# タグ付きコーパスの管理/検索ツール「茶器」の現状*

松本 裕治，浅原 正幸，河部 恒，高橋由梨加
奈良先端科学技術大学院大学
{matsu, masayu-a, kou-k, yurika}@is.naist.jp

投野 由紀夫,　大谷 朗,　森田 敏生
明海大学　大阪学院大学　総和技研
y.tono@meikai.ac.jp,　ohtani@utc.osaka-gu.ac.jp,　morita@sowa.com

## 1　はじめに

近年の統計的言語処理の進展により，高精度の形態素解析／統語解析が現実的になり，種々の言語処理応用に利用され始めている．一方，言語解析をそのまま他システムに利用するのではなく，言語研究の資料として解析済みデータを観察したり，言語処理のさらなる精度向上のためのタグ付きデータ作成を目指すには，大規模なテキストデータの解析結果の閲覧や，しばしば発見されるタグ付け誤りの検出や修正のための支援環境の構築が重要である．

本稿で紹介するタグ付きコーパス管理／検索ツール「茶器」が目指すのは，言語学研究を支援するための種々の言語表現（単語列，品詞列，文節係り受け構造，および，それらの組合せ）を柔軟に検索するための機能と，一定の誤りパターンの網羅的な検索と修正のための機能を提供する支援システムであり，平成 15 年より 3 年計画で科研費の支援を受けて，研究を続けている．本稿では，茶器の現在の機能と今後の予定について紹介する．なお，本システムの公開情報は，http://chasen.naist.jp/hiki/ChaKi/ にあり，システム本体が入手可能となっている．

## 2　コーパス管理/検索ツールに求める機能

タグ付きコーパスを管理し検索するシステムにとして，我々が必要と考えている機能をまとめる．

### 2.1　タグ情報の種類および検索機能

文書コーパスに対して，現在のシステムが想定しているアノテーション情報，および，それらがどのような形式で検索対象となるかを以下にまとめる．なお，形態素タグおよび統語タグについては，形態素解析システム「茶筌」[5]，および，係り受け解析システム「南瓜」[4] の出力を想定している．基本的に検索対象は文である．

**書誌情報**：コーパスの書誌情報 (コーパス名，作者，出典等の情報)，および，文の属性情報 (話者，文脈情報等)．現在は，これらの情報は，構造化されたフォーマットを仮定しておらず，文字列データとしてコーパスごとあるいは文ごとに格納される．これらの情報は，文字列部分一致による検索の対象となる．

**文字列**：コーパスを構成する生の文データ．任意の文字列，および，基本的な正規表現による検索の対象となる．

**単語 (形態素) 列**：文を構成する形態素列．各形態素には，表層形，読み (発音)，原形，品詞 (品詞細分類)，活用型，活用形などの様々な文法情報を持たせることができ，それら任意の部分情報の指定した検索を可能とする．

**複合語，固有表現，文節，基本句**：形態素と統語構造の中間に位置する構造のアノテーション．文節は日本語の統語単位であり，基本句 (base phrases) は英語の統語単位と仮定している．複合語は，形態素と文節 (基本句) の中間に位置すると考えている．日本語においては，単語の単位の認定を一律に行うことが難しく，応用によっても，基本となる単語の考え方は異なる．我々が対象としている辞書 [2] は，登録されている語が複合語の場合，その構成要素となっている形態素へのポインタ列を定義することができる．現在予定している仕様では，形態素情報がタグ付けされたオリジナルのコーパスと，すべての形態素の構成語からなるもっとも

*Progress Report of ChaKi: An Annotated Corpus Maintenance/Retrieval Tool, Yuji Matsumoto, Masayuki Asahara, Kou Kawabe, Yurika Takahashi, Yukio Tono, Akira Ohtani, Toshio Morita

細かい単位の形態素列としてのコーパス，の２つの視点により検索を可能にすることを考えている．固有表現 (named entities) は，形態素や文節と直接の包含関係を持たない単位である．すなわち，固有表現は１つ以上の形態素列からなり，文節を越える場合もあれば，文節の途中を区切りとするものもあり得る．固有表現タグは，他の文法情報とは別のデータベーススキーマによって記述される．それぞれの構造を構成する任意の形態素 (列) を用いた検索を想定している．

