W-5-3

デモンストレーション

アラステア バトラー　　窪田 愛　　窪田 悠介
国立国語研究所　　国立国語研究所　筑波大学

ajb129@hotmail.com, ai.kubota@ninjal.ac.jp, kubota.yusuke.fn@u.tsukuba.ac.jp

1　はじめに

デモンストレーションでは、コーパスの利用を助けるためのユーザーフレンドリなインターフェースについて解説する。これは、あらかじめアノテーション体系その他の予備知識が無い人でもコーパスデータにアクセスできるよう、いくつもの手段を提供することを試みるものである。それらの手段とは、とりわけ以下の 3 つである。

- 2 節で解説するパターン・ブラウザ。主要な句、節、構造体のリストを列挙したものである。
- 3 節で解説する KWIC/コンコーダンス・ブラウザ。文字列にもとづく検索の結果を表示したものである。
- 4 節で解説する、原テクストの各文に対し統語解析情報をリンク付けしたもの。

続いて 5 節で述べるように、該当する文のアノテーション結果について、様々なタイプの情報を選んで表示できるようにする。

検索を行うには 2 つのやり方がある。1 つは単純に文字列を入力するものであり、単語分割に関する知識すら必要でない。もう 1 つは XPath (XML ドキュメントの特定部分の指定に用いられる言語で、木構造に対するクエリとして使用される) による解析木を用いるもので、これにはあらかじめの定義されたパターンの中から選ぶものと、利用者が自分で作るものとがある。後者について、6 節で解説する。

以上のように多様な解析木の表示のうち、1 つのものから他のものへと移ることは簡単に行える。例えば、解析木に相当する文の前後の文脈が原テクストでどうなっているかを即座に調べることが出来る。あるいは、検索結果として得られた解析木に対し利用者が改変を加えることで、新しい解析木パターンによる検索が容易に行える。

2　パターン・ブラウザ

解析木を検索する直接的な方法として、あらかじめ与えられた XPath 解析木のリストから選んでクリックするやり方がある。これにより、句や節がどのように使用されているかを知ることが出来るし、特定の構造体について調べることも出来る。例えば、主名詞に相当するトレースが直接目的語の役割を果している関係節の用例をコーパスから抽出することが可能である。

図 1 において、左端のリストにあらかじめ用意された句や構造体のパターンが掲げられており、その 1 つをクリックすることでマッチする文が右上部に表示される。ここでは機能表示の ADV (副詞句または副詞節) が選ばれており、その結果、該当する文が 20 文ずつ表示される。その中から 1 つの文を選ぶ

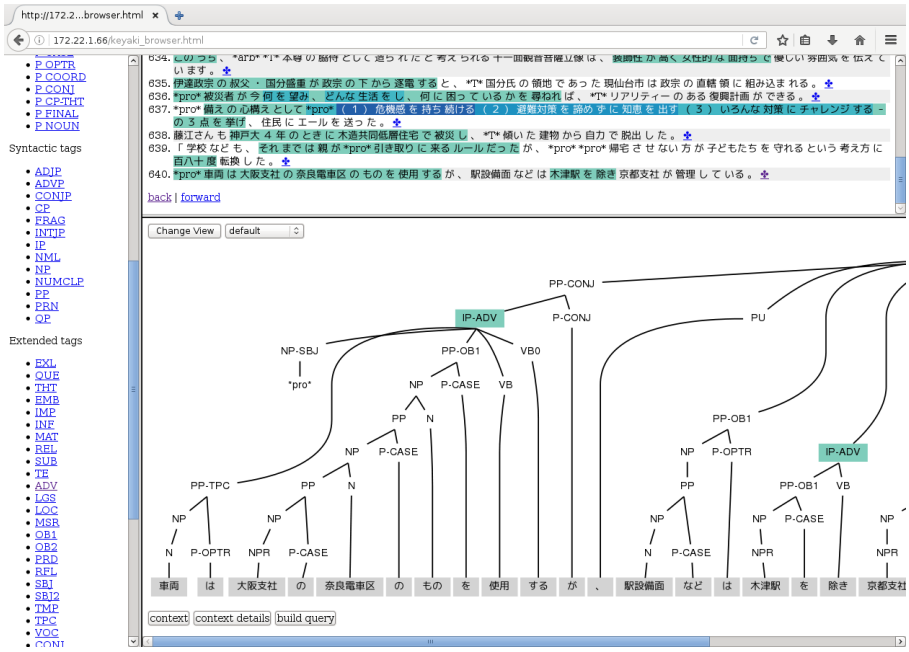

図 1: パターン・ブラウザ

と、右下部にその統語解析木があらわれる。検索対象である IP-ADV (ここでは複数出現する) がハイライトされている。

3　頻度表示および KWIC

文字列を入力することにより、コーパスの中でそれがどのようにアノテートされているか、また解析の種類ごとにその頻度を知ることも出来る。これは特に、当該の文字列がどのように単語分割されるかを知るのに用いることが出来る。

