『国語研日本語ウェブコーパス』の検索系

浅原 正幸 ♣♢　河原 一哉 ♠　武井 裕也 ♠　舛岡 英人 ♠
大場 寧子 ♡　鳥井 雪 ♡　森井 亨 ♡　前川 喜久雄 ♣♢
加藤 祥 ♢　小西 光 ♢
人間文化研究機構 国立国語研究所 ♣ 言語資源研究系 ♢ コーパス開発センター
♠ 株式会社 Preferred Infrastructure ♡ 株式会社 万葉
masayu-a@ninjal.ac.jp

1 はじめに

国立国語研究所（国語研）では、2011 年より 100 億語規模のウェブコーパス『国語研日本語ウェブコーパス』[1] の開発を進めてきた。収集においては約 1 億 URL を 3 か月おきに 1 年間にわたってクロールすることで数百億語規模のテキストデータを四半期ごとに構築した。組織化においては、解析器を用いて日本語テキスト抽出・文分割・形態素解析・係り受け解析を行った。本稿では、収集・組織化を行ったテキストに対する検索系『**梵天 (BonTen)**』について述べる。具体的には、『現代日本語書き言葉均衡コーパス』(BCCWJ)[2] に対する検索系である『**少納言**』[5]・『**中納言**』[6] の文字列検索や『**ChaKi.NET**』[3] の String Search 相当の機能である文字列検索、『**中納言**』の短単位検索や『**ChaKi.NET**』の Tag Search 相当の機能である短単位検索、『**ChaKi.NET**』の Dependency Search 相当の機能である係り受け部分木検索からなる。Web アプリケーションとして動作し、ブラウザがあれば他のソフトウェアのインストールは不要である。フロントエンドは『**ChaKi.NET**』を参考にして、万葉 [1] 社にウェブ上で動作するインターフェイスの構築を依頼した。バックエンドには Preferred Infrastructure[2] 社の『Sedue for Bigdata』[3] を用い、100 億語規模の検索に耐えうる検索系を構築した。さらに、語彙表・n-gram データの公開だけでなく、n-gram データに対する検索系の提供を検討している。

以下、2 節では格納している『国語研日本語ウェブコーパス』の概要について示す。3 節では検索系『**梵天**』の機能について紹介する。4 節ではまとめと今後の公開予定について示す。

[1] http://everyleaf.com/
[2] https://preferred.jp/
[3] https://preferred.jp/product/sfbd/

2 『国語研日本語ウェブコーパス』

『国語研日本語ウェブコーパス』はウェブを母集団として 100 億語規模を目標として構築した日本語コーパスである。ウェブアーカイブの構築で用いられる Heritrix-3.1.1 [4] クローラを運用することで 1 年間 3 か月おきに固定した約 1 億 URL のウェブページを収集している。得られたウェブページは nwc-toolkit-0.0.2 [5] を用いて、日本語文抽出と正規化を行う。正規化時に、文末認定に用いた句点・感嘆符・疑問符を削除する仕様となっている。コピーサイトの問題を緩和するために、文単位の単一化（文の異なりを用いる）を行う。形態素解析器 MeCab-0.996 [6] と辞書 UniDic-2.1.2 [7] を用いて形態素解析を行う。さらに UniDic 主辞規則に基づく [8] 係り受け解析器 CaboCha-0.69 [9] により係り受け解析を行う。

研究の動機として、言語研究への利活用のみならず、次世代に現在の言語使用実態を残すための日本語ウェブアーカイブ構築があげられる。Heritrix が出力する WARC 形式 (ISO 28500:2009) のデータと構造化した拡張 CaboCha 形式のデータは国語研所内のサーバ上に保存（オンラインバックアップ）するほか、未来の研究に資するためにテープに保存（オフラインバックアップ）する。

しかしながら、収集したデータを現在の研究者に提供することが求められている。著作権の問題があり、データをそのまま外部の研究者に提供することは難しい。文字列のみならず、形態論情報や係り受け構造に基づく検索系を構築し、例文とともに元データが含ま

[4] http://webarchive.jira.com/wiki/display/Heritrix/Heritrix/
[5] http://code.google.com/p/nwc-toolkit/
[6] http://mecab.googlecode.com/svn/trunk/mecab/doc/
[7] https://osdn.jp/projects/unidic/
[8] ./configure --with-posset=UNIDIC
[9] https://taku910.github.io/cabocha/

れる URL へのリンクを含めて提示するサービスを構築した。2016 年度に 2014 年 10-12 月期収集データ (2014-4Q データ) を検索系に格納したものを公開する予定である。格納データの基礎統計は表 1 のとおりである。

表 1: 基礎統計:2014-4Q データ

収集 URL 数	83,992,556	8399 万 URL
文数（のべ数）	3,885,889,575	38 億文
文数（異なり数）	1,463,142,939	14 億文
国語研短単位数	25,836,947,421	258 億単位

