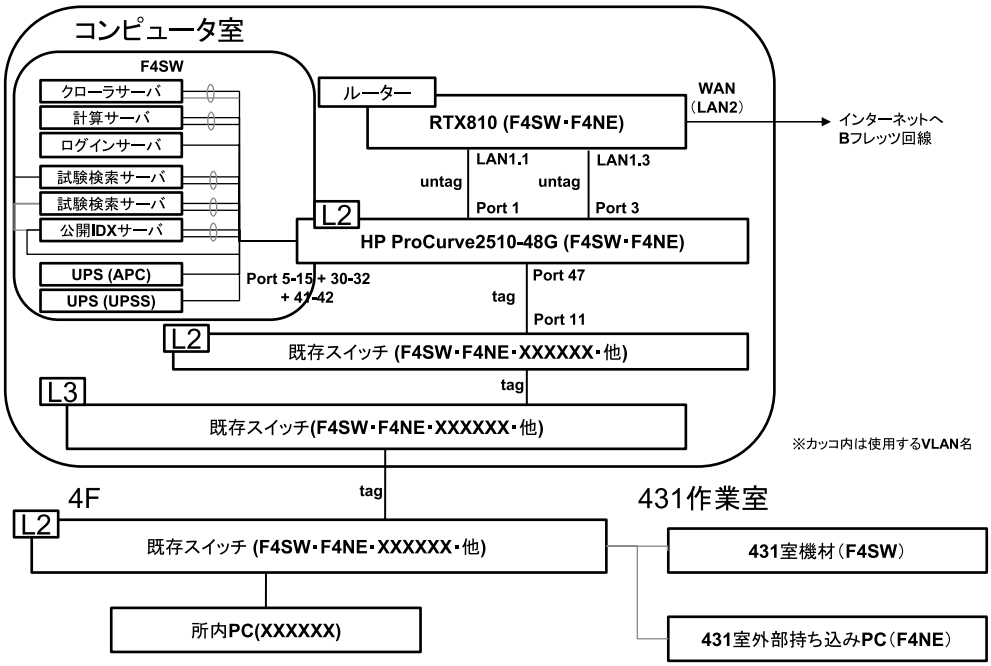

図 2　ネットワーク構成（物理構成）

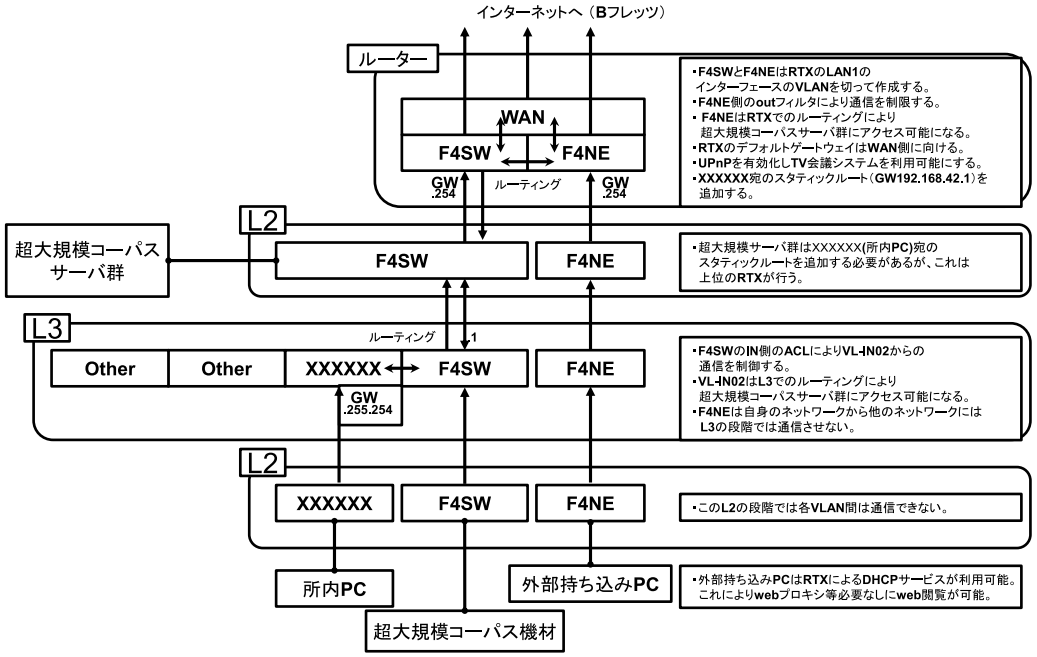

図 3　ネットワーク構成（論理構成）

図中 XXXXXX とあるのが国語研の既存のセグメントである。スイッチは基本的にはギガビットスイッチを利用している。F4SW が超大規模コーパス開発用サーバほか各種機材を格納するセグメントで，F4NE が外来研究者が持ち込む機材を接続するためのセグメントである。外来研究者が持ち込む機材はルーターの DHCP サービスにより自動的にアドレスが付与されるほか，超大規模コーパスサーバ群への接続が許可され，B フレッツ経由で外部ネットワークに接続することができるが，国語研の既存のネットワークへの接続は許可されておらず，論理的に切断されている。一方, 国語研所内の既存機材から超大規模コーパスサーバ群に接続することは可能である。

物理的にコンピュータ室にあり論理的に F4SW 内にあるサーバ群はギガビットスイッチに対し 2 系統（Bonding）で接続しており，無停電電源装置に接続している。一方，作業室にある機材については，1 系統のみで接続しており，無停電電源装置にも接続していない。作業室は勤務時間内しか空調が運転されないため，夏期の冷房期間中 4 か月間電源を入れない。

続いてサーバ機材について述べる。収集・組織化のために計算機室に導入した主要サーバ機材は以下の 3 台である。

- ccd-lserv-40：ゲートウェイ兼ログインサーバ

  Dell PowerEdge R415, AMD Opteron 4112 2.2GHz 4-core 2-CPU, 16GB mem., SAS 300GB HDD × 2, RAID 1, Ubuntu 11.10 Server
- ccd-lserv-41：クローラサーバ

  Dell PowerEdge R515, AMD Opteron 4180 2.6GHz 6-core 2-CPU, 32GB mem., SSD 50GB × 10,

