Web を母集団とした超大規模コーパスの設計

浅原 正幸 (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)

A Design of Web-scale Japanese Corpora

Masayuki Asahara (Center for Corpus Development, NINJAL)
Kikuo Maekawa (Dept. Corpus Studies/Center for Corpus Development, NINJAL)

1. はじめに

国立国語研究所では 2006 年-2010 年の期間に 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』(BCCWJ)(前川 (2007); 前川・山崎 (2008)) を構築し、2011 年より一般公開している。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、国立国語研究所では 2011 年から 6 か年の期間で、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。本発表では、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。

2. 先行研究

Web スケールの言語資源として、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・ウェブ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。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 (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」コーパス（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。

自然言語処理を研究している機関においては、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。

* masayu-a@ninjal.ac.jp

表 1　主な Web スケールの言語資源

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。2007 年 7 月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。月毎に母集団を設定。
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 年公開。以下は 2012 年 8 月公開版について) 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) 、 楽天ゴルフのレビュー (32 万レビュー)、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」コーパス第二弾 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・大学・官公庁	
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。
京大	「京都大学格フレーム」(Ver 1.0) 2009 年 3 月公開。 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	「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。
個人	
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 395GB), N-gram コーパス (文字, 形態素) を配布。

例えば、情報通信研究機構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))。

官公庁においては、国立国会図書館 (NDL) は官公庁自治体のウェブ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。

個人でも矢田 (2010) が形態素解析用辞書 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。

表 1 に、一般に公開されている主な Web スケールの言語資源を示す。さまざま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。

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可用性をあげるためにどのような工夫を行うかについてくわしく説明する。

表 2　超大規模コーパスの概要設計

- 収集
  - クローラ
    - Heritrix 3.1 系
- 構造化
  - 正規化技術
    - nwc-toolkit
  - 形態素解析
    - MeCab/UniDic (国語研短単位, UniDic 品詞体系)
    - CRF++ (国語研長単位, UniDic 品詞体系)
    - JUMAN (益岡・田窪品詞体系)
    - 教師なし形態素解析 (単語分かち書きのみ)
  - 係り受け解析
    - CaboCha/京都大学テキストコーパス
    - CaboCha/BCCWJ アノテーション基準
  - レジスタ分析
    - BCCWJ メタデータ相当情報推定
    - スパムサイト等判定
    - クラスタリングによる文体論的分析
- 利活用
  - 検索アプリケーション
    - 文字列検索 (+レジスタによるファセット分析)
    - 品詞検索 (中納言相当)
    - 係り受け検索 (ChaKi 相当)
  - 語彙表・n-gram データ
    - 語彙表 (出現形, 形態論情報を含む)
    - n-gram データ (基本形, 形態論情報を含まず)
    - 係り受け部分木 (基本形, 形態論情報を含まず)
  - 言語解析器
    - UniDic 未登録語調査
    - 頻度・共起情報を用いた言語解析器の改善
- 永続保存
  - ファイル形式
    - WARC 形式 (ISO-28500)
  - 情報アクセス
    - Open Source Wayback (ハーベスト)
    - NutchWAX (検索)
  - キュレーション
    - Web Curator Tool

我々が構築する予定の超大規模コーパスの概要設計について、**収集・構造化・利活用・永続保存**の四つの観点から解説する：

収集： Web 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。

構造化： Web 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・レジスタ推定を行い、言語コーパスとしての信頼性を高める。

利活用： 構造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。

永続保存： 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。

表 2 に概要設計を示す。以下、各項目について詳説する。

3.1　収集

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。本研究では継続的に収集を行うために、バルク収集が可能なクローラ Heritrix [(1)] を運用する。Heritrix クローラは、wayback machine と呼ばれる Web アーカイブに実績を持つ米国 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・.xml 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。日本語であれば スパムサイト (splog) であろうと機械翻訳結果であろうと収集を行い保存する。外国語などのデータは後述するレジスタ推定などにより定期収集対象から除外するが、収集したものを削除することは行わない。

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 5000 万 URL 規模の第二次収集テストを行い、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 年第四四半期から本収集（第一期）開始し、2013 年 1 月現在、本収集（第二期）を行っている。今後一年かけて同様の本収集を行い、URL の更新頻度推定などを行う。二年目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。コーパスの分布を特徴づける統計量のうち、分散が大きくなるように、また、尖度が小さくなるようにするが、歪度については制御しない。

3.2 構造化

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**、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。

正規化：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」[(2)] 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) [(3)] が公開されている。