3 検索系 『梵天』の機能

本節では検索系『梵天』の機能について解説する。基本的に既存の検索系の機能を可能にすることを目標として開発を進めているが、100 億語規模に耐えうる検索系にするために、既存の検索系で利用可能ないくつかの機能については制限がある。

以下では各機能について概観する。参考のため開発中のユーザインターフェイスの画面を示す。公開版ではより使いやすいデザインのユーザインターフェイスを提供する予定である。

3.1 文字列検索

文字列検索はもっとも基本的な検索機能で、検索文字列を指定し適合する文を表示するものである。絞り込み機能として、URL のドメインの末尾 2 パート [10] を指定できる。

『梵天』では『少納言』・『中納言』・『ChaKi.NET』などで利用可能な正規表現（ワイルドカード・文字クラス・否定文字クラスなど）が利用できない。データ自体が文単位で組織化しているために、文をまたいだクエリを発行することができない。さらに過去の検索クエリを保存するヒストリ機能を有しない。

文字列検索は認証なしのものと認証ありのものの 2 つを準備する。認証系を介した利用者には、より高機能な結果表示を提供する。結果表示については 3.4 節に説明する。

[10]Internet top-level domain (例: .com, .jp) と second-level domain (例: .co.jp, ac.jp)。

3.2 短単位検索

短単位検索は、『中納言』の短単位検索や『ChaKi.NET』の Tag Search 相当の機能である。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短単位連接をクエリとした検索が可能である。

図 1 に短単位検索クエリ発行画面を示す。

この例では、サ変名詞・動詞「する」の連用形・接続助詞「て」の 3 つの短単位からなるクエリが指定されている。3 つの形態論情報 Box が配置され、各形態論情報 Box の 2 つの整数値のうち、左が最左出現位置・右が最右出現位置を表す。各形態論情報 Box は出現文字列（表層系）・品詞・活用型・活用形・語彙素・語彙素読み・発音形出現形が指定できる。

図 1: 短単位検索クエリ発行画面

『梵天』では『少納言』・『中納言』・『ChaKi.NET』などで利用可能な正規表現（ワイルドカード・文字クラス・否定文字クラスなど）が利用できない。さらに『ChaKi.NET』で利用可能な Case Intensive（英文字の大文字・小文字の区別をしない）機能が利用できない。

3.3 係り受け部分木検索

係り受け部分木検索は、『ChaKi.NET』の Dependency Search 相当の機能である。係り受け解析結果に含まれる形態論情報・文節内相対位置・文節間相対位置・文節係り受け関係に基づく部分木をクエリとした検索が可能である。

図 2 に係り受け部分木検索クエリ発行画面を示す。

この例では、代名詞の直後に係助詞「は」が後置する文節が、「猫」と判定詞の「だ」が含まれる文節にかかる部分木構造が指定されている。2 つの文節 Box（左：緑・右：オレンジ）内にそれぞれ 2 つの形態論情報 Box が配置され、文節 Box の左上に文節 ID 0 と 1 がふられている。文節 ID 0 が文節 ID 1 にかかっていることが、左の文節 Box の白い小さな箱に指定

図 2: 係り受け部分木検索クエリ発行画面

されている。2 つの文節 Box の間は「<」で結ばれ、これは各文節がこの順で出現していることを表す。緑の文節内の 2 つの形態素の間は「–」で結ばれ、これは各形態素がこの順で連接していることを表す。

Box 間の関係を規定するほかの記号として「^」(文頭 or 文節頭)「$」(文末 or 文節末)を指定することができる。図中の「+」はその外側に Box を追加するための記号である。

3.4 結果表示とダウンロード

検索結果の表示として 通常の KWIC 表示と文節係り受け木に特化した表示の 2 種類を準備する。検索結果は適切な分量のテキストを画面上に表示し、他の検索結果をページネーション機能で遷移して表示する機構が含まれる。

図 3 に文節係り受け木に特化した表示の例を示す。短単位に空白区切りで分割され、文節単位に角括弧でまとめられた表示を行う。各短単位にマウスオーバーすることにより形態論情報がポップアップで表示される。さらに各文節をマウスオーバーすることにより、当該文節(図中黄色背景)・その文節にかかる文節(図中水色背景)・その文節がかかる文節(図中ピンク色背景)の 3 種類を色分け表示する。

検索結果は 2 通りの方法でダウンロードできる。1 つ目は、検索クエリに適合する文(最大 10 万件)の文字列のみを TSV 形式でダウンロードする機能である。2 つ目は、画面上表示されている文の係り受け解析結果を CaboCha 形式でダウンロードする機能である。後者のデータは、『ChaKi.NET』 などで読み込み、検索や統計処理を行うことを想定している。ダウンロードファイルの改行コードは CRLF (Windows), LF(Linux), CR(Mac OS) の 3 種準備し、文字コードは UTF-8 固定とする。

3.5 n-gram データ

Web コーパスの先行事例では、著作権などの問題を回避するために n-gram データを頒布しているものが多い。『国語研日本語ウェブコーパス』でも語彙表 (1-gram データ) や n-gram データを頒布する。収集期ごとのデータの公開を検討している。