RAID 6, Ubuntu 11.10 Server

- ccd-lserv-42：計算サーバ
  Dell PowerEdge R815, AMD Opteron 6180 2.6GHz 8-core 4-CPU, 512GB mem., SSD 100GB × 5, RAID 6, Ubuntu 11.10 Server

ccd-lserv-40 は各種プロジェクト管理用の機材として利用する。ccd-lserv-41 がクローラサーバでクローラプログラム Heritrix が動作する。ccd-lserv-42 は計算サーバで主として MeCab, JUMAN, CaboCha, KNP をはじめとする言語解析器が動作する。ccd-lserv-41 と ccd-lserv-42 についてはそれぞれ実効 16TB（2TB × 12，内 2 台がホットスワップ，RAID 6 構成）のストレージが接続されている。これらのサーバ群は先に述べたギガビットスイッチ 1 台に 2 系統(Bonding) で接続する。サーバ群とは別に安価なワークステーション（ccd-lserv-01 〜 03）を作業室に常時 3 台設置しており，各種実験環境の試験やデータの多重化などに用いられている。このようにプロジェクトが終了する 2016 年くらいには個人もしくは小規模な研究室でも調達可能な水準の計算機構成で収集・組織化を進めている。

次に 2012 年度後半から調達を進めている利活用・保存のためのサーバ機材について述べる。利用者系のサーバとして 2 台の試験用サーバと 4 台の公開用サーバ(インデックスサーバ× 2 台，検索用サーバ× 2 台）とストレージを，保存系機材として BTO テープドライブと Open Source Wayback や NutchWAX 系のサービスを実現するサーバの調達を予定している。2013 年度末までに利用者系の以下の 3 台の機材を調達する。

