Ｗｅｂ版コーパス検索アプリケーション「中納言」の公開

中村壮範（マンパワージャパン株式会社）
小木曽智信（国立国語研究所）

1. はじめに

国立国語研究所を中心に構築が行われている「現代日本語書き言葉均衡コーパス」（以下、BCCWJ と呼ぶ）は、2010 年度を以て開発をほぼ終え 2011 年中に一般公開を開始する予定である。公開にあたっては、その形式の一つに Web オンラインサービスが予定されている。

これまでに公開されている BCCWJ の検索ツールは「BCCWJ 検索デモンストレーションサイト」や全文検索システム「ひまわり」など、表層の文字列を対象としたものであった。しかし、BCCWJ には形態素解析辞書 UniDic を用いて形態論情報が付与されることになっている。形態論情報が付与されたデータを検索することができるようになれば、表層の文字列にとらわれず、見出し語や品詞などを基に用例を収集することが可能になるため、コーパスを利用する上で有益である。そこで、品詞などの短単位情報を検索条件に指定して検索を行うことができる Web 検索アプリケーション「中納言」を開発した。「中納言」は BCCWJ の人手修正済みコーパスの作成などに用いているコーパス修正ツール「大納言」を基にし、検索機能に特化して、インターフェイスを Web 用に改めたものである（小木曽・中村 2009）。

2. 中納言の特徴

中納言の画面を図 1（次ページ）に示す。中納言の主な特徴は以下の通りである。

1) Web アプリケーションであるため、インターネットが利用できる環境と標準的なブラウザがあれば、特別なソフトをインストールすることなく利用することができる。
2) 「短単位検索」「文字列検索」の 2 種類の検索方法を提供している。「短単位検索」とは BCCWJ に付与された短単位情報について条件を指定して検索を行う機能、「文字列検索」とは検索条件に文字列や正規表現を使用して表層の文字列の検索を行う機能である。
3) 検索結果として、文脈、品詞などの短単位情報のほか、サンプルのタイトルや著者などの情報を表示することができる。
4) 「短単位検索」時には共起条件を指定した検索を行うことができる。
5) 検索結果は、タブ区切りテキスト形式でダウンロードすることができる。

3. 中納言の検索機能

2.で述べたように「中納言」には形態論情報を組み合わせた「短単位検索」と、単位境界を意識せずに利用することが可能な「文字列検索」が可能になっている。以下、これらの検索機能について解説する。

3.1. 検索時の指定項目

中納言の画面上部に表示される操作画面（図 2）では、「検索方法（短単位検索・文字列検索）」「検索対象コーパス」「文脈の文字数」「文脈内の短単位区切り記号」「検索対象（固定長・可変長）」などを指定することができる。

「検索対象コーパス」は BCCWJ のサブコーパスに相当し、ジャンル等によって分割されたものを個別に指定することができる。「検索対象（固定長・可変長）」は、BCCWJ のサンプル取得方法に合わせて設定されたもので、長さを 1000 字に固定した固定長サンプルと、節や章など文章の意味上のまとまりをとりだした可変長サンプルに対応している（BCCWJ の設計の詳細は山崎（2007）参照）。

3.2. 短単位検索

BCCWJ のデータには形態素解析辞書 UniDic による形態論情報が付与されている。UniDic では、表記が異なっても同じ語であれば、一つの見出し語にまとめるという方針を取り、語を階層化した形で辞書登録している。この階層の最上位を語彙素と呼んでおり、この語彙素の下に語形、更に語形の下に書字形という階層が設けられている（伝ほか 2007）。