以下では、単語分割に関して曖昧な文字列「だろう」を例として取り上げる。図 2 を参照のこと。

図 2: 「だろう」の頻度と KWIC 表示

図の右上にあるボタンにより、検索対象の文字列の解釈として、Liberal, Character, Mine, および Strict の 4 つのモードの中から選ぶことが出来る。それぞれの違いは以下の通りである。

Character: これがディフォールトである。入力された文字列の最初の字が単語の最初であり、文字列の最後が単語の最後であるという以外、単語分割に関しての何の条件も課されない。「だろう」に対して

は、[だ][ろ][う]、[だろ][う]、[だ][ろう]、および [だろう] の 4 通りの分割が理論上可能なことになる。図 2 では、[だろ][う] が 29 例、[だろう] が 55 例データ中に出現するとの結果が得られている。他の分割の例はあらわれない。

Liberal: 文字列の最初および最後の字が必ずしも単語の最初および最後と一致しなくともよい。

Strict: 単語分割に関して、利用者が厳格な条件を課して使用する。例えば、「だろう」と入力すると [だろう] の例のみを検索し、また「だろ　う」と入力すると [だろ][う] のみを検索する。

Mine: Strict のように厳格な単語分割条件を使用するが、文字列の最初および最後は単語の境界と一致しなくてもよい。

また、これらのボタンの左方にある Fine をチェックすると、図 3 のように機能タグ SUB や MAT をともなう、より詳しい統語解析情報が表示される。

図 2 や図 3 の上方左端の頻度を表す数字をクリックすれば、図の下部のように KWIC (keyword-in-context) 形式でより詳しいアノテーション情報が得られる。図 2 を下方にスクロールすると図 4 となる。このようにして、[だろう] はイ-形容詞および動詞に後続する MD としてのみ出現すること、また [だろ][う] は名詞またはナ-形容詞に後続する AX としての「だろ」および MD としての「う」からなることが分かる。

図 3: 「だろう」のより詳しい解析情報表示

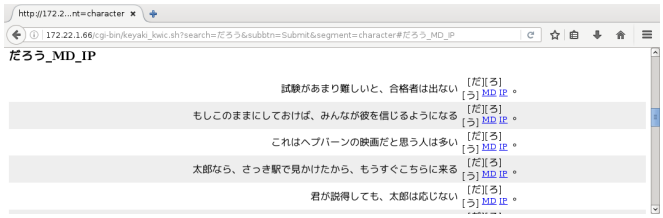

図 4: 「だろう」の KWIC 検索結果

4　原テクストからの解析木表示

コーパスにアクセスするもっとも直接的なやり方は、ジャンルごとにソートされたテクスト中の文からリンクをたどっていく方法である。図 5 にテクストのリストを示す。

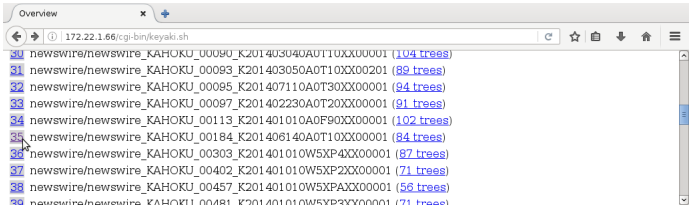

図 5: テクストのリスト

図 5 の左端の番号をクリックすると、図 6 のように原テスクトを見ることができる。図の左端の番号をクリックすることによって、それぞれの文の解析木を表示できる。

これに対して、図 5 の右端の木の数の表示をクリックすると、図 7 のようにアノテーション情報 (名詞の品詞および統語レベル投射情報) をともなう原テクストを見ることができる。解析木へのリンクも付加されている。

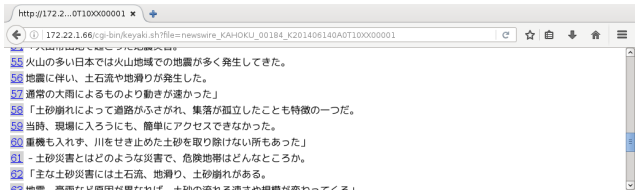

図 6: テクストの内容

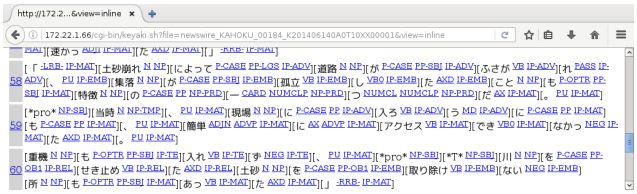

図 7: テクスト中のアノテーション情報

5　解析木の表示

以上では、様々な手段によりどのようにして文の統語解析情報が得られるかについて述べた。得られた統語解析情報にもとづいて、当該の表現があらわれる文脈を見ることもできるし、それを再利用して新たな検索を行うこともできる。また図 8 に示すように、ディフォールトの解析木表示の他に、ほかの表示方法をプルダウン・メニューで選ぶこともできる。

