情報処理学会論文誌 Vol.59 No.2 299–305 (Feb. 2018)

『国語研日本語ウェブコーパス』とその検索系『梵天』

浅原 正幸[1,a)] 河原 一哉[2] 大場 寧子[3] 前川 喜久雄[1]

受付日 2017年4月21日, 採録日 2017年11月7日

概要：国立国語研究所は言語研究に資する 258 億語規模のウェブコーパス『国語研ウェブコーパス』を構築した．コーパスの構築は，ページ収集・言語解析・保存・検索系の構築の 4 種類の部分工程からなる．本稿では，『国語研ウェブコーパス』を概説するとともに，その検索系である『梵天』の機能について紹介する．この検索系は 100 億語規模のテキストコーパスを文字列だけでなく，形態素列・係り受け部分木に基づく問合せが可能である．

キーワード：コーパス，ウェブアーカイブ，検索系，アノテーション，ユーザインタフェース

'NINJAL Web Japanese Corpus' and Its Search System 'BonTen'

Masayuki Asahara[1,a)] Kazuya Kawahara[2] Yasuko Ohba[3] Kikuo Maekawa[1]

Received: April 21, 2017, Accepted: November 7, 2017

Abstract: The National Institute for Japanese Language and Linguistics, Japan (NINJAL) compiled a web corpus for linguistic research comprising 25 billion words. The project is divided into four parts: page collection, linguistic analysis, development of the corpus concordance system, and preservation. This article presents a corpus concordance system named 'BonTen', which enables a ten-billion-scaled corpus to be queried by string, a sequence of morphological information or a subtree of the syntactic dependency structure.

Keywords: corpus, web archive, search system, annotation, user interface

1. はじめに

国立国語研究所は 100 億語規模の日本語テキストコーパスである『国語研日本語ウェブコーパス』('NINJAL Web Japanese Corpus': 以下 'NWJC') [1] を構築し，検索系『梵天』を介してコーパス中の例文を問い合わせる環境を公開した．このシステムは『現代日本語書き言葉均衡コーパス』('the Balanced Corpus of Contemporary Written Japanese': 'BCCWJ') [2] 用に作成された検索ツール『中納言』[3] の文字列検索・短単位検索（品詞列検索）相当の機能と，コーパス管理システム『ChaKi.NET』[4], [5] の Dependency Search（係り受け検索）相当の機能を有している．フロントエンドのユーザインタフェースは株式会社万葉に委託し，バックエンドは 'Sedue for Bigdata' を用いた検索系を株式会社レトリバに委託した．本検索系は 258 億語からなる NWJC を効率的に検索することができる．

本稿では，NWJC を概説するとともに，その検索系である『梵天』の機能について紹介する．2 章では NWJC の構築過程と基礎統計について示す．3 章では検索系『梵天』の各種機能について述べる．4 章では検索系の評価を行う．5 章にまとめと今後の方向について示す．

2. 『国語研日本語ウェブコーパス』（NWJC）

『国語研日本語ウェブコーパス』（NWJC）は，言語研究に資する言語資源としてウェブを母集団として構築した 100 億語規模のテキストコーパスである．

ページ収集は Heritrix クローラを用いた[*1]遠隔採取（remote harvesting）による．1 年間で 3 カ月ごとに 4 回，1 億

[1] 人間文化研究機構国立国語研究所
NINJAL, Tachikawa, Tokyo 190–8561, Japan
[2] 株式会社レトリバ
Retrieva, Inc., Chiyoda, Tokyo 102–0071, Japan
[3] 株式会社万葉
Everyleaf Corporation, Chiyoda, Tokyo 101–0051, Japan
[a)] masayu-a@ninjal.ac.jp

