「現代日本語書き言葉均衡コーパス」を対象とした全文検索システム

曽根孝明，小原京子，斎藤博昭
慶應義塾大学理工学部
sone@nak.ics.keio.ac.jp

1　はじめに

本稿では,「現代日本語書き言葉均衡コーパス」[1][1]のモニター公開データを対象とした全文検索システム（BCCWJ KWIC on Web）について報告する．

本システムはブラウザベースの全文検索を提供することによって，言語学者の言語分析，特に語法・句法研究を支援することを目的とする．従来のデスクトップアプリケーションではなくウェブアプリケーションとすることにより，ユーザーによるインストールの手間を省き，メンテナンスも容易なものとすることができる．

本稿では文末指定，KWIC 表示との切り替え，正規表現による検索，検索対象文書の選択などシステムの機能とその実装について紹介する．BCCWJ KWIC on Web は実際に日本語フレームネットプロジェクト[2] における言語分析や，意味アノテーション対象文の選定に使用されているものであり，機能面での改良にはアノテータの意見が反映されている．

2　システム使用イメージ

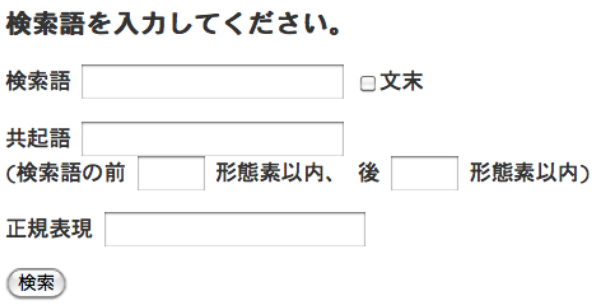

図 1: 検索語入力画面

[1]以下 BCCWJ とする
[2]http://jfn.st.hc.keio.ac.jp/

図 1 に示すのが，検索語の入力画面である．検索語を入力し，検索ボタンをクリックすることで，検索結果が表示される．検索語入力画面では，文末指定，共起語の指定，検索結果の絞り込みのための正規表現の指定ができる．詳細についてはそれぞれ 3.3 節，3.4 節，3.5 節で紹介する．

図 2 が，検索結果の画面である．デフォルトでは，KWIC 形式での表示となっている．結果の表示形式については，3.7 節で説明する．

3　機能の紹介

本節では，BCCWJ KWIC on Web の機能について紹介する．

3.1　活用を考慮にいれた検索

本システムは検索語として形態素の基本形を入力とし，用言の場合，全活用形を網羅的に検索する．例えば，図 2 からもわかるように,「嬉しい」で検索をした場合,「嬉しい」の他に「嬉しく」「嬉し」などにもヒットする．これは，言語分析やアノテーションの対象としては，終止形だけでなくすべての活用形を必要とするためである．

3.2　文情報の表示

選択した文の MeCab による形態素解析結果，CaboCha による構文解析結果，BCCWJ 中での前後の文を表示する．これらの機能により，形態素解析と構文解析の結果を言語分析者が目で確認することができ，また，代名詞の照応関係など文脈を把握することができる．

