# タグ付きコーパス管理/検索システム「茶器」
# 使用説明書

version 2.1

August 24, 2007

松本裕治，浅原正幸（奈良先端科学技術大学院大学）
橋本喜代太（大阪府立大学），投野由紀夫（明海大学）
大谷朗（大阪学院大学），森田敏生（総和技研）

Annotated Corpus Management System ChaKi: User's Manual
Yuji Matsumoto, Masayuki Asahara, Kiyota Hashimoto, Yukio Tono, Akira Ohtani, Toshio Morita


version 0.5 March 17, 2004
version 0.51 April 23, 2004
version 0.6 August 23, 2005
version 0.62 November 10, 2005
version 0.63 January 17, 2006
version 0.64 February 11, 2006
version 1.0 March 31, 2006
version 2.0 April 2, 2007
version 2.1 August 24, 2007

# 目 次

# 1　はじめに

コーパスに基づく自然言語処理の進展に伴い，品詞情報や統語情報，さらに詳細なタグ付きコーパスの蓄積が進んでいる．また，種々の統計的言語解析システムの進歩により，テキストに対する自動タグ付与がかなりの精度で行えるようになってきた．

タグ付きコーパスは，言語学/言語処理研究の基本データとしてだけでなく，統計的機械学習に基づく言語処理の高性能化のためにも貴重な資源である．前者のためには，柔軟な検索機能および統計解析を行うための加工機能を持ったタグ付きコーパス支援システムの存在が重要である．

「茶器」は，タグ付きコーパスの検索および管理を支援する目的で作成されたツールである．文字列，単語列，および，係り受け関係による検索機能を備えている．単語列による検索では，単語の表層形以外に，読み，品詞や活用形などの文法情報を指定して検索を行うことができる．係り受け関係による検索では，文節内の単語列の指定と文節間の係り受け関係を指定した文検索が可能である．また，コーパス内の単語の頻度や前後文脈における単語の頻度など，簡単な統計処理を行うことができる．茶器は，タグ付きコーパスを関係データベースシステム（MySQL を使用）に格納し，検索要求を記述し結果を表示するためのインタフェースを提供する．対象言語は，多言語を目指しており，日本語，英語，中国語のデータを取り扱うことが可能である．

本使用説明書では，茶器の基本的な機能について解説する．

## 2　環境のインストール

本マニュアルでは，既にインストール済みの茶器の使用法について説明する．茶器を利用したコーパスの検索およびコーパス修正には次の作業が必要である：

- MySQL のインストール
- ChaKi GUI のインストール
- データ整形ツール (形態素解析ツール，係り受け解析ツールなど) のインストール
- データ整形ツールにより解析されたコーパスのシステムへの格納ツール

これらの詳細については，茶器の配布パッケージに含まれるインストーラを参照のこと．

# 3　検索に関する基本機能と実行例

茶器は，コーパスのデータベース化のための MySQL とデータベースへの問い合わせおよび結果の表示を行うインタフェース部分からなる．本説明書は主として後者のインタフェース部の説明書である．MySQL の実装と品詞タグ付きコーパスの MySQL への格納を事前に行う必要があるが，その詳細については，付録を参照のこと．

## 3.1　準備

### 3.1.1　コーパス定義ファイル

茶器をインストールしたフォルダ内にコーパス定義ファイルを作成する．例えば，sanshiro というコーパスが MySQL に既に格納されているとする．その場合，例えば，sanshiro.def というファイルを作成し，その中身を以下のようにする:

```
corpusname=sanshiro
server=localhost
user=root
password=okage
```

”corpusname” は格納されたコーパス名を指す．”user” および”password” には，MySQL の実装時に使用したユーザ名とパスワードを指定すること．

コーパス定義ファイルには，次のような項目を指定することができる．

**品詞定義リスト:** 検索時に品詞を指定する際，品詞名を直接指定するのではなく品詞一覧から選択したい場合，品詞の一覧を記述したファイルを事前に用意し，それを品詞定義ファイルの中で次のように指定すればよい．

```
poslist=RWCP.pos
```