[*1] http://webarchive.jira.com/wiki/display/Heritrix/Heritrix/

© 2018 Information Processing Society of Japan

URL を固定してクロールする．1 年ごとに，各ページの更新頻度およびリンク先の被リンク数に基づき，再サンプリングを行い，収集を行う．

言語解析は，正規化と形態素解析と係り受け解析からなる．正規化は，HTML タグを取り除き，文字コードの統制を行い，日本語文抽出を行う処理である．収集したページは nwc-toolkit-0.0.2 [*2] を用いて，正規化を行う．このソフトウェアは，先行事例である Google 日本語 n-gram などで採用されている方法を実装したもので，既存のウェブコーパスの構築と同等の処理を行っている．文字コードは UTF-8 に変更したうえで正規化形式 KC により正規化を行う．文境界を句点・ピリオド・エクスクラメーションマーク・クエスチョンマークなどで分割したうえで，切り出した範囲の長さ・文字種比率などにより日本語か否かを判定する．ウェブ上の重複・コピー・定型句の問題に対処するために，収集した日本語文は延べではなく異なりでコーパスを構築する．形態素解析は MeCab-0.996 [*3] と UniDic-2.1.2 [*4] を用いる．係り受け解析器として CaboCha-0.69 [*5] の UniDic 主辞規則 [*6] を用いる．

保存は Heritrix クローラが出力する WARC 形式によりアーカイブを行う．WARC 形式は，国際標準化機構の規格 ISO 28500:2009 で決められた形式で，Open Source Wayback とよばれる閲覧ソフトウェアに格納可能な形式である．

収集したテキストデータは，次章に述べる検索系『梵天』により，文型パターンデータベースとして一般公開した．検索系は 2014 年 10 月から 12 月に収集したデータ（2014-4Q データ）について，文の異なりをとったものを公開する．このデータの基礎統計を**表 1** に示す．

格納する例文中，品詞が「名詞-固有名詞-人名-姓」と「名詞-固有名詞-人名-名」については，表層形・語彙素・発音形などを「=」で匿名化する．これは，個人に対する誹謗中傷表現対策として行うものである．個人名で検索することに制約をかけているが，元テキストの URL を参照することにより，原文を確認することが可能である．

表 1 NWJC の基礎統計：2014-4Q data
Table 1 Basic statistics of NWJC: 2014-4Q data.

URL の数	83,992,556
文（延べ）	3,885,889,575
文（異なり）	1,463,142,939
文節数	8,736,741,719
国語研短単位数（形態素数）	25,836,947,421

*2 https://code.google.com/archive/p/nwc-toolkit/
*3 http://taku910.github.io/mecab/
*4 http://unidic.ninjal.ac.jp/
*5 https://taku910.github.io/cabocha/
*6 CaboCha コンパイル時に `./configure --with-posset=UNIDIC` と指定することで UniDic に適応することができる．

3. 検索系『梵天』の機能

本章では検索系『梵天』の機能について紹介する．基本的な検索機能として，文字列検索・品詞列検索・係り受け検索がある．一般利用者は文字列検索のみが利用できる．それ以外の検索機能（多機能版）は講習会参加者のみに利用を制限している．また，検索結果の表示機能とダウンロード機能についても紹介する．

3.1 文字列検索

『梵天』はコーパスを検索する最も基本的な機能として，文字列検索を提供する．指定した文字列を含む文を返す機能であり，最も高速に結果を返す．**図 1** に一般利用者向けの検索系の画面を示す．URL ドメインの末尾 2 パートを指定して絞り込み検索をすることが可能である．

『梵天』の制約事項として，文字列に対する選言・グループ化・量化（クリーネ閉包）などの正規表現による検索条件指定ができない．さらに，文型パターンを検索系のスニペットとして提示するのみで，文を超える文脈を指定した検索や提示はできない．

3.2 品詞列検索

品詞列検索は，検索系『中納言』の短単位検索機能や『ChaKi.NET』の Tag Search 機能に相当する，形態素列に基づく検索機能である．**図 2** に検索クエリの例を示す．

グレーの箱が文中の形態素に対応する．箱中の上の数字

図 1 文字列検索
Fig. 1 String search query.