**統語構造** ： 現在は，日本語については文節係り受け構造，英語や中国語については，単語あるいは基本句の係り受け構造を仮定している．係り受け構造の任意の部分構造を記述したパターンによって検索を可能とする．

### 2.2 検索結果の表示機能

基本的な検索形態として，文字列検索，形態素 (列) 検索，係り受け構造検索，固有表現検索を想定しており，検索結果は，KWIC(Key Word In Context) 形式で文ごとに表示される．係り受け構造については，文節列による表示以外に木構造による表示機能を提供する．

### 2.3 統計解析機能

検索結果に対し，表層形，原形，品詞などの視点を指定して，出現頻度，比率などの基本的な統計情報を表示する機能．検索結果と周辺の他の単語との共起尺度として，相互情報量を始めとする何種類かの統計値を計算し，提示する機能.

### 2.4 コーパス管理機能

**辞書との連携** ： 形態素タグ付きコーパスは，辞書へのポインタ列として表現される．これにより，コーパスへのタグ付けの基本となっている辞書に含まれない形態素がコーパスに出現することがなくなり，逆に，タグ付けによってコーパスに新たに出現することになった形態素は必ず辞書に含まれることが保証される．

**コーパスの誤り修正** ： コーパス中にタグ付け誤りが発見された場合に，上記の検索機能を利用して，類似の誤りを発見することを支援する．検索されて集められた誤り事例集合を，一括して，あるいは，部分的に選択して修正する機能を提供する．

### 2.5 その他の機能

上記以外に想定している機能，および，特徴についてまとめる．

**多言語対応** ： 特定の言語に限定せず，様々な言語のコーパにも対応可能であること．現在は，日本語と英語を対象にしているが，中国語への適用を予定している．

**誤り検索の支援** ： 現在は，タグ付け誤りの検出に特化した機能の構築は予定していない．しかし，機械学習を利用した誤り検出については既にいくつか提案があり，我々にも経験があるので [7]，外部の誤り検出システムとのインタフェースを構築する形で誤り候補抽出の機能を取り込みたいと考えている．

**配布に対する制限** ： 本システムは無償で配布が可能なソフトであることを前提としている．現システムでは，フリーソフトである関係データベースシステム MySQL[1]をコーパスの格納と検索に利用しており，その他の部分はプロジェクト内で構築しているので，システム全体としてもフリーソフトウェアとして配布する予定である．また，プラットホームとしては，利用人口の多い Windows 上で稼動するシステムとして開発している．

**言語処理ツール** ： 共通に利用可能なタグ付きコーパスの量は多くないため，現実の利用としては，利用者が個別に所有しているタグ付きコーパスを対象とするか，あるいは，利用者が所有する生コーパスに自動タグ付与したコーパスを対象に本システムを用いることを想定している．現時点では，茶筌 [5] と南瓜 [4] を日本語の分かち書き＋品詞タグ付けと係り受け解析に用い，これらのシステムの解析出力をそのまま取り込むためのフィルターを用意する．また，英語や中国語コーパスのタグ付けの自動化のために，茶筌を英語/中国語対応に拡張するとともに，英語辞書，中国語辞書の作成に着手している [3].