品詞リスト（上の例では “RWCP.pos” という名前のファイルを仮定している）の記述法については，付録を参照のこと．

**単語情報の表示指定:** 単語および係り受け関係検索時には，単語のもつ属性情報として表層形以外の次のような項目を指定することができる．

- 日本語：表層形（morph），読み（reading），発音（pronunciation），基本形（base），品詞（pos），活用型（ctype），活用形（cform）
- 英語：表層形（morph），基本形（base），品詞（pos）
- 中国語：表層形（morph），品詞（pos）

茶器では，デフォルトとして，日本語コーパスが持つ７つの情報を単語が持っていると仮定しているが，対象コーパスが一部の情報を欠く場合もあるし，多言語のコーパスのようにその一部しか情報を持たない場合もある．検索時に単語の情報として指定する項目を限定したい場合は，コーパス定義ファイルの中に，利用者が表示したい項目だけ（例えば，表層形と品詞だけ）を次のように記述すればよい．

```
attrs=morph
attrs=pos
```

特に属性情報を明示的に記述しない場合は，７つの属性情報すべてを用いると仮定される．

### 3.1.2 茶器の実行とコーパス指定

茶器の実装を行ったフォルダ内の ”ChaKi.exe” を実行する．図 1 のようなシステムが現れる．

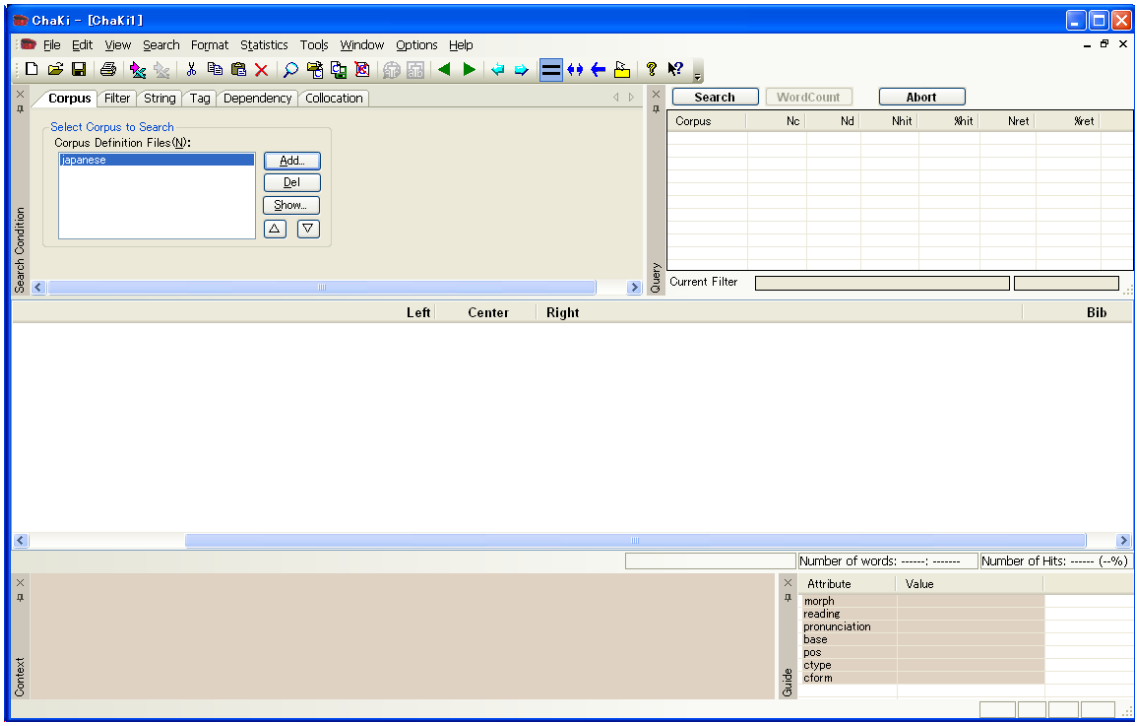



図 1: 茶器初期画面

システムは，いくつかの部分からなっている．最上部には，各種コマンドを指示するツールバーがあり，その下の本体が５つの部分よりなる．最初が，コーパスの指定部である．「Add..」ボタンを押すと，コーパス定義ファイルを聞いてくるので，事前に定義しておいたコーパス定義ファイル（上述の .def を拡張子にもつファイル）を選択する．複数のコーパスを指定することができる．追加したコーパスは「Del」ボタンによる削除することができる．「Show」ボタンは，コーパスの詳細情報を表示するために用いる．その下の上下の三角記号によるコーパスの順番を入れ替えることができる．コーパス指定部の上部には，コーパス検索全般に関するフィルター機能を指定するボタン (Filter)，および，種々の検索ボタン (String, Tag, Dependency, Word List, Collocation) がある．コーパス指定部の右側には，検索時に，コーパス毎の検索数等の統計情報が表示される．