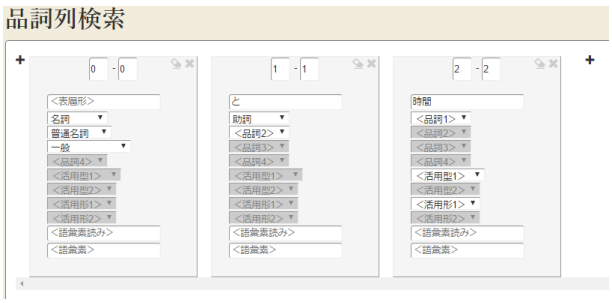


図 2 品詞列検索
Fig. 2 Short unit search query.

© 2018 Information Processing Society of Japan

が，形態素の相対位置を表す．相対位置が ‘0’-‘0’ の形態素が含まれる文節が KWIC の中央に表示される．中央に対して左文脈の形態素は，負の値の相対位置を入れることにより指定する．中央に対して右文脈の形態素は，正の値の相対位置を入れることにより指定する．各形態素に付与される 2 つの相対位置のうち，左の数値は最小（最左）相対位置を，右の数値は最大（最右）相対位置を指定する．

各形態素は次の情報を指定することが可能である：⟨表層形⟩ は文中に出現した形を指定する；⟨品詞 1⟩，⟨品詞 2⟩，⟨品詞 3⟩，⟨品詞 4⟩ は，UniDic 品詞体系の 4 階層を指定する；⟨活用型 1⟩，⟨活用型 2⟩ は，UniDic 活用型体系の 2 階層を，⟨活用形 1⟩，⟨活用形 2⟩ は，UniDic 活用形体系の 2 階層を指定する；⟨語彙素読み⟩，⟨語彙素⟩ は，UniDic の語彙素を指定する．文字列検索と同様に各項目について正規表現を利用することができない．品詞・活用型・活用形は▽のアイコンをクリックすることにより可能な要素を選択することができる．

形態素を示す箱の列の左右にある ‘+’ のアイコンをクリックすることにより，文脈を指定する形態素を増やす．‘×’ のアイコンをクリックすることにより，形態素を示す箱を消す．消しゴムのアイコンをクリックすることにより，形態素の指定項目を初期化する．

図 2 の例では，「と」「時間」の直前に出現する「名詞-普通名詞-一般」を検索する．

3.3 係り受け検索

係り受け検索機能は ‘ChaKi.NET’ の ‘Dependency Search’ 機能に相当するものである．文節係り受け解析結果に対する部分木構造に含まれる形態論情報や相対位置に基づく問合せが可能である．**図 3** に検索クエリ例を示す．緑とオレンジの色付きの箱が文節を表す．各色付きの箱の左上に文節の識別子と当該文節の係り先の文節の識別子を示す．文節（色付きの箱）もしくは形態素（グレーの箱）の前後に相対位置を示す記号を入力する．^ 記号は，文頭（色付きの箱の外側）もしくは文節頭（色付きの箱の内側）を意味する．$ 記号は，文末もしくは文節末を意味する．– 記号は，2 つの文節・形態素が隣接してこの線形順序であることを意味する．< 記号は，2 つの文節・形態素がこの線形順序であることを意味する．箱を追加する際は + 記号を選択することによる．

図 3 の例では 2 つの文節を定義しており，最初の文節は文頭に出現し，「形状詞」+「な」の 2 形態素からなり，その文節の係り先が「時間」を含む文節である部分木構造を問い合わせる．

図 3 係り受け検索
Fig. 3 Dependency search query.