図 9 に、もう 1 つの表示方法の例を示す。動詞「入ろ」のノードにおいて、格フレームが文中の名詞句や空要素に相当するインデクスと共に示されている。また、主語に相当する空要素 PRO が示されている。ディフォールトではこれらの情報は表示されない。また、通常用いるフラットな木構造に代わって、二分木にもとづく表示も可能である。これを図 10 に示す。

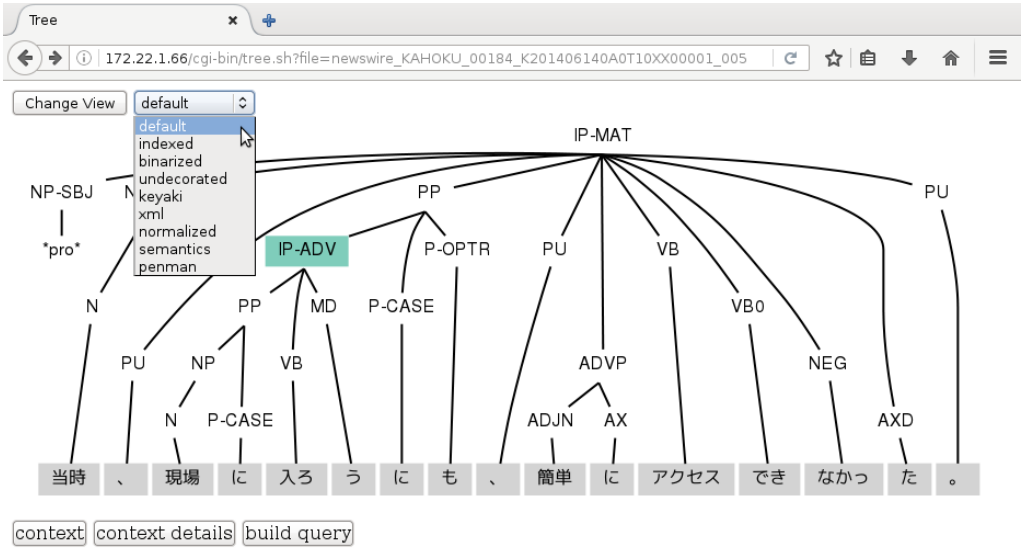

図 8: ディフォールトの解析木情報

6　解析木の検索

解析木の表示の下部には build query というボタンがある。これを押すことによって同じ木のテーブル形式の表示があらわれ、新たな検索のために利用することができる。一例として、図 8 のテーブル形式の表示は図 11 となる。

これにもとづき、利用者はプルダウン・メニューから選び句、節、や単語を表すワイルド・カード、さらには否定を組み合わせて検索用クエリを作成することができる。図 12 に、IP-ADV に関連するプルダウン・メニュールを示す。メニューの内容は上から順に、以下の意味である。