- ccd-lserv-50, ccd-lserv-51：試験用検索サーバ× 2 台
  Dell PowerEdge R320, Intel Xeon E5-2407 2.2GHz 4-core 1-CPU, 64GB mem., SAS 2TB HDD, 100GB SSD × 2, RAID 0, Red Hat Enterprise Linux v6.3
- ccd-lserv-52：公開用インデックスサーバ× 1 台
  Dell PowerEdge R320, Intel Xeon E5-1410 2.8GHz 4-core 1-CPU, 64GB mem., SAS 2TB HDD × 2, RAID 1, Red Hat Enterprise Linux v6.5

これらのサーバ群も先に述べたギガビットスイッチ 1 台に 2 系統 (Bonding) で接続するほか，外部の技術者が作業できるようにルーター側からログインできる経路を整備している。

最後に環境整備に関する課題について述べる。一つ目の課題は保存媒体である。次節で述べるように，当初想定していた収集容量をやや上回る規模で収集が進んでいる。保存容量がひっ迫しないように随時 2TB 〜 3TB の HDD に収集済みデータ・解析済みデータを格納している。HDD を保存媒体にするのはあまり得策とは言えず，現在 BTO テープドライブによる保存を検討している。二つ目の課題は頻度集計時の計算用機材の空間計算量である。解析には時間計算量のみが問題となり計算サーバ 8 core × 4 CPU レベルでも対応できている一方，頻度集計時の処理に単純な手法では計算サーバ 1 台で処理していると空間計算量が厳しくなる傾向がある。この問題については機材で解決する方法とは別にソフトウェアによる解決を検討している。三つ目の

課題は外部計算機資源の利用である。収集には時間がかかるが現状解析と頻度集計は比較的短期間で完了する。解析と頻度集計は年間を通して必要な処理ではないために必要に応じて大規模分散並列処理環境を短期間利用する方がよい。さらに保存媒体についても物理的な媒体ではなく外部計算機資源にも格納することを検討したい。

4.2 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 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情報システム・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 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 年ごとに変更する方針にした。2012 年 10 月より 3 か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。

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 2　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 2 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 ごと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。

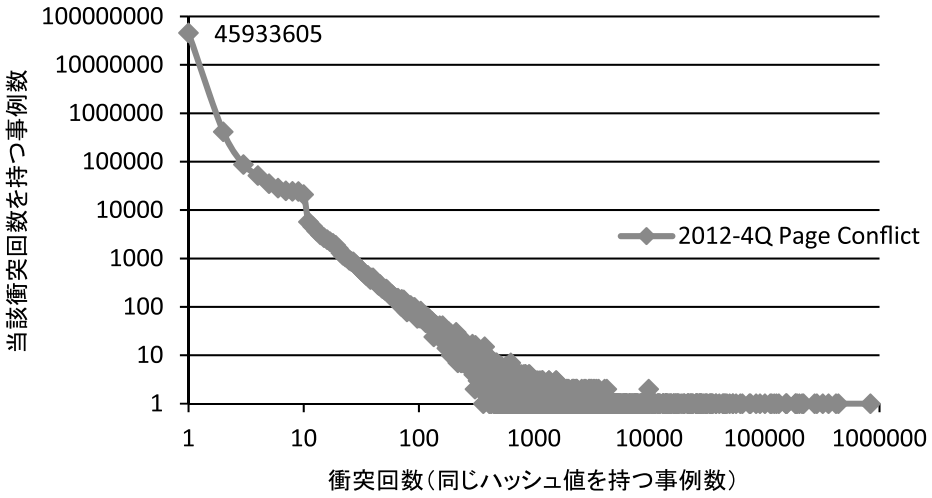

図 4　2012 年第 4 四半期収集ページの重複

図 4 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，当該 URL は robots.txt や，「ソフト 404」と呼ばれるもので，サーバ上にはなく，そのことを 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当該ページがないことを示すコンテンツを返していると判断できる。

表 3　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 69 億 URL	6,610,763,700 66 億 URL	7,064,611,259 70 億 URL	7,222,958,033 72 億 URL
リンク先（異なり）	892,135,930 8.9 億 URL	843,166,672 8.4 億 URL	865,694,816 8.6 億 URL	855,684,918 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 3 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と比べてリンクが約 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。