3.4 検索結果表示

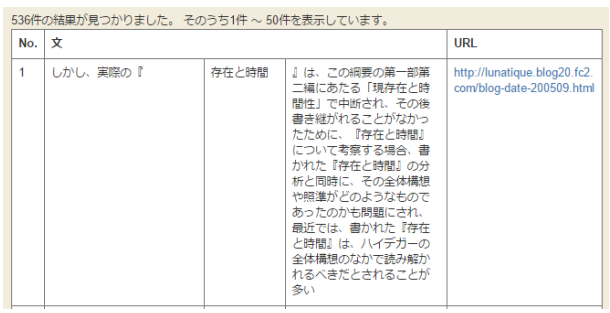
検索結果表示として一般利用者向けの簡易表示（**図 4**）と登録者向けの形態論情報表示（**図 5**）の 2 種類を準備する．いずれの検索結果も，ヒット件数と例文 50 文を表示する．ページ下部の「次へ」をクリックすることにより，51 文目から 50 文を表示するページネーション機能が利用できる．ヒット件数が多く検索に時間がかかる場合は，精緻な検索を打ち切って「約 XXX 件」と表示されることもある．さらに文が含まれる URL が表示される．同じ文が複数の URL に表示されるほか，表示できない数である場合には「XXX 件」のように，いくつかの URL とともに件数が表示される．簡易表示（図 4）は例文の情報を検索結果のキーワードを中央に配置し，左に前文脈を，右に後文脈を配置する．

形態論情報表示（図 5）は，形態素単位に分かち書きを行い，品詞情報（大分類）を形態素の上に表示する．文節境界を [] により示す．形態素にマウスカーソルをかざすことにより，形態論情報をポップアップで表示する．また当該形態素が含まれる文節は黄色でハイライトされる．さらに，黄色の文節に係る文節（係り元文節）を青で表示し，黄色の文節が係る文節（係り先文節）を赤で表示する．

3.5 ダウンロード機能

検索結果は最大 10 万件までダウンロードすることができる．TSV 形式（3 種類）か CaboCha 形式のいずれかの形式でダウンロードが可能である．このうち，文字コード UTF-16LE で改行コードを CRLF にした TSV を拡張子 .csv にしてダウンロードすると，Windows OS 上でダブルクリックで Microsoft Excel のファイルとして開くことができる（**図 6**）．

TSV 形式は「文のみ」「KWIC 文字列」「形態論情報つき」


536件の結果が見つかりました。 そのうち1件 ～ 50件を表示しています。

No.	文			URL
1	しかし、実際の『	存在と時間	』は、この綱要の第一部第二編にあたる「現存在と時間性」で中断され、その後書き継がれることがなかったために、『存在と時間』について考察する場合、書かれた『存在と時間』の分析と同時に、その全体構想や照準がどのようなものであったのかも問題にされ、最近では、書かれた『存在と時間』は、ハイデガーの全体構想のなかで読み解かれるべきだとされることが多い	http://lunatique.blog20.fc2.com/blog-date-200509.html

図 4 検索結果表示（簡易表示）
Fig. 4 Displaying the retrieval results (simple).

© 2018 Information Processing Society of Japan

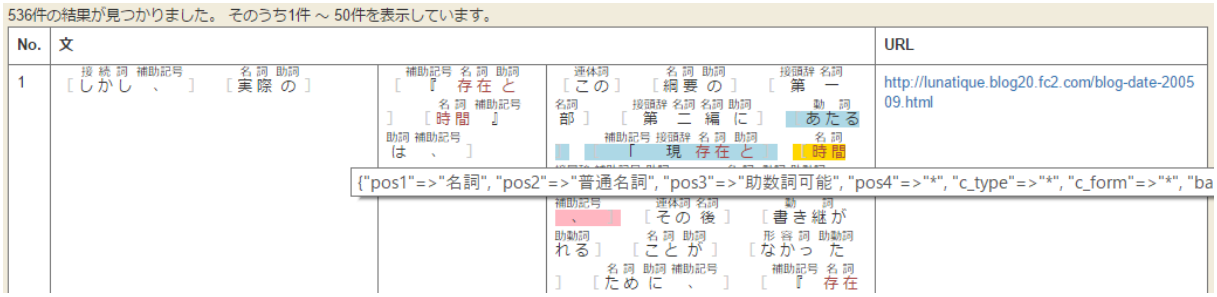


図 5　検索結果表示（形態論情報表示）
Fig. 5　Displaying the retrieval results (rich).