短単位検索では、この情報を生かした柔軟な検索条件指定が可能になっている。短単位検索時の操作画面を図 3 に示す。

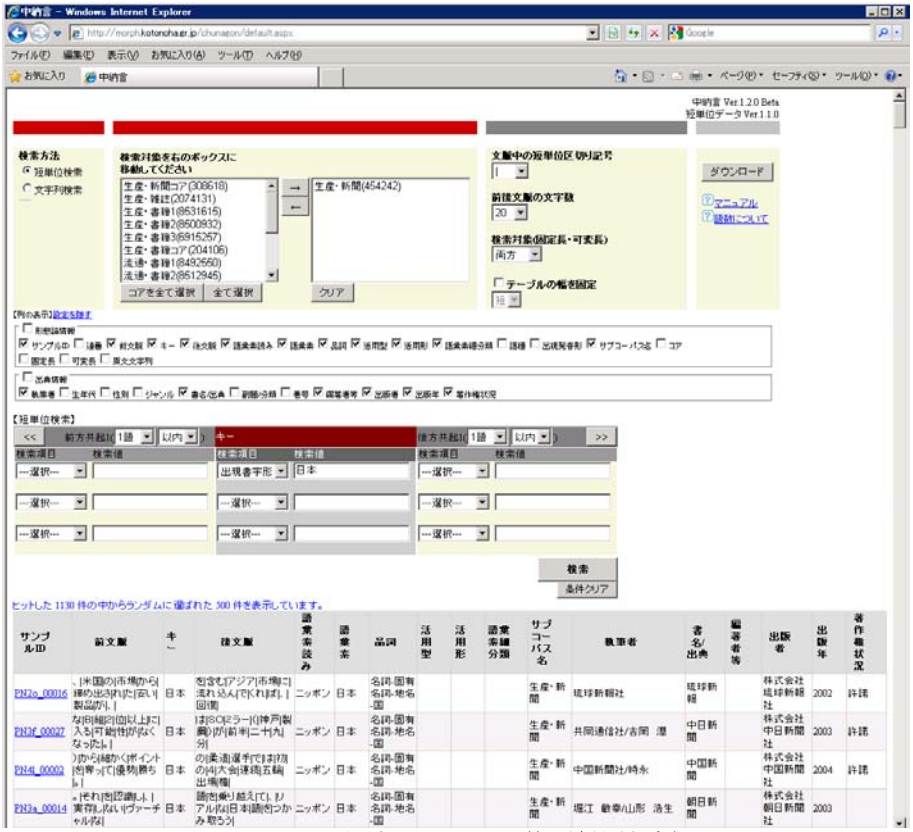

図１　中納言の画面（短単位検索）

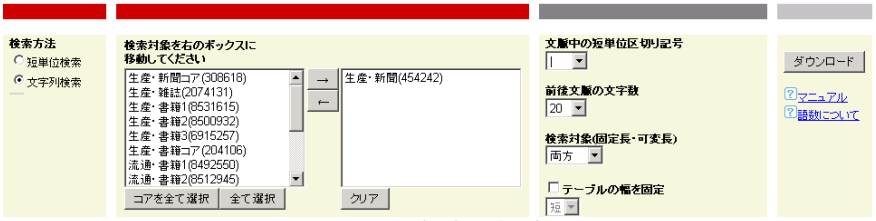

図２　検索時の指定項目

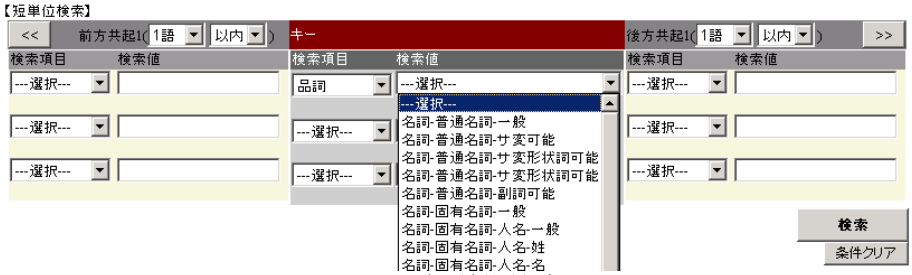

図３　短単位検索

A 検索項目指定

検索項目はドロップダウンにより選択することができる。選択肢には「出現書字形」「品詞」「語彙素」「語彙素読み」「活用形」「活用型」がある。

B 検索値指定

検索項目に「出現書字形」「語彙素」「語彙素読み」を指定した場合には検索値をテキストボックスに直接入力する。検索項目に「品詞」「活用型」「活用形」を指定した場合には、検索値を指定するテキストボックスがドロップダウンリストに変化するため、そこから選択する。選択肢が表示されるので、ユーザーが UniDic の品詞体系を完全に把握している必要はない。