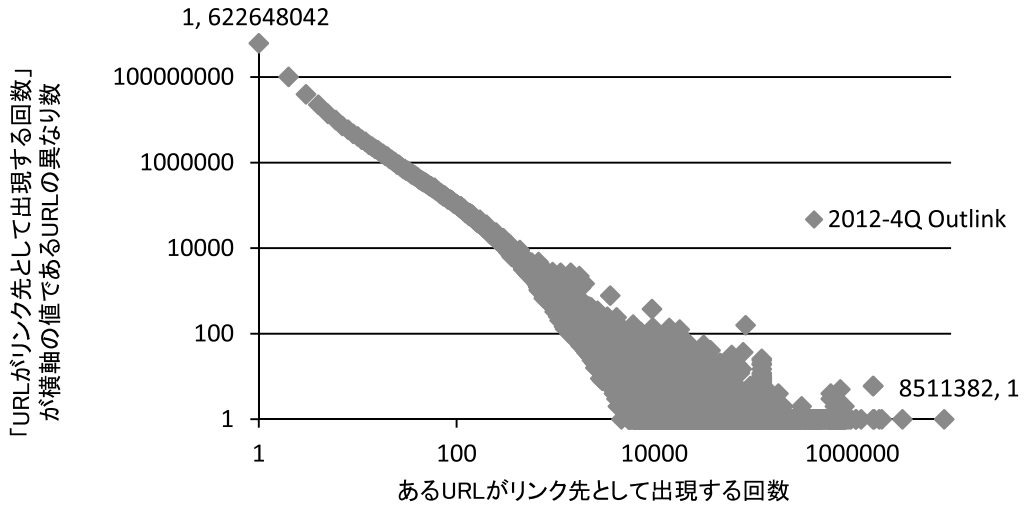

図 5　2012 年第 4 四半期収集リンク先の統計

図 5 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 2 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。

4.3 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。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 4　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	65,736,027,334 657 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

表 4 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 〜 657 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 6 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。

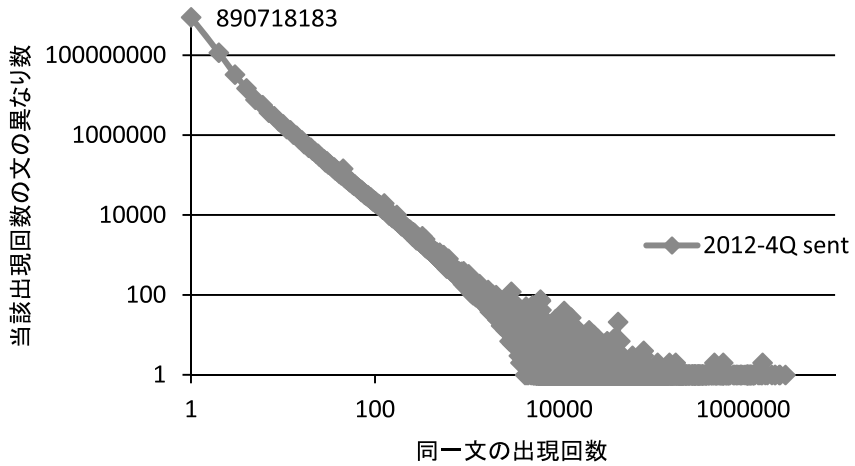

図 6　2012 年第 4 四半期収集文の同一性

表 5　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 書籍 *	451,273	394,782	0.87	186,908	176,792	0.95
OB 書籍 *	222,437	203,467	0.91	23,236	22,242	0.96
OC 知恵袋	42,506	45,082	1.06	369,004	368,572	1.00
OL 法律	38,827	34,761	0.90	6,004	5,864	0.98
OM 会議録	140,422	116,295	0.83	26,948	26,052	0.97
OP 広報紙	257,796	142,660	0.55	22,372	14,636	0.65
OT 教科書	64,100	47,680	0.74	5,992	5,340	0.89
OV 韻文	18,977	17,807	0.94	1,656	1,632	0.99
OW 白書	146,402	126,140	0.86	28,072	24,876	0.89
OY ブログ	117,816	76,205	0.65	224,064	213,480	0.95
PB 書籍 *	392,301	308,827	0.79	176,812	163,476	0.92
PM 雑誌	301,399	227,089	0.75	30,248	25,812	0.85
PN 新聞	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