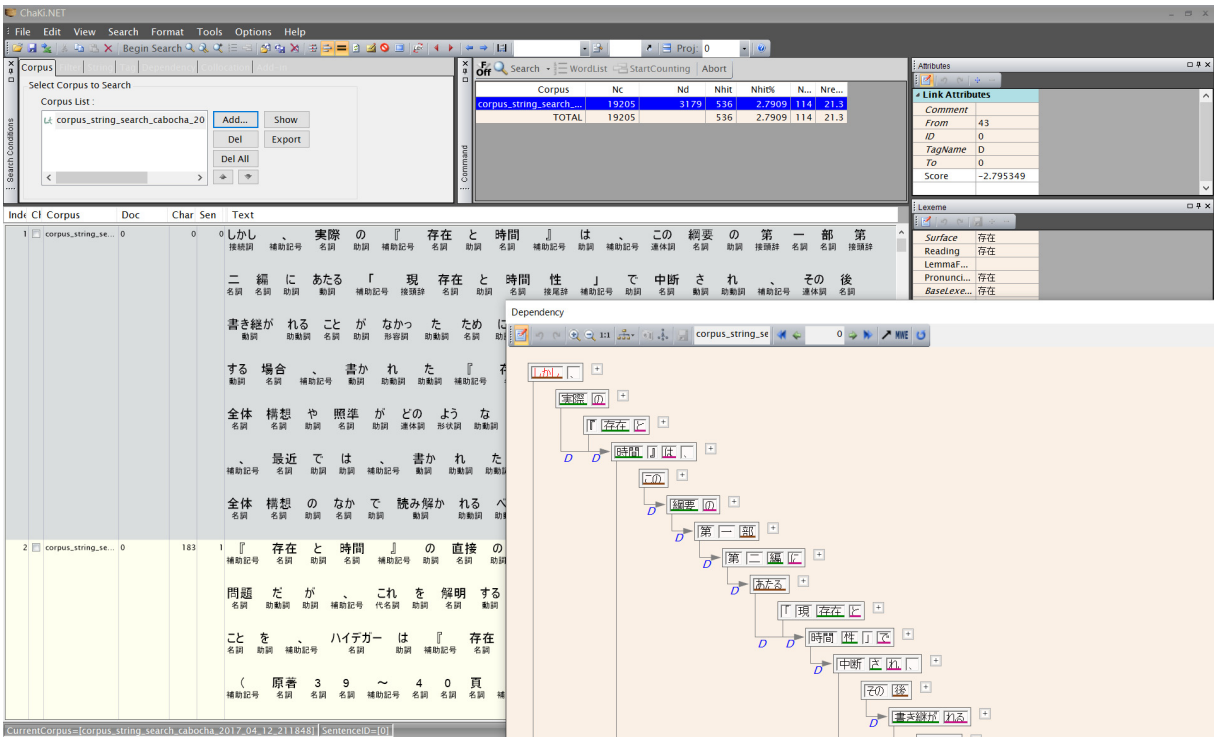


図 7　ChaKi.NET による CaboCha 形式の可視化
Fig. 7　Visualisation of the CaboCha format results by ChaKi.NET.

図 6　ダウンロード機能
Fig. 6　Download function.