コーパス指定部の下が，KWIC 表示部であり，検索結果を中心として前後の文字列（単語列）が表示される．一番下が文脈表示部である．検索結果の一つを指定して，コーパス中のその前後の文を表示することができる．文脈表示部の右側のウィンドウは，単語情報の表示部であり，マウスが置かれた位置の単語の属性情報が表示される．

上記のそれぞれのウィンドウは，独立したウィンドウとして切り離して表示することが可能であり，位置を自由に変更することが可能である．

## 3.2 検索結果出力に対するフィルター機能

次章以下で説明する検索は，コーパス指定部で指定されたすべてのコーパスに対して行われるが，図 2 に示すフィルター画面より，書誌情報に対する制約と，検索結果の表示数，および，検索対象の文の範囲などを指定することができる．（なお，本機能は，まだ未実装である）

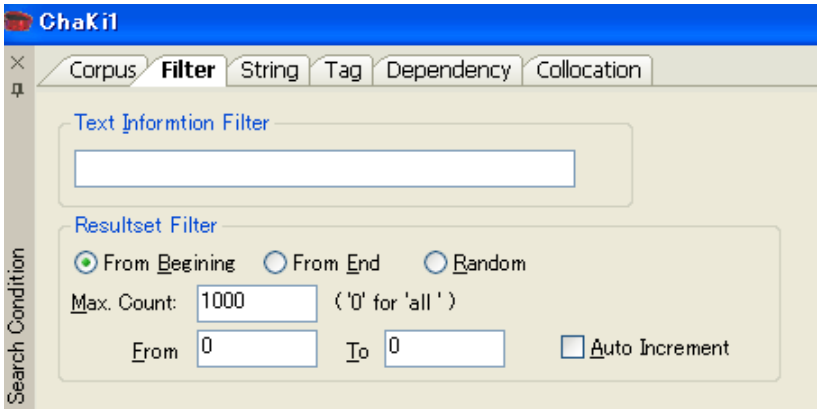



図 2: コーパスに対するフィルター指定画面

# 4 検索に関する基本機能と実行例

## 4.1 文字列検索機能 (String Search)

「String」ボタンを押すことにより，文字列検索モードになる.「Search Expression」に検索したい文字列を入力し，Query ボックス内の「Search」ボタンを押すことにより，検索が実行される．図 3 は,「三四郎」という文字列を入力して検索した結果を示している.