また、ssgnc [11] を用いた、N-gram 検索ツールを公開する。N-gram 頻度表に対し、出現順序不問 (Unordered)・出現順序合致 (Ordered)・フレーズ一致 (Phrase)・完全一致 (Fixed) の問い合わせが可能である。図 4 に n-gram 検索ツールの画面例を示す。

『中納言』では n-gram 表示はできない。代わりに BCCWJ の語彙表を Web ページにて頒布しているほか、BCCWJ と『日本語話し言葉コーパス』頻度に基づいて編纂した辞書 [4] が出版されている。**『ChaKi.NET』**は WordList 機能を用いて、自由な文脈長の頻度表や、係り受け関係に基づく共起対リストを作成することができる。

4 おわりに

本稿では『国語研日本語ウェブコーパス』の検索系の基本機能について概説した。発表ではデモを行う。

2016 年度中の公開を予定している。n-gram データと文字列検索のうち通常の KWIC 表示のものについては認証なしのものを公開する予定である。文字列検索(係り受け解析結果表示)・短単位検索・係り受け部分木検索の 3 つの機能については、**『中納言』**の認証系を介しての登録制公開を検討している。公開の前には利用者向けの講習会を行う予定である。

[11] https://code.google.com/p/ssgnc/

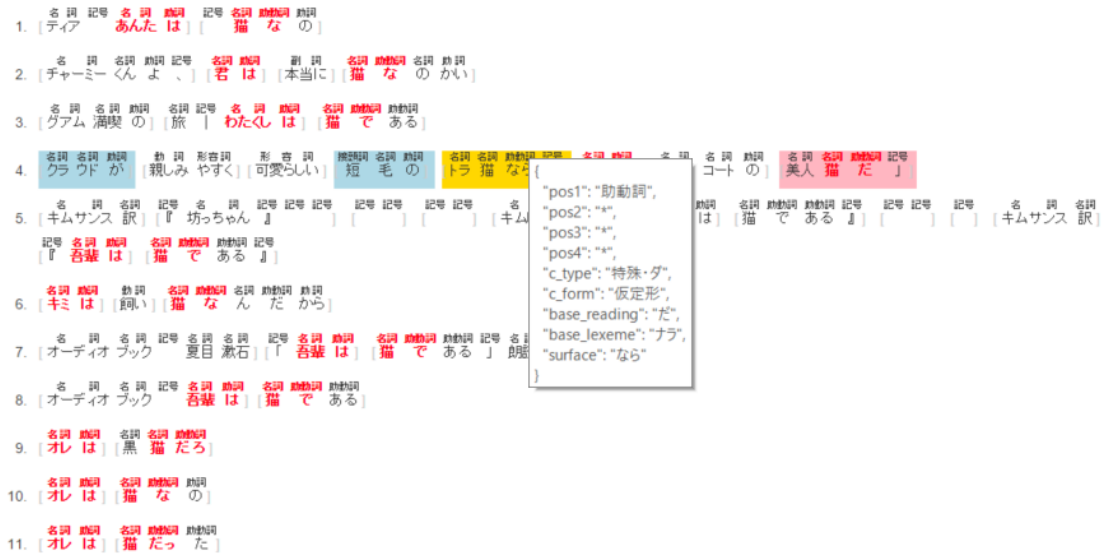

図 3: 文節係り受け木に特化した KWIC 表示

【出力例】

国立 感染 症 研究 所 4320
国立 環境 研究 所 3690
国立 研究 所 1540
国立 情報 学 研究 所 1410
、 国立 感染 症 研究 所 1210
国立 感染 症 研究 所 の 1060
国立 教育 政策 研究 所 849
国立 国語 研究 所 809
国立 環境 研究 所 の 779
国立 健康 ・ 栄養 研究 所 714
国立 衛生 研究 所 646
国立 極地 研究 所 564
...

図 4: n-gram 検索ツール

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] Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan. Alexandria, Vol. 25, No. 1-2, pp. 129–148, 2014.

[2] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced Corpus of Contemporary Written Japanese. Language Resources and Evaluation, Vol. 48, pp. 345–371, 2014.

[3] Yuji Matsumoto, Masayuki Asahara, Kou Kawabe, Yurika Takahashi, Yukio Tono, Akira Otani, and Toshio Morita. ChaKi: An Annotated Corpora Management and Search System. In Proceedings from the Corpus Linguistics Conference Series, 2005.

[4] Yukio Tono, Makoto Yamazaki, and Kikuo Maekawa, editors. A Frequency Dictionary of Japanese. Routledge, 2013.

[5] 国立国語研究所コーパス開発センター. コーパス検索アプリケーション『少納言』. http://www.kotonoha.gr.jp/shonagon/, 2015 年 12 月 25 日確認.

[6] 国立国語研究所コーパス開発センター. コーパス検索アプリケーション『中納言』. https://chunagon.ninjal.ac.jp/, 2015 年 12 月 25 日確認.