## 3 「茶器」の現状

前節で挙げた機能のうち，現在完成している機能について本節で説明する．

### 3.1 タグ情報の種類と検索機能

文字列，形態素列，および，係り受け関係による検索と検索結果の表示機能が既に実装されている．文字

[1] www.mysql.com/

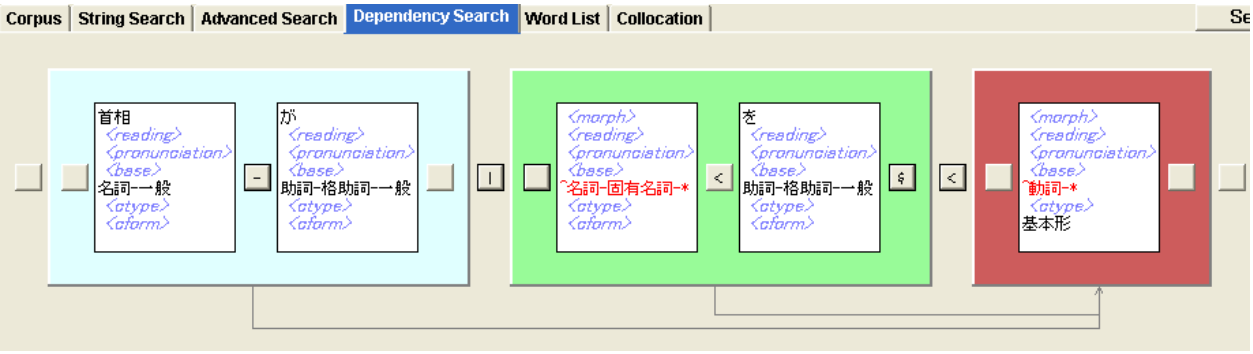


図 1: 係り受け構造の検索要求の例

列，形態素列による検索画面と検索結果の表示に関する基本機能については，前年の報告 [6] を参照されたい．図 1 は，係り受け構造を指定した検索要求の記述の例を示している．図では，文節を囲むボックスと形態素を表すボックスが入れ子になって表現されており，各文節には自動的に異なる色が割り当てられている．この例は 3 つの文節からなり，最初の文節は，「名詞-一般」という品詞を持つ「首相」という表層形の語と「助詞-格助詞- 一般」の「が」を含み，これらが連続して現れる (これらの形態素間の “—” という記号が直接の連接を表す) ことを記述している．2 つ目の文節は,「名詞- 固有名詞」と「助詞-格助詞-一般」の「を」を含み，これらが必ずしも連続する必要はないが，この順序で現れる (これらの形態素間の記号 “<” は順序のみの制約を表す) ことを記述している．最後の文節は，活用形が「基本形」の「動詞」を含むことだけが指定されている．形態素や文節の前後にある小さいボックスは，それらの絶対位置や相対的な位置関係を示すための正規表現に類似の制約を記述している．このように表現された 3 つの文節間の係り受けが，下部の矢印によって示されている．この例では，最初の 2 つの文節が最後の文節に係るという係り受け構造を指定している．この検索要求により，このような係り受け構造を含む例文だけが検索結果として得られる．

### 3.2 検索結果の表示機能

検索結果の表示は，KWIC 形式で，一文一行に表示される．形態素情報については，図 1 の形態素ボックスに示されているように多数の情報があるが，KWIC 表示部では，2 種類までの情報を同時に表示できる．形態素に関する全情報を見たい場合は，マウス位置にある形態素の詳細をバルーン表示する機能や，別 window に表示する機能がある．文字列以外の検索については，検索要求中のいずれか一つの形態素が中心語として指定され，KWIC 表示では，それが中心位置に置かれる．中心語以外の検索要求にマッチした形態素は，それぞれに対応する色指定ができ，指定された色で表示されるので，視覚的に対応が取りやすくなっている．形態素列の表示では形態素ごとに，係り受け解析の表示では文節ごとに異なる色を用いることができる．係り受け木を表示する機能はまだ実装されていない．

### 3.3 統計解析機能

中心語の前後 5 単語について，出現頻度，相互情報量，log-log score，Z score などの統計的共起尺度の表示が可能である．また，中心語を起点とする左右の N-gram 出現数のカウントを行う機能も提供されており，形態素に基づく簡単な統計処理を行うことができる．前後文脈の出現頻度表は，エクセルにエクスポートすることもできるので，検索結果のさらなる加工や統計処理を行いたい利用者は，この機能を利用してエクセル上で作業することも可能である．

### 3.4 コーパス管理機能

形態素タグ付きコーパスは，辞書と同期して管理される．辞書は事前に用意して指定してもよいが，辞書を指定しない場合には，形態素タグ付きコーパスに出現した形態素集合が辞書として自動的に作られる．形態素タグ付きコーパスは，システム内では辞書項目へのポインタ列として格納される．

誤り訂正機能については，現在は形態素タグの修正のみ実装されている．検索結果の全体，あるいは，マウスで選択した文の集合に対して，TagEdit という機能を利用すると，指定された文集合中の共通部分が表示され，その形態素列を任意の形態素列に変更することができる．ただし，表層文字列を変更することは