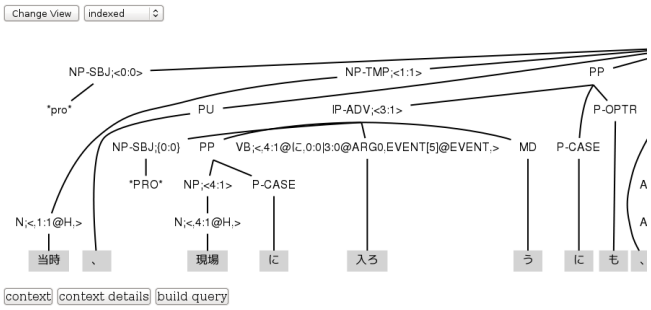

図 9: インデクス付きの解析木の表示

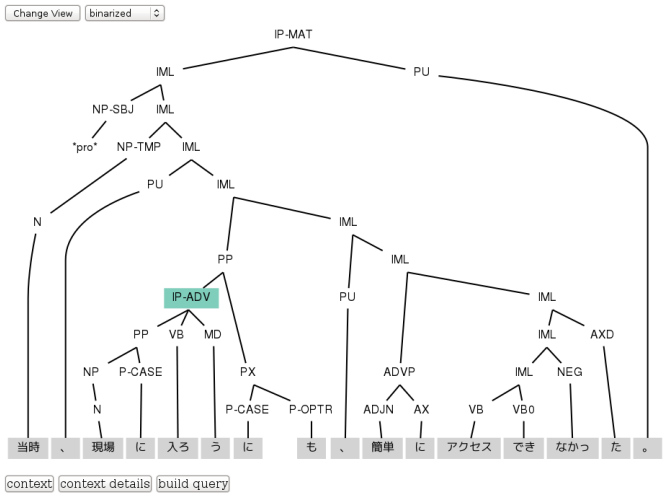

図 10: 二分木の表示

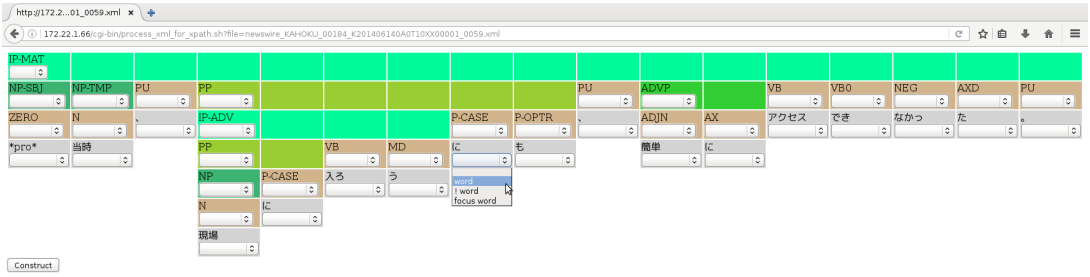

図 11: 検索用解析木作成のためのフォーム

(a) 検索パターンからこのノードを除外する (ディフォールト)。
(b) ノードに付けられたタグのすべて選ぶ。すなわち、IP-ADV がパターンに含まれる。
(c) より一般的に、IP となる構成素を検索する。
(d) IP-ADV の構成素を支配しない PP を検索。
(e) より広く、IP を支配しない PP を検索。
(f) IP-ADV ではない構成素を検索。
(g) IP 以外の構成素を検索。

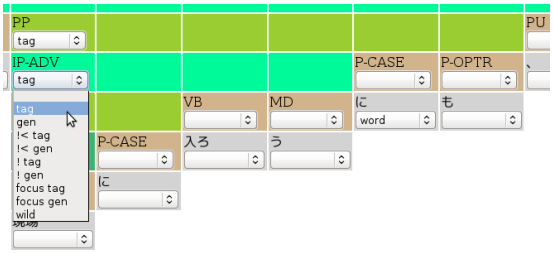

図 12: 検索用解析木作成の詳細

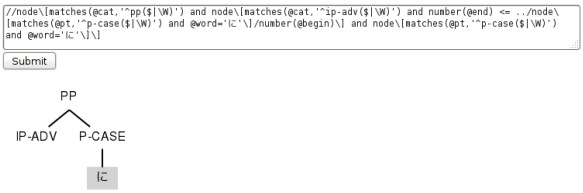

図 13: 解析木を用いた検索

(h) 検索結果中の IP-ADV をハイライトする。
(i) 検索結果中の IP をハイライトする。
(j) ワイルドカードによる検索。

Construct のボタンを押すと、このようにして規定したノード、単語、およびそれらの間の関係に相当する XPath のパターンが作成される。このパターンは図 13 のように表示され、さらに利用者が手を加えることも可能である。Submit のボタンを押して検索を行う。

検索の結果を図 14 に示す。マッチする文のリストが並び、検索対象部分がハイライトされている。また、完全な解析木を表示するためのリンクも提供される。

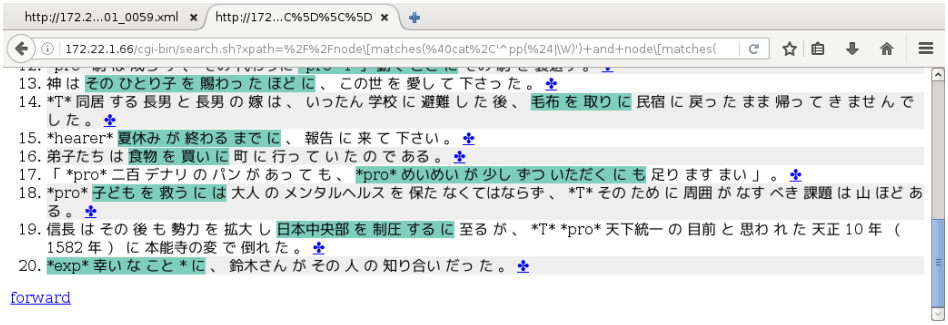

図 14: 検索の結果