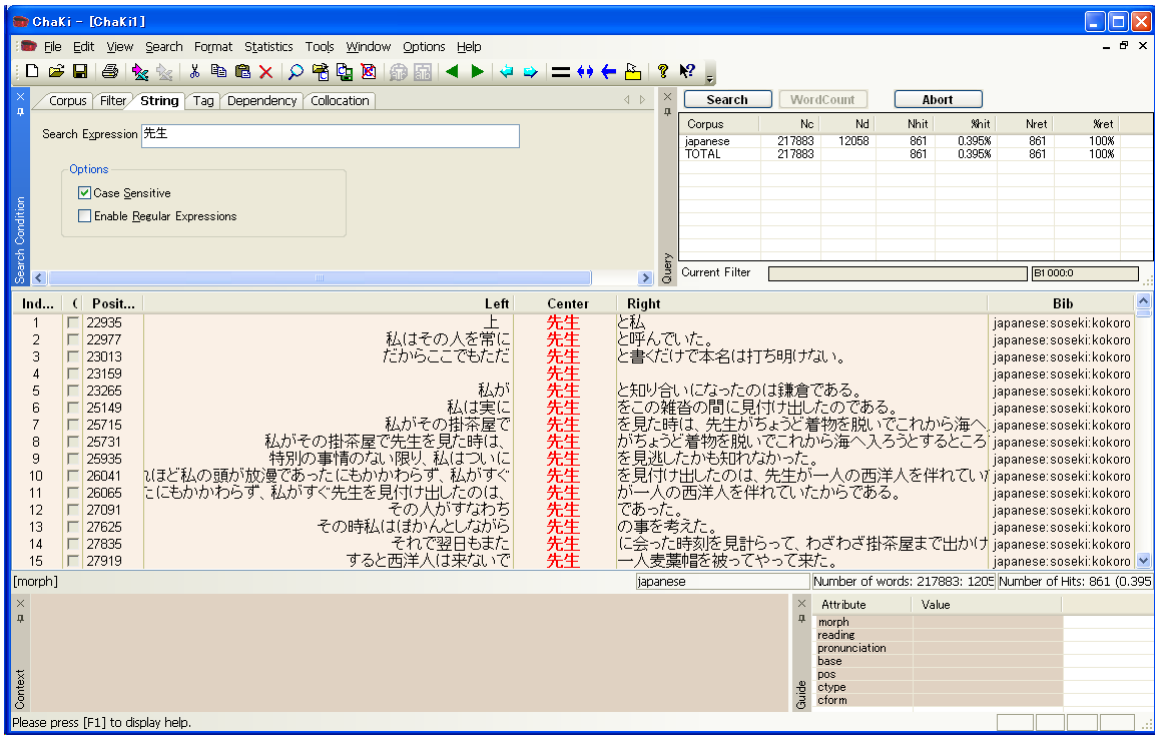



図 3: 文字列検索画面

図のように，KWIC 表示部に，検索文字列が赤色で示され，その前後の文字列が示される．その一つを選んでダブルクリックすると，その文の前後の文が文脈表示部に示される．図を見てわかるように，文脈表示部でも検索文字列と当該文が色づけされて表示される．文脈表示部では前後 2 文ずつを表示するが，そ

の数を変更することができる．ツールバーの「Options - Context Options」を開くと，前後文脈の文数を指定するウィンドウが現れるので，前文脈，後文脈として表示したい文数を指定すればよい．

文字列検索では，正規表現による文字列指定を行うこともできる．正規表現を用いる場合には，「Search Expression」の下に表示されている「Options」内の「Enable Regular Expressions」のボックスをクリックしておくこと．現在は，文字列の否定 (negation) はサポートしていない．また，英語の検索については，大文字と小文字を区別するかどうかを指定することができる．デフォルトでは，大文字と小文字の区別を行わないが，同じ「Options」内の「Case Intensive」ボタンをクリックしておけば，大文字と小文字を区別した検索を行うことができる．

## 4.2 単語列検索機能 (Tag Search)

品詞タグ付けされたコーパスに対し，様々な情報を指定して，コーパス内の表現を検索することができる．現在は，個々の単語（形態素）に次の情報（属性）が付与されていると仮定している．ただし，すべての情報が必要ということではなく，下記のものが付与されうる最大の情報である．

「出現形 (morph)」,「読み (reading)」,「発音 (pronunciation)」,「原形 (base)」,「品詞 (pos)」,「活用型 (ctype)」,「活用形 (cform)」

これらの情報は，任意の文字列，および，正規表現によって指定可能である．

品詞は階層構造を持つことができ，階層の区切りはハイフンによって示される．例えば,「名詞-固有名詞-組織」は，３階層からなる品詞名である．品詞の一覧は適当な名称のテキストファイル，例えば POSList.txt というファイルに記述する．茶器のデフォルトでは，茶筌で用いているＩＰＡ品詞体系の品詞一覧が格納されている．任意の単語は，上記属性の組として指定される．詳細については，付録Ａを参照のこと．

### 4.2.1 単語列検索の検索要求の記述

「Tag」のボタンを押すと，コーパス指定部の位置に，次のような単語列検索命令のための単語ボックスが表示される．

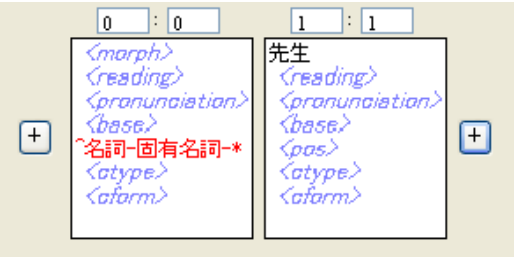



図 4: 単語列検索の質問例

属性名の部分をクリックすると入力用の箱が表示される．正規表現入力したい場合は,「R」というボタンをクリックする（正規表現は，赤色文字で入力される）．品詞定義ファイルが指定されている場合，(pos) 属性を選ぶと，階層構造に沿った品詞名入力を行うことができる．