*LB：図書館（流通実態）書籍サブコーパス
OB：ベストセラー書籍サブコーパス
PB：出版（生産実態）書籍サブコーパス

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ごと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。

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 データの作成を進めている。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。

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 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

	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 6 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。稿末の「付録」の表 7・表 8 にそれぞれのデータの頻度順位の上位 10 件を提示する。研究目的により n-gram データに求めることは異なるであろうから，付録ではそれぞれの性質のみを議論し，どちらがよいかという評価は行わない。今後，共同研究者等に利用してもらいながら適切な組織化とはどのようなものであるのかを画策していきたい。

また, 近年格解析･述語項構造解析についても実用に耐えうる速度と性能のものがオープンソースソフトウェアとして入手可能になりつつある。解析を行うまではそれほど難しくはないが，このような高次アノテーションに対してどのような利用者系を構築するかは重要な研究課題であると考える。

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。4.3 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，4.3 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。4.3 節の末尾や「付録」で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行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。自動解析であって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は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は解明できない問題が何であるのかを示すことが重要だと考える。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。

参照文献

バイドゥ株式会社（2010a）『Baidu ブログ・掲示板時間軸コーパス』，(http://www.baidu.jp/corpus/).

バイドゥ株式会社（2010b）『Baidu 絵文字入りモバイルウェブコーパス』，(http://www.baidu.jp/corpus/).

今井新悟・赤瀬川史朗・プラシャント パルデシ（2013）「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予稿集』，199–206.

情報通信研究機構（2011）『日本語係り受けデータベース Version 1.1』，(https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html #A-8).

Kamide, Yuki and D.C. Mitchell (1999) Incremental pre-head attachment in Japanese parsing. *Language and Cognitive Processes* 14(5/6): 631–662.

河原大輔・黒橋禎夫（2006）「高性能計算環境を用いた Web からの大規模格フレーム構築」『情報処理学会研究報告 自然言語処理』2006-NL-171(12): 67–73.

Kilgarriff, Adam, Siva Reddy, Jan Pomikálek, and Avinesh Pvs (2010) A Corpus Factory for many languages. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 904–910.

国立国会図書館『インターネット資料収集保存事業』，(http://warp.da.ndl.go.jp/search/).

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 (http://www.gsk.or.jp/catalog/GSK2007-C/).

京都大学大学院情報学研究科黒橋研究室（2008）『京都大学格フレーム（Ver 1.0）』，(http://www.gsk.or.jp/catalog/GSK2008-B/).

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(1)]: 15–25.

益岡隆志・田窪行則（1992）『基礎日本語文法』改訂版．東京：くろしお出版.

持橋大地・菊井玄一郎・北研二（2005）「言語表現のベクトル空間モデルにおける最適な計量距離」『電子情報通信学会論文誌』J88-D-II (4): 747–756.

持橋大地・山田武士・上田修功（2009）「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.

Pomikálek, Jan and Vít Suchomel (2012) Efficient web crawling for large text corpora, *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, 39–43.

楽天技術研究所（2010）『楽天データセット』，(http://rit.rakuten.co.jp/rdr/index.html).

関根麻緒（2010）「国立国会図書館のインターネット情報の制度的収集」『図書館雑誌』104(5): 288.

Shinzato, Keiji, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi (2008) TSUBAKI: An open search engine infrastructure for developing new information access. *Proceedings of Third International Joint Conference on Natural Language Processing (IJCNLP2008),* 189–196.

ヤフー株式会社（2007）『Yahoo! 知恵袋データ（第 1 版）』.

ヤフー株式会社（2011）『Yahoo! 知恵袋データ（第 2 版）』，(http://www.nii.ac.jp/cscenter/idr/yahoo/chiebkr2/Y_chiebukuro.html).