C 共起範囲指定

キーとなる短単位の前方・後方それぞれ 1～5 語まで、またはキーとなる短単位を含む文の文頭から文末までを共起範囲として指定した検索ができる。共起語についても、上記 1)、2)に示した検索条件を指定できる。

UniDic による形態論情報を用いることができるため、図 3 の「短単位検索」の検索項目指定で「語彙素」または「語彙素読み」を指定することによって、検索語の異語形や異表記形を網羅的に検索することができる。例えば、検索条件で検索項目を「語彙素」、検索値を「矢張り」と指定することで、「やはり」「やっぱり」「やっぱ」「やっぱし」「矢張り」など、「矢張り」という語彙素見出しを持つ全ての語形、及びその語形見出しを持つ全ての書字形を検索することが可能である。

3.3. 短単位検索の内部処理

ここで中納言の短単位検索時の内部処理について説明する。短単位検索時は処理の高速化のために様々な検索補助用のデータを使用しているが、ここではコーパス全体からランダムサンプリングした小規模なデータベースを利用した高速化について説明する。

全短単位データが格納された normalDB に対し、normalDB からランダムにサンプルを抜き出して normalDB の約 1/100 のレコード数になるようにした smallDB を用意する（図 4）。condition1～3 はそれぞれ短単位検索の前方共起 1・キー・後方共起 1 に入力された検索条件である。

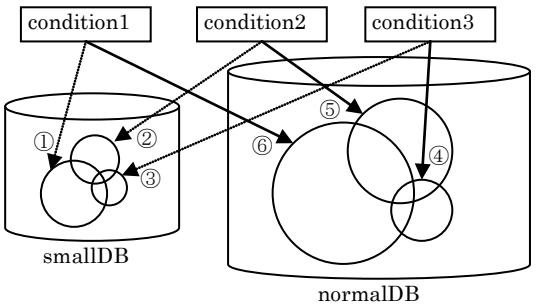

図 4 短単位検索処理の概念図

この上で、次のような手順を踏むことにより高速化を実現している。

1) condition1～3 の順に smallDB 内を検索し、それぞれの検索条件のヒット件数を求める（①～③）。ここで仮に condition3 の検索ヒット件数が最も少なく、condition1 の検索ヒット件数が最も多いという結果が得られたとする。また condition1～3 での共起検索を行い、smallDB におけるヒット件数を求める。
2) smallDB でのヒット件数が最も少ない検索条件（condition3）が normalDB においても最もヒット件数の少ない検索条件であると仮定して、condition3 で normalDB 内を検索する（④）。
3) condition3 に隣接する検索条件（condition2）で normalDB 内の 2)で求められた範囲についてのみ検索を行う（⑤）。
4) 残りの検索条件（condition1）で normalDB 内の 3)で求められた範囲についてのみ検索を行う（⑥）。

このように検索条件が複数ある場合に、検索ヒット件数が少ないと考えられる条件から検索を行うことで、検索の初期の段階から検索対象をある程度絞ることができる。これによって condition2、condition1 の検索が効率的に行われることになり検索処理が高速化される。

また、smallDB を使用することで各検索条件のヒット件数の概数を高速で求めることができるほか最終的な検索ヒット件数の概数も高速で求めることができるので、検索ヒット件数が膨大になる場合の回避処理などを検索処理内に組み込んでいる。