の 3 種類が選択でき，最後の「形態論情報つき」はインラインタグとして各種形態論情報を表示することができる．CaboCha 形式は，コーパス管理ソフトウェア ‘ChaKi.NET’ に取り込むことにより，係り受け木の可視化（**図 7**）・正規表現による再問合せ・語彙表作成・n-gram 統計取得・コロケーション展開などができる．

3.6　検索履歴機能

文字列検索・品詞列検索・係り受け検索それぞれに検索履歴機能がある．**図 8** に検索履歴機能の画面を示す．再度まったく同じ検索を繰り返す「再検索」と 2 種類の「編集して検索」機能を有する．「編集して検索（UI）」は，ユーザインタフェース上に検索クエリを再表示する．「編集して検索（json）」は，json 形式の検索式を表示し書き換えることによる．

4.　評価

4.1　検索速度評価

本節では応答時間に基づく検索速度評価を行う．評価は家庭用インターネット回線から研究所内にあるサーバに接続することによって行う．応答時間はブラウザ Chrome の拡張機能 Page load time 1.2.4 に基づく．

表 2 は文字列検索の 2 つの表示機能について応答時間

© 2018 Information Processing Society of Japan

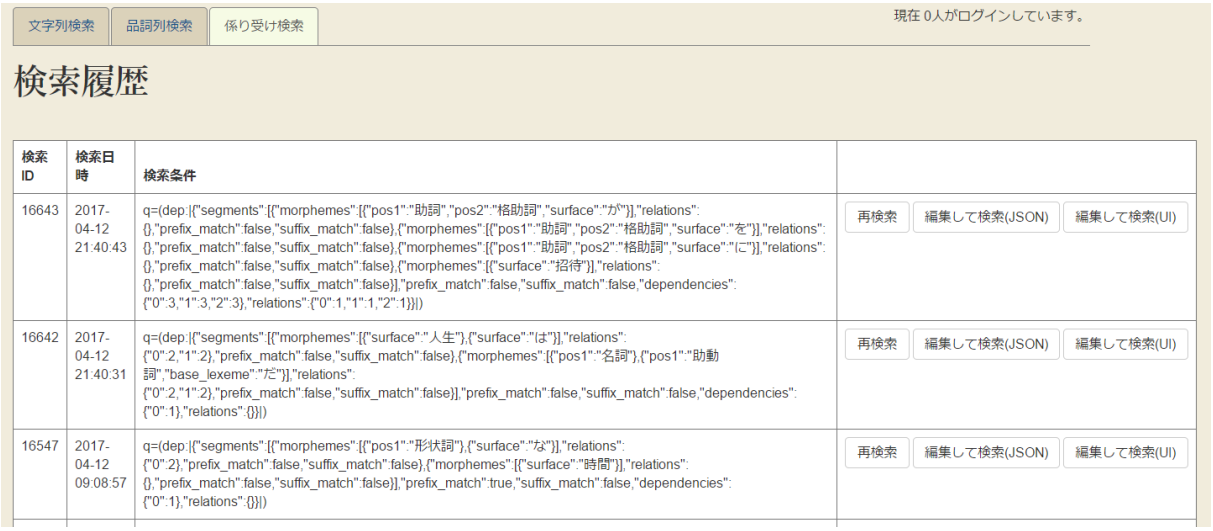


図 8　係り受け検索の検索履歴機能
Fig. 8　History function of dependency search query.

表 3　品詞列検索評価（形態論情報表示）
Table 3　Short unit search query evaluation.

検索クエリ		
ヒット件数	2,469,285 件	12,887 件
応答時間	190 秒	31.7 秒

表 2　文字列検索評価（簡易表示と形態論情報表示）
Table 2　String search evaluation.

検索クエリ	ヒット件数	応答時間	
		簡易表示	形態論情報表示
さくら	498,422 件	1.26 秒	18.3 秒
国立国語研究所	667 件	1.23 秒	4.07 秒
あるくない	40 件	0.78 秒	1.61 秒

を評価したものである．簡易表示は結果を高速に返すが，形態論情報表示は各種アノテーションを組みあげるために時間がかかる．またヒット件数に応じて，応答時間に差が出る．

表 3 に，品詞列検索の評価結果を示す．左の例は「名詞-固有名詞-地名」の品詞に「人」がついている形態素 2 つ組を検索したものである．ヒット件数が約 250 万件であり，応答に約 3 分強かかる．右の例は「研究」「所」の 2 形態素前（左）もしくは 3 形態素前（左）に「国立」が出現する例を検索したものである．ヒット件数が 12,887 件であるが，応答に 31.7 秒かかる．