単語ボックスは，左右の「＋」が書かれたボタンをクリックすることで任意の個数作ることができる．上の図３の例では,「先生」という見出しの直前に「名詞-固有名詞-*」を品詞とする単語が現れるパターンを検索しようとしている．単語ボックスの上の数字が入った２つの小さい箱は，単語の出現位置を表しており，0：0 が中心語を表し，他は中心語からの相対位置を表す．２つの箱によって出現位置に幅を持たせることができる．上の例では,「名詞-固有名詞-*」の位置が「0：0」であり,「先生」の位置情報 1：:1 が中心語の直後の位置のみを指定するが，これを例えば「1：3」とすれば，直後から３単語後の位置までに「先生」

が出現するすべてのパターンが検索対象となる．中心語に幅を持たせることも可能ではあるが，KWIC 表示が複雑になるので，中心語は「0：0」と指定することが好ましい．

### 4.2.2 単語列検索の結果表示

図 5 が，検索結果の例である．KWIC 表示窓には中心語と前後の単語が表示される．

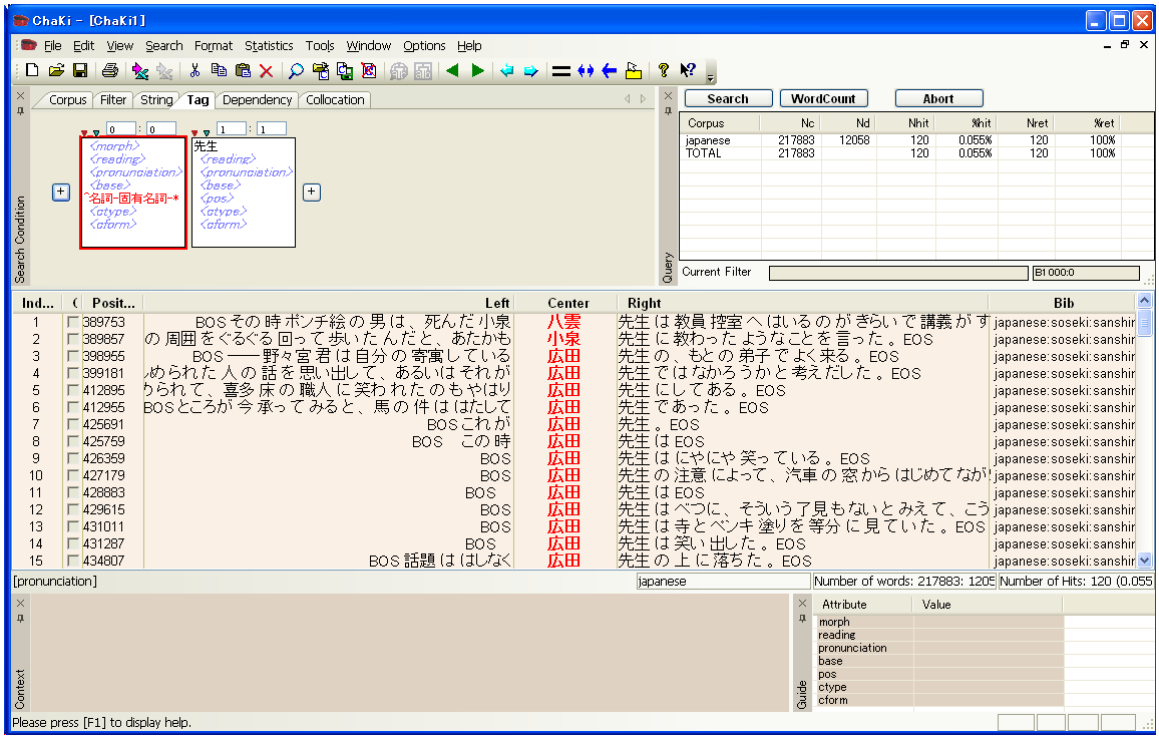



図 5: 単語列検索の結果画面

KWIC表示部では，各単語に１つあるいは２つの属性を表示することが可能である．２つの属性を表示するには，ツールバーから「View - View Attribute」を順に選択すればよい，表示すべき属性の選択は，ツールバーから「Options - KWIC Display Row Setting…」から指定することができる．