3.4. 文字列検索

中納言のもうひとつの検索方法に文字列検索がある。文字列検索では検索したい文字列を指定することで短単位の境界を意識せずに文字列を全文検索することができる。したがって、短単位

の区切りが分からない場合に、まずは文字列検索によって短単位の区切りを調べ、次に行う短単位検索での語の検索条件指定を行いやすくする、といった短単位検索の補助的な使い方をすることができる。

4. 検索条件の保存と再利用

現在のところ、中納言の短単位検索の検索条件を保存する方法としては、検索条件が入力された画面をキャプチャして JPG などの画像ファイルとして保存するか、テキストファイルへのメモのような形で保存するしかない。また、保存した検索条件で再度検索を行うためには、保存した検索条件を中納言の画面上で再度入力または選択する必要があり、この作業自体が作業者にとって手間になるだけでなく、入力ミス・選択ミスが起こりやすくなるため、検索の再現性という点で問題がある。

そこで、検索条件の指定方法を記述する簡易言語（X-CQL）を規定して、この形式による検索条件のエクスポート・インポートを可能にするよう準備を行っている。これにより検索条件の保存や検索結果の再現が容易になる。記述には XML 形式を用いている。

X-CQL の記述例として「助動詞「らしい」(接尾辞ではない)が名詞を連体修飾する用例」を抽出するための X-CQL を以下に示す（記述方法は検討中のものであり今後変更される可能性がある）。

```
<x-cql application="中納言" version="1.0.1">
<corpus selected="PB OB LB"/>
  <condition0 品詞="助動詞%" 語彙素読み="ラシイ" 活用形="連体形" />
  <condition1 品詞="名詞%" />
</x-cql>
```

これにより、検索対象コーパスの指定、キーとその前後の形態論情報による条件指定など、中納言の画面上で行える条件指定をすべて記述することができる。

実際に短単位検索において X-CQL を使用する場合には、次のような手順によることになる。

1) 中納言の画面上で検索条件を入力すると、X-CQL が画面上に表示される。作業者はこれをテキストファイルにコピー&ペーストすることで検索条件の保存を行う。
2) X-CQL はテキストエディタなどを使用してユーザーが独自に記述することもできる。
3) X-CQL は中納言の短単位検索モードの画面上から取り込むことができる。中納言は XML パーサにより X-CQL のチェックおよび変換を行い、検索条件を画面上で入力および選択した場合と同様に検索処理を行う。

5. おわりに

以上、中納言の詳細について述べた。中納言は 2009 年 9 月下旬から特定領域研究「日本語コーパス」のメンバーに対して公開を開始している。現時点での検索対象となるデータは「BCCWJ 領域内公開データ（2009 年度版）DVD」の XML データ約 8000 万語である。1 億語以上の本格的な公開は 2011 年 7 月頃を予定している。

参考文献

小木曽智信・中村壮範（2009）『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装』国立国語研究所内部報告書 LR-CCG-08-04

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・原裕（2011）国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』形態論情報規程集　第 4 版』

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (2007)「コーパス日本語学のための言語資源－形態素解析用電子化辞書の開発とその応用－」『日本語科学』、vol.22、 pp.101-123.

山崎誠（2007）「『現代日本語書き言葉均衡コーパス』の基本設計について」『特定領域「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集』、pp.127-136.

小木曽智信・中村壮範（2010）「「現代日本語書き言葉均衡コーパス」のための形態論情報データベースについて」『第 16 回公開シンポジウム「人文科学とデータベース」論文集』、pp.45-52.

関連 URL

KOTONOHA 検索デモンストレーションサイト　http://www.kotonoha.gr.jp/demo/

全文検索システム『ひまわり』（国立国語研究所「言語データベースとソフトウェア」）
http://www2.ninjal.ac.jp/lrc

付記

本発表は科研費・特定領域研究「日本語コーパス」による成果の一部を含むものである。