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） NICT Web クローラ：(https://alaginrc.nict.go.jp/resources/nictmastar/software/crawler-info/crawleroutline.html).
（4） 日本語ウェブコーパス用ツールキット：(http://code.google.com/p/nwc-toolkit/).
（5） 形態素解析器 MeCab-0.996：(http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html).
（6） 形態素解析用辞書 UniDic-2.1.2：(http://sourceforge.jp/projects/unidic/).
（7） 形態素解析器 JUMAN-7.0：(http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN).
（8） 京都大学テキストコーパス 4.0：(http://nlp.ist.i.kyoto-u.ac.jp/index.php? 京都大学テキストコーパス).
（9） 日本語係り受け解析器 CaboCha-0.67：(http://code.google.com/p/cabocha/).
（10） 日本語係り受け解析器 KNP-4.1：(http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP)
（11） 日本語述語項構造解析器 ChaPAS-0.742：(https://sites.google.com/site/yotarow/chapas/).
（12） NAIST テキストコーパス 1.5：(http://cl.naist.jp/nldata/corpus/).
（13） SVMlin：（多クラスのトランスダクティブ学習が可能）(http://vikas.sindhwani.org/svmlin.html).
（14） BACT：（部分木を特徴量とする決定株を弱学習器としたブースティング）(http://chasen.org/~taku/software/bact/).
（15） コーパス検索アプリケーション「中納言」1.1.0：(http://chunagon.ninjal.ac.jp/).
（16） コーパスアノテーション支援環境「ChaKi.NET」version 2.4：(http://sourceforge.jp/projects/chaki/releases/).
（17） 頻出部分木マイニングプログラム FREQT-0.22：(http://chasen.org/~taku/software/freqt/).
（18） IIPC（International Internet Preservation Consortium）：(http://netpreserve.org/).
（19） ISO 28500:2009, Information and documentation—WARC file format: (http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717).
（20） Wayback Machine—Internet Archive: (http://archive.org/web/web.php).
（21） Open Source Wayback-1.6.0: (http://archive-access.sourceforge.net/projects/wayback/).
（22） Nutch Web Archive eXtension-0.13: (http://archive-access.sourceforge.net/projects/nutch/).
（23） Web Curator Tool-1.6: (http://webcurator.sourceforge.net/).

付録

表 7・表 8 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。

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（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

表 8　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 MeCab/IPADIC 2012-4Q_ 文単位での重複性 排除有	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
	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
	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
	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
	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 MeCab/IPADIC 2012-4Q_ 文単位での重複性 排除無	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
	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
	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
	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』 MeCab/IPADIC ［工藤・賀沢 2007］	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
	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
	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-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，多く見られるようになる。一方，本言語資源で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 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 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常に起きる問題で，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。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。そして人文系の研究者はこのような現象の問い合わせについて何件出現するという回答が欲しいのだろうか？

Page Collection and Linguistic Annotation Issues in Ultra Large-Scale Web Corpus Construction

ASAHARA Masayuki[a]　IMADA Mizuho[b]　YASUDA Sachi[b]
KONISHI Hikari[c]　MAEKAWA Kikuo[d]

[a]Center for Corpus Development, NINJAL
[b]Postdoctoral Research Fellow, Center for Corpus Development, NINJAL
[c]Adjunct Researcher, Center for Corpus Development, NINJAL
[d]Department of Corpus Studies / Center for Corpus Development, NINJAL

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. During the page collection stage, crawling began during the fourth quarter of 2012. We crawled 100 million URLs every three months as fixed-point observations. During the linguistic annotation, normalization (HTML tag removal and character encoding conversion), Japanese morphological analysis (word segmentation and part-of-speech tagging), and Japanese dependency analysis were performed on the data that were crawled in the timespan of one year, specifically from the fourth quarter of 2012 to the third quarter of 2013. In this paper, we present the basic statistics of the crawled data and discuss possible theoretical and practical implications of the language resources. Additionally, we address issues encountered during the page collection and linguistic annotation stages, and offer tentative solutions.

Key words: corpus development, web archive, linguistic information indexing, linguistic analyzers