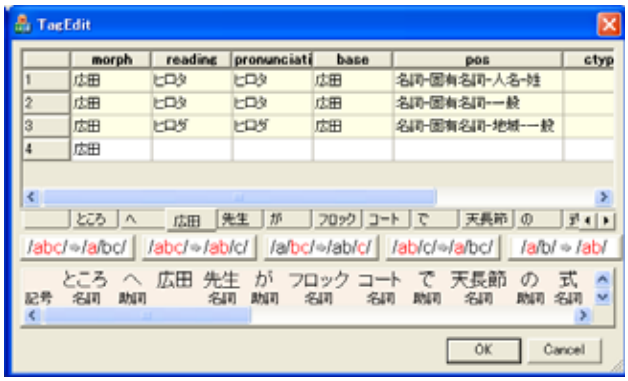


図 2: 形態素誤りの修正画面

許されない．変更できるのは，分かち書きの修正，および，形態素情報の修正である．形態素情報は，直接書き換えるのではなく，辞書に含まれる形態素を選択する形で行われるので，辞書との連携が維持される．なお，このインタフェースを通して新たな形態素を定義するこも可能であり，その場合には，辞書に新しい形態素が登録され，それへのポインタが作られる．図 2に形態素誤り修正インタフェースを示す．この例は,「広田」という語に対して，正しい品詞情報をタグつけようとしているところであり，辞書に登録されている 3 つの候補と新らしい形態素を定義するための 4 つ目の候補が示されている．利用者は，正しい形態素を選択するか，新しい形態素情報を入力することで，コーパスの修正を行うことができる．

現システムの実装は，MySQL(version 4.1) を用いてタグ付きコーパスと辞書を格納し，ユーザインタフェース部は VisualC++を用い，MySQL への検索要求の生成と結果の表示等の処理は Ruby を用いている．

## 4 今後の予定

現在のシステムは，無償での配布と実装の容易性を主として考え，汎用のデータベースを用いており，効率については多くの注意を払っていない．100 万語程度の形態素タグ付きコーパスについては，充分な速度で動いているが，係り受け構造や複雑な検索要求については，効率上の問題が残っている．

係り受け構造については，現在は基本的な検索機能をもつだけである．係り受け木の表示や係り受け解析誤りの修正機能については，来年度に実施する予定である．また，複合語とその構成要素，固有表現とその他のタグ情報のように，互いにオーバーラップする可能性のあるタグについては，これらを同時に表示するか，あるいは，画面を切り替えて表示する方法の検討を進めている段階である．

## 5 おわりに

現在開発中の，タグ付きコーパスの管理と検索を行う汎用のツール「茶器」の開発指針と現状について紹介した．本プロジェクトは平成 15 年度から 3 年計画で継続しており，来年度が最終年にあたる．現在の最新版は，年度内にフリーソフトウェアとして公開する予定であり，利用者からのフィードバックを得たいと考えている．来年度中は，新規機能が追加されるごとに，新しいバージョンとして適宜公開したいと考えている．

**謝辞:** 本システムの構築に協力いただいた研究室の学生および修了生に感謝する．なお，本研究は，文部科学省科学技術研究補助金 基盤研究 B「言語研究のためのコーパスの作成と利用に関する研究」(研究期間: 平成 15 年度〜17 年度, 課題番号: 15300046) の支援を受けて行ったものである．

## 参考文献

[1] 浅原正幸, 他, 「語長変換を考慮したコーパス管理システム」, 情報処理学会論文誌 Vol.43,No.7, pp.2091-2097, 2002.

[2] 浅原正幸, 高橋由梨加, 松本裕治, 「異表記同語情報を付与した辞書の整備」, 本論文集, 2005.

[3] ゴーチュイリン, 鄭育昌, 浅原正幸, 松本裕治, 「中国版茶筌の開発」, 本論文集, 2005.

[4] 工藤拓, 松本裕治, 「チャンキングの段階適用による日本語係り受け解析」, 情報処理学会論文誌 Vol.43,No.6, pp.1834-1842, 2002.

[5] 松本裕治, 「形態素解析システム『茶筌』」, 情報処理, Vol.41, No.11, pp.1208-1214, 2000.

[6] 松本裕治, 他,「タグ付きコーパスの格納/検索ツール：茶器」, 言語処理学会第 10 回年次大会発表論文集, pp.405-408, 2004.

[7] Tetsuji Nakagawa, Yuji Matsumoto, “Detecting Errors in Corpora Using Support Vector Machines,” COLING 2002, pp.709-715, 2002.