品詞の表示については，すべての階層情報を表示することは煩雑であることが多いので，最上位階層の品詞名のみの表示に限定することが可能である．上記の「KWIC Display Rows Setting」のウィンドウ内にそれを指定するチェックボックスがあるので，好みに応じて指定すればよい．単語について完全な情報を見たい場合には，ツールバーから「View - Popup 」Attributes」を選んでおけば，KWIC 表示窓の単語にマウスを持っていくと，全属性情報を含むポップアップウィンドウが表示されるようになる．

## 4.3 係り受け関係検索 (Dependency Search)

「Dependency」を選択することにより，文節係り受け解析済みコーパスの検索を行うことができる（日本語の場合は，南瓜による係り受け解析結果を，英語の場合は，単語係り受け解析結果をデータベースに格納しておくことが必要）．

### 4.3.1 係り受け検索の検索要求の記述

図 6 が係り受け解析の検索質問の例を示している．色のついたそれぞれのボックスが一つの文節を表し，その中に単語情報のボックスが含まれている．

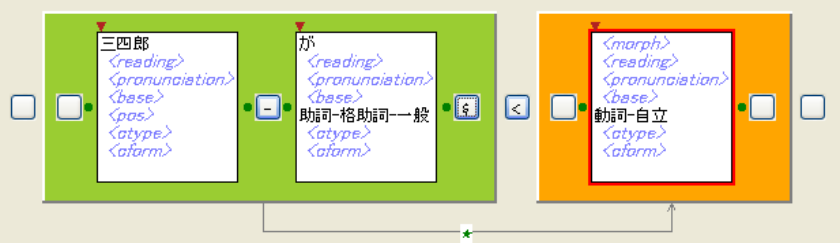



図 6: 係り受け検索質問の例

この例は,「三四郎　が」を含む文節が「動詞-自立」という品詞を持つ語を含む文節に係っているパターンを表している．各単語ボックスの左上部に小さな赤い逆三角がある．検索結果のＫＷＩＣ表示において中心位置に置きたい語を，この逆三角を選ぶことによって決める．図 6 では,「動詞-自立」が選択されている．

単語の両側にある小さなボックスは，文節内での相対位置を表す．文節ボックスの両側にある小さなボックスも同様である．この小さなボックスをマウスでクリックすると以下の記号を選択することができる．各記号の意味は次の通りである．

- 記号なし：　特に制約を指定しない
- ＋：　新しい単語（文節）ボックスを追加する
- －：　両側の単語が直接隣接関係にある（単語（文節）ボックス間でのみ選択可）
- ＜：　両側の単語の相対位置のみ指定，間に他の単語が存在してもよい
- ^:　文節（あるいは文）内の先頭の位置（文節，文の最も左のボックスにのみ指定可）
- $：　文節（あるいは文）内の末尾の位置（文節，文の最も右のボックスにのみ指定可）

上の例では,「三四郎」と「が」が文節内で直接隣同士の位置にあることを指定している．

文節間の係り受け関係を指定するには，一つの文節内でマウスを左ボタンをクリックし，そのまま係り先の文節までマウスを移動してボタンをはなせば，両文節を結ぶ矢印が挿入される．係り受け関係を解除するには，矢印上のマウスを置いて右ボタンをクリックすると「Delete this arrow」と表示されるので，それを選択すればよい．

### 4.3.2 係り受け検索の結果表示

係り受け検索の結果は，単語列検索と同様に KWIC 表示されるが，文節のまとまりが下線で示される．検索で指定した文節については，検索質問の文節ボックスと同じ色の下線が示される (図 7)．

KWIC 画面で文を選択して,「Tools」オプション，あるいは，マウスの右クリックで「TreeEdit」を選ぶと，係り受け解析木が別画面（図 8）が表示される．後述するように，この画面は，係り受け解析の誤り修正を行うのにも利用される．