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。

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。UniDic

が採用している国語研短単位は形態論的な情報に基づき、単位に斉一性があり、音韻的な情報が豊富であり、音韻論・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、係り受けなどの統語分析を行う研究者は、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ため、益岡・田窪文法に基づく品詞体系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、言語処理の研究者で Web 上に頻出する辞書に登録されない形態を中心に分析するものもいる。

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[4] /UniDic[5] による国語研短単位解析、汎用チャンカー CRF++ [6] による国語研長単位解析、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ほか (2009) の四つを利用する。

係り受け解析：形態論情報において多様な品詞体系・単位が選択されるように、係り受けアノテーション基準もコーパス間で差異がある (浅原 (2013))。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 CaboCha [9] （益岡・田窪品詞体系に基づく形態論情報）・BCCWJ 基準 (浅原・松本 (2013)) に基づいて学習した CaboCha （UniDic 品詞体系に基づく形態論情報・国語研文節単位）の二つを利用し、双方の基準による係り受け木を作成する。

レジスタ推定：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、尖度を小さ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文法的なものを排除するための分類（(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 [10] による境界事例分析と、ランダムフォレスト法やブースティング法[11] 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 (持橋ほか (2005)) の検討を行う。

3.3 利活用

構造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。

検索アプリケーション：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」[12] 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」[13] 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。

語彙表・n-gram 頻度情報：3 カ月おきにクロールするデータに対して構造化を行ったうえで、語彙表 (1-gram 頻度情報; 形態論情報を含む; 出現形に基づく)・文字列上の n-gram 頻度・形態素列上の n-gram 頻度情報 ($n \geq 1$; 形態論情報を含まない; 基本形に基づく)・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。尚、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[14] 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。

言語解析技術への利用：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 永続保存

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)[15] における各国国立図書館の活動動向を見ながら、保存のための構造化を行う。

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[16] で保存する。WARC ファイルは Internet Archive が公開している Wayback Machine [17] と同じ機能を持つハーベストソフトウェア Open Source Wayback[18] と、情報検索システム NutchWAX (Nutch Web Archive eXtension) [19] により構造化し、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)[20] 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、アノテーションを効率的に行う環境を構築する。

最後に、長期保存可能な記憶媒体を機構内外に確保し、収集し構造化したデータの保存に努める。

4. おわりに

本稿では、現在国立国語研究所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」で整備を進めている Web スケールのコーパスの概要設計を解説した。表 3 に現状の工程表を示す。

以下、進捗について示す。2011 年度後半に計画立案を行った。2012 年度は主に収集技術・テキストの正規化技術・形態素解析技術・文字列検索技術・保存技術の調査を行った。収集技術に関しては実際にクローラの運用テストを行いながら運用規則の策定を行い、現在クローラの本運用を開始している。今後定期的に運用規則を見直しながら収集作業をすすめていきたい。またテキストの正規化・形態素解析の構造化環境を構築し、係り受け解析のテスト環境を構築中である。2013 年度は、テキストの正規化技術・形態素解析関連技術を運用レベルにあげ、文字列検索技術の調達を開始する。係り受け解析技術の既存技術については調査を行うとともに年度末までに運用レベルにする。技術調査としてはレジスタ分析技術と品詞・係り受け構造に基づく検索技術を対象とする。2014 年度以降、細部については修正の可能性もあるが、大方はこの工程表に準じて構築をすすめる予定である。

本研究に関する意見・要望・疑問点などについては第一著者まで。

謝辞

本研究は国立国語研究所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」によるものである。本研究を行うにあたり、情報通信研究機構ユニバーサルコミュニケーション研究所の諸氏および統計数理研究所の持橋大地氏よりさまざまな技術指導をいただいた。国立国語研究所コーパス開発センターの諸氏から設計時点での有益なコメントをいただいた。ここに記して謝意を表す。

参考文献

Shinzato, K., T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi (2008). “Tsubaki: An open search engine infrastructure for developing new information access.” *IJCNLP-2008*.

浅原正幸 (2013).「係り受けアノテーション基準の比較」 第 3 回コーパス日本語学ワークショップ.

浅原正幸・松本裕治 (2013).「『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション」 第 19 回言語処理学会年次大会 (NLP2013).

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」情報処理学会自然言語処理研究会 171-12 巻, pp. 67–73.

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/catalog.html.

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/catalog.html.