図 2 からも分かるように，これらの情報はブラウザの下部に表示され，画面遷移を伴うことなく確認することができる．

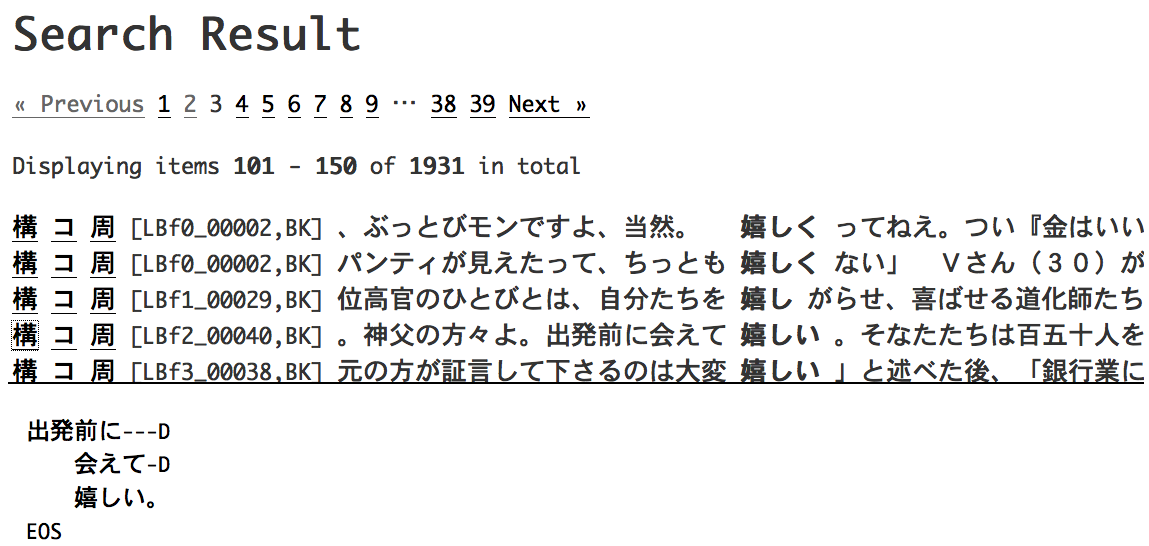

図 2: 「嬉しい」を検索した時の結果

3.3 文末指定

BCCWJ KWIC on Web では「検索語が文末に出現する」という条件での検索が可能である．目下「文末に出現する」とは，文区切りである「。」（句点）から 3 形態素以内に検索語があらわれることとしている．

この条件を指定をすることで検索結果から主節以外に検索語が出現する文を除去することができる．例えば用言に対しアノテーションを行う場合には，それが主節に出現する文のみを抽出することが可能である．

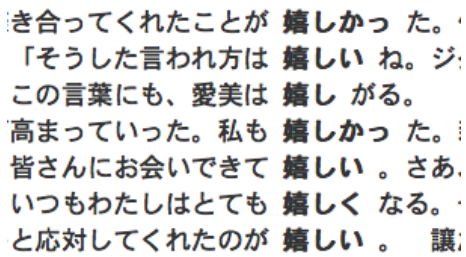

図 3: 文末指定をした場合の表示

例えば,「嬉しい」で検索したとき，文末指定をしないと,「自信たっぷりな葉月の心を、少しでも揺るがせたことが**嬉しく**て、秋乃は、気分よく葉月のそばを離れた。」や「老人は**嬉し**そうに目を細めながら、ゆっくりと飲み始めた。」などが結果に含まれてしまうが，文末指定をすることで，これらの文を除外できる．

3.4 共起語の指定

検索語と共起する形態素を指定して検索することができる．その際，検索語の前後何形態素以内に出現するかを指定できる．

例えば，検索語として「嬉しい」，共起語として「ば」が検索語の前 1 形態素以内に出現するという条件で検索した場合，図 4 のようになる．

3.5 正規表現による絞り込み

正規表現による絞り込みが可能である．

正規表現としては，Ruby と同等の正規表現が使用できる．例えば，

本当 (は | に).*嬉し

という正規表現によって絞り込めば,「本当」のあとに「は」か「に」が続き，0 回以上の任意の文字が繰り返されたあとに,「嬉し」が続くものが取り出せる．

例えば，この条件で絞り込みを行うと「入手したときは**本当に 嬉し** く」や「**本当は**メチャクチャ **嬉し** いのである」などが該当することになる．

3.6 検索結果からランダムに文を選択

アノテーション対象文を選択する際に，検索結果すべてに目を通すのは現実的ではない．そこで検索結果からランダムに文を選択するか否かと，ランダムに選択する文の数を設定できるようになっている．

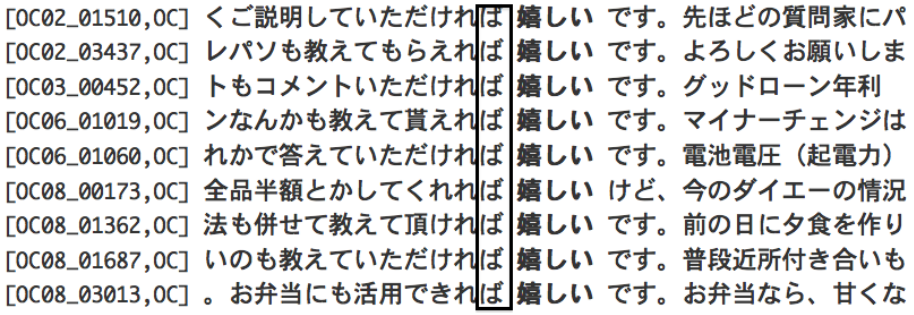

図 4: 共起語を指定しての検索

3.7 KWIC 表示の切り替え

検索対象の形態素が中心に表示される KWIC(KeyWord In Context) 形式 (図 5) と，検索対象の形態素を含む文全体が表示される文単位での表示 (図 6) を選択することができる．