表 4 に，係り受け検索の評価結果を示す．上の例は「教える」という形態素が含まれる文節に格助詞「が」「を」「に」をそれぞれ含む文節がこの順序で出現する例文を返答する．ヒット件数が 1,439 件であり，応答に 151 秒かかる．下の例は「教える」という形態素が含まれる文節に格助詞「が」「に」「を」をそれぞれ含む文節がこの順序で出現する例文を返答する．ヒット件数が 6,319 件であり，応答に 153 秒かかる．

4.2　検索数

一般利用者版は 2017 年 3 月 7 日に公開した．2017 年 8 月 7 日までの 5 カ月間に約 175,000 件の検索があった．多機能版は講習会参加者のみに公開している．試験公開の 2016 年 9 月末から 2017 年 8 月 3 日までに 33 回講習会を実施し（うち 2 回は Youtube Live で実施），約 380 人が参加した（複数回参加者を含む）．多機能版の利用は 2017 年 4 月以降利用状況を**表 5** に示す．ダウンロードを含む文字列検索についてはよく利用されているが，品詞列検索・係り受け検索については利用方法が浸透しておらず，利用は限定的である．

© 2018 Information Processing Society of Japan

表 4 係り受け検索評価（形態論情報表示）
Table 4 Dependency search query evaluation.

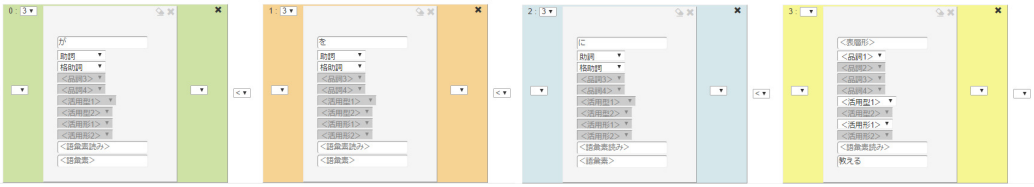
検索クエリ		
ヒット件数	1,439 件	応答時間 151 秒
検索クエリ		
ヒット件数	6,319 件	応答時間 153 秒

表 5 多機能版検索数
Table 5 Use statistics of rich display.

月	文字列検索	品詞列検索	係り受け検索
2017 年 4 月	2,001	64	57
2017 年 5 月	1,734	75	14
2017 年 6 月	2,845	41	6
2017 年 7 月	3,867	279	38

5. おわりに——まとめと今後の展開

本稿では，『国語研日本語ウェブコーパス』（NWJC）とその検索系『梵天』について紹介した．『梵天』は 258 億語の NWJC を，文字列検索・品詞列検索・係り受け検索の 3 種類の方法で検索することができ，検索した結果は 10 万件までダウンロードできる．一般公開版は http://bonten.ninjal.ac.jp/ から利用できる．2017 年 8 月現在，インデックスサーバ 3 台と検索サーバ 2 台で運用している．一般公開版の文字列検索については想定している 3 文字以上の文字列検索の場合は同時に 40 人が人手によるクエリを発行する程度の規模耐性を有している[*7]．また，システムが落ちた場合も自動で再起動する機構が実装されている．

多機能版は現在講習会参加者のみにアカウントを発行している．限定公開している主な理由は，利用規約の同意とデータの利用に関する教育と高機能版の利用方法に関する教育の 3 つからなる．1 つ目の利用規約の同意はダウンロード機能などを利用するために必要な手続きであると考える．2 つ目のデータの利用に関する教育は，元の文を引用する際には必ず元 URL を確認したうえで利用するなど，基本的なネットマナーに関するものである．3 つ目の高機能版の利用方法に関する教育は，検索クエリの組み上げ方に関するものである．表 5 に示すとおり，品詞列検索や係り受け検索に関しては，なかなか人文系の利用者には利用されていない．利用する場合も，検索クエリによっては組合せ爆発により検索件数が膨大になり，サーバに負荷がかかり検索時間がかかったり，場合によってはシステムが落ちたりすることもある．このようなトラブルを未然に防ぐために講習会を受けたうえでの利用をお願いしている．上に述べたとおり，限定公開の範囲もオンサイトの講習会だけでなく Youtube Live による講習会でも利用可能にした．今後，利用規約への同意と動画教材を見ることにより利用可能にしていく予定である．