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

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/nictmastar/resource-info/abstract.html#A-8.
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, pp. 288.
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/.
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, pp. 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), pp. 15–25.
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, pp. 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.
矢田晋 (2010).『日本語ウェブコーパス 2010 (NWC 2010)』, http://s-yata.jp/corpus/.
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.
楽天技術研究所 (2010).『楽天データセット』, http://rit.rakuten.co.jp/rdr/index.html.

関連 URL

(1) クローラ Heritrix-3.1.1：http://webarchive.jira.com/wiki/display/Heritrix/Heritrix

(2) Google Web 日本語 N グラム第 1 版 README：http://www.gsk.or.jp/catalog/GSK2007-C/GSK2007C_README.utf8.txt

(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-0.995：http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

(5) 形態素解析用辞書 UniDic-2.1.1：http://sourceforge.jp/projects/unidic/

(6) 汎用チャンカー CRF++-0.57：http://crfpp.googlecode.com/svn/trunk/doc/index.html

(7) 形態素解析器 JUMAN-7.0： http://nlp.ist.i.kyoto-u.ac.jp/nl-resource/juman/juman-7.0.tar.bz

(8) 京都大学テキストコーパス 4.0：http://nlp.ist.i.kyoto-u.ac.jp/nl-resource/corpus/KyotoCorpus4.0.tar.gz

(9) 日本語係り受け解析器 CaboCha-0.66：http://code.google.com/p/cabocha/

(10) SVMLin：http://vikas.sindhwani.org/svmlin.html（多クラスのトランスダクティブ学習が可能）

(11) BACT：http://chasen.org/~taku/software/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）

(12) コーパス検索アプリケーション「中納言」1.0.5：http://chunagon.ninjal.ac.jp

(13) コーパスアノテーション支援環境「ChaKi」version 2.3：http://sourceforge.jp/projects/chaki/releases/

(14) 頻出部分木マイニングプログラム FREQT-0.22： http://chasen.org/~taku/software/freqt/

(15) IIPC(International Internet Preservation Consortium)：http://netpreserve.org/

(16) ISO 28500:2009, Information and documentation – WARC file formathttp://www.iso.org/iso/catalogue_detail.htm?csnumber=44717

(17) Wayback Machine – Internet Archive：http://archive.org/web/web.php

(18) Open Source Wayback-1.6.0：http://archive-access.sourceforge.net/projects/wayback/

(19) Nutch Web Archive eXtension-0.13：http://archive-access.sourceforge.net/projects/nutch/

(20) Web Curator Tool-1.6：http://webcurator.sourceforge.net/

表 3　超大規模コーパスプロジェクト：工程表

年	2011	2012				2013				2014				2015				2016
四半期	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q
準備	⇒ 計画立案																	
	⇒ 機材調達 (初回)			⇒ (2 回目)				⇒ (3 回目)				⇒ (4 回目)					
収集	(クローラ関連)																	
		⇒ クローラ運用テスト																
				⇒ クローラ本運用開始														
		⇒ 運用規則策定			⇒ 運用規則見直し (初回)				⇒ 運用規則見直し (2 回目)				⇒ 運用規則見直し (3 回目)					
構造化	(正規化)		⇒ 正規化技術調査															
						⇒ 正規化技術運用開始												
	(形態素解析)				⇒ 形態素解析技術調査（既存技術）													
								⇒ 形態素解析運用開始（既存技術）										
	(係り受け解析)					⇒ 係り受け解析技術調査（既存技術）												
								⇒ 係り受け解析運用開始（既存技術）										
									⇒ 係り受け解析技術調査・開発（新規技術）									
	(レジスタ分析)			⇒ BCCWJ メタデータ関連調査														
						⇒ splog 検出技術調査												
								⇒ クラスタリングによる文体論的分析技術調査										
									⇒ 実装・並列化									
										⇒ レジスタ分析技術運用開始								
利活用	(検索アプリケーション)		⇒ 文字列検索技術調査															
						⇒ 文字列検索技術調達開始												
						⇒ 品詞検索技術調査				⇒ 品詞検索技術調達開始								
						⇒ 係り受け検索技術調査				⇒ 係り受け検索技術調達開始								
															⇒ 内部公開			
																⇒ 外部公開		
	(語彙表・n-gram データ)								⇒ 語彙表作成開始									
									⇒ n-gram データ作成開始									
											⇒ 係り受け部分木データ作成開始							
	(言語解析器)							⇒ 未登録語調査 (初回)				⇒ (2 回目)				⇒ (3 回目)		
													⇒ 言語解析器の改善					
保存			⇒ 技術調査															
						⇒ Open Source Wayback 運用開始												
										⇒ NutchWAX 運用開始								
															⇒ 保存媒体の確保			