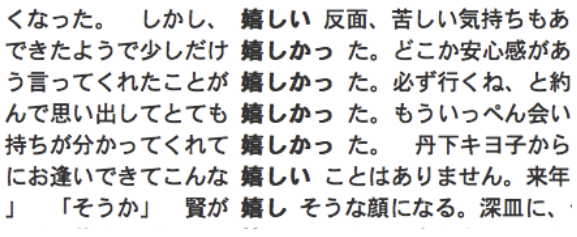

図 5: KWIC 形式での表示

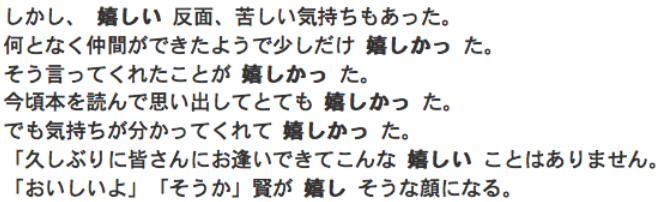

図 6: 文単位での表示

3.8 検索対象文書の選択

BCCWJ には，「国会会議録」，「書籍」，「白書」，「Yahoo!知恵袋」の 4 ジャンルのデータが含まれている．

設定によって，検索対象ジャンルを 1 つ，または複数選択できる．

4 実装上の工夫

MeCab により形態素解析を行い，転置インデックスを作成している．転置インデックスとは，各単語ごとに出現する文書とその位置を記録してあるリストである．転置インデックスを作成することで，特定の単語を含む文を探すためにコーパス全体を走査する必要がなくなる．

インデックスは，通常の検索用のインデックスとは別に文末検索用のインデックスも作成し，高速化をはかっている．

転置インデックスを格納するデータベースとしては代表的な Key-Value ストアである Tokyo Cabinet[3]を使用している．Tokyo Cabinet を使用した理由としては，高速に動作すること，直感的に使用できること，バックアップが容易であることがあげられる．

Web アプリケーション・フレームワークとしては Ruby on Rails[4]を用いている．

5 従来ツールとの比較

従来の検索用ツールはウェブアプリケーションではなく，「ひまわり」[5][2] などデスクトップアプリケーションとして構築されているものが多い．本システムは，ウェブアプリケーションとして構築されている．この場合のメリット・デメリットは以下であると考えられる．

メリット

デスクトップアプリケーションの場合，各ユーザーがインストールやアップデート作業を行う必要があ

[3] http://1978th.net/tokyocabinet/
[4] http://rubyonrails.org/
[5] XML 文書を対象とした全文検索システムであり，BCCWJ モニター公開データ配布 DVD にも収録されている．

る．これはコンピュータでの作業に慣れていないユーザーにとっては負担である．また，ツール作成側にとっても，インストールやアップデート用ドキュメントの作成の負担が生じる．一方，ウェブアプリケーションの場合，ブラウザを使って特定の URL にアクセスするだけでよい．ブラウザは通常の OS であれば備わっているため，別途インストールする必要はない．このため，ユーザーにとっての負担はほぼ存在しないに等しい．

さらに，メンテナンスの観点からみてもサーバー側のみを一括で管理すれば良く，研究メンバー内でのツールやコーパスのバージョンの相違などの問題が生じない．特に BCCWJ のように更新される可能性のあるコーパスの場合，一括で管理できる利点は大きい．

デメリット

インストールが必要なデスクトップアプリケーションと異なり，Web サイトにアクセスすることで誰でも全文検索ツールが使えてしまう．これはコーパスの利用規約等の観点から問題がある．そのため，パスワード等を設定しアクセス可能なユーザーを厳重に管理する必要がある．

6　今後の展望

現状では，本システムが検索語として入力にとっているのは各形態素の基本形である．そのため，検索を行った場合，すべての活用形が結果に含まれる．これは網羅的に検索できるという点で利点がある．

その一方で，未然形である「走ら」だけを対象にすることは簡単ではない．さらに,「にもかかわらず」などの複数形態素にわたる文字列，形態素の一部分を構成している文字列なども検索対象にすることは難しい[6]．

以上の理由から，現状の検索方法の他に，表層文字列を対象とした検索方法を追加することを考えている．

また，現状では出版年月・出版社・ジャンルなどの書誌情報，著者の性別や出生年度など著者情報を検索対象にしていない．この点に関しても今後実装予定である．

[6] これらの問題は共起語の指定や正規表現による絞り込みを使えば可能であるが，検索速度，ユーザの利便性等の観点から望ましくない．

参考文献

[1] 前川喜久雄. KOTONOHA 『現代日本語書き言葉均衡コーパス』 の開発. 日本語の研究, Vol. 4, No. 1, pp. 82–95, 2008.

[2] 山口昌也, 田中牧郎. 多様な構造化テキストに対応した全文検索システム「ひまわり」(デモンストレーション, 日本語学会 2004 年度秋季大会研究発表会発表要旨). 日本語の研究, Vol. 1, No. 2, pp. 144–145, 2005.