NWJC の文字表・語彙表・n-gram データ・word2vec [6] 訓練済み分散表現データ nwjc2vec [7] をクリエイティブ・コモンズ・ライセンス表示（CC BY）に基づき配布している[*8]．講習会での要望に基づき，複合語リストや敬語表現のリストなど検索系で得られないデータを所内サーバで整形したうえでオープンデータとして公開している．引き続き，NWJC に基づくさまざまなデータを公開していきたい．

謝辞 本研究は国立国語研究所コーパス開発センター「超大規模コーパスプロジェクト」（2011–2015）によるものです．検索系およびコーパスの整備にあたっては，株式会社ベストシステムズ・デジタルテクノロジー株式会社・徳永拓之氏・武井裕也氏・舛岡英人氏・伊藤敬彦氏・鳥井雪氏・森井亨氏・田中祐樹氏・櫻井達生氏・小田井優氏・依光奏江氏・小芝美由紀氏・榎本誠氏・加藤祥氏・今田水穂氏・小西光氏の支援を受けました．本研究の一部は国際会議 COLING-2016（Demo Session）で発表したものである．

参考文献

[1] Asahara, M., Maekawa, K., Imada, M., Kato, S. and Konishi, H.: Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL,

*7 公開当初，漢字の頻度情報を得るために，1–2 文字による検索を繰り返していた利用者がいた．文字の頻度表（1-gram，2-gram，3-gram）を提供することで対応した．
*8 http://nwjc-data.ninjal.ac.jp/

© 2018 Information Processing Society of Japan

Japan, *Alexandria*, Vol.25, No.1-2, pp.129–148 (2014).
[2] 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*, Vol.48, pp.345–371 (2014).
[3] 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』，入手先 ⟨https://chunagon.ninjal.ac.jp/⟩（参照 2017-03-31）.
[4] Matsumoto, Y., Asahara, M., Kawabe, K., Takahashi, Y., Tono, Y., Otani, A. and Morita, T.: ChaKi: An Annotated Corpora Management and Search System, *Proc. Corpus Linguistics Conference Series* (2005).
[5] Asahara, M., Matsumoto, Y. and Morita, T.: Demonstration of ChaKi.NET – Beyond the corpus search system, *Proc. COLING-2016* (*Demo Session*), pp.49–53 (2016).
[6] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *ICLR Workshop paper* (2013).
[7] 浅原正幸，岡　照晃：nwjc2vec：『国語研日本語ウェブコーパス』に基づく単語の分散表現データ，言語処理学会第 23 回年次大会発表論文集，pp.94–97 (2017).

浅原 正幸　（正会員）

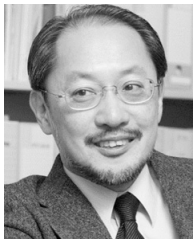
1998 年京都大学総合人間学部卒業．2003 年奈良先端科学技術大学院大学博士後期課程修了．国立国語研究所准教授．博士（工学）．

河原 一哉

2001 年電気通信大学電気通信学部電子情報学科卒業．2016 年株式会社レトリバを創業．代表取締役社長．

大場 寧子

1996 年東京大学文学部卒業．2007 年株式会社万葉設立．代表取締役社長．



前川 喜久雄

1984 年上智大学大学院外国語学研究科博士後期課程（言語学）中途退学．国立国語研究所教授・コーパス開発センター長．博士（学術）．

© 2018 Information Processing Society of Japan