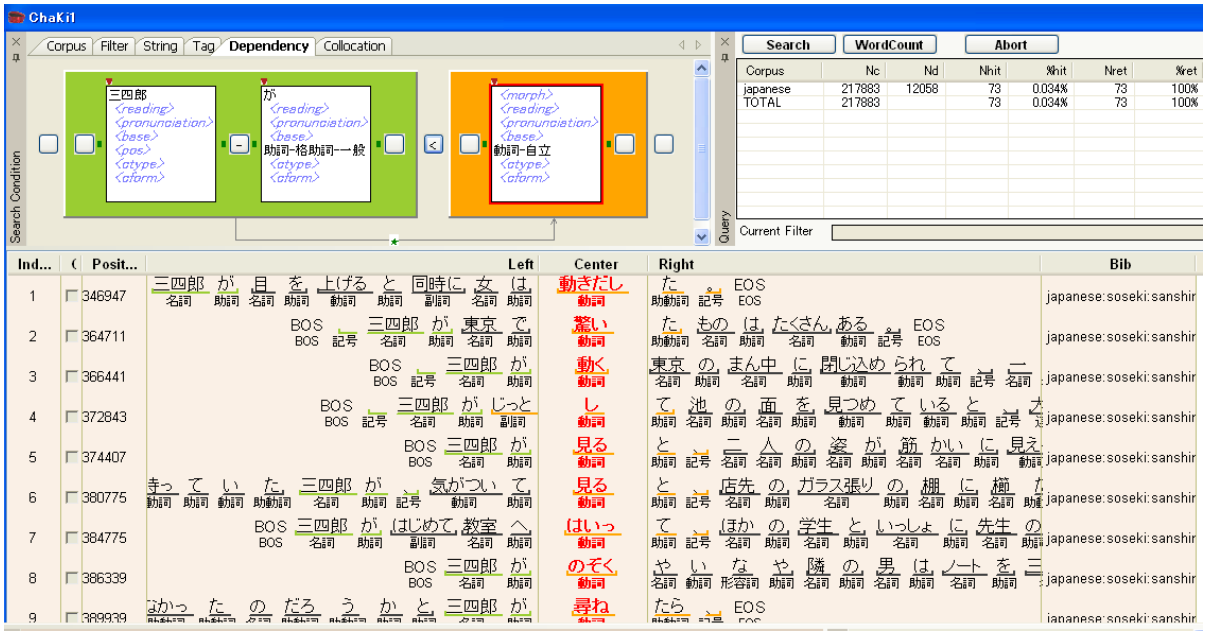



図 7: 係り受け関係検索の KWIC 表示

## 4.4 単語統計 (Word Count)

前節の単語列検索や係り受け検索では，ヒットしたすべての検索結果がデータベースから読み出されて KWIC 表示窓に出力されるため，ヒット数が多い場合，データの転送に時間を要する場合がある．単語列検索および係り受け関係検索において，Query ボックス内の「Search」ボタンの代わりに「WordCount」ボタンを押すことによって，中心語の簡単な単語統計を取ることができ，検索ヒットする数や，どのような単語が検索されるかの一覧を見ることが簡単にできる．検索要求の記述は，単語列検索や係り受け関係検索とまったく同じである．

図 9 は，図 6 の検索質問と同じものを，このモードで実行した結果表示されるウィンドウを示している．

このように,「三四郎 が」が直接係る自立動詞は 73 個あり,「あげる」の連用形が 1 回,「する」の連用形が 5 回現れたことなどがわかる．morph から cform までの情報は，考慮しないようにすることも可能である．表中のそれぞれの属性の前にプラス (＋) の記号があるが，それをマウスでクリックすることにより，属性を考慮しないようにできる．例えば，base と ctype 以外を off にし (つまり活用変化を無視したい場合)，頻度によってソートした結果を図 10 に示す．

各行が文に含まれていたかを知るためには，その行の左端のマスでマウスの右ボタンを押し (図 11),「List Occurences」を選択すれば，対応する行の具体例の検索が行われる．Word List の表は，茶器の「File」メニュー内の「Send to Excel」を選択することにより，Excel 可読の csv ファイルとして出力することが可能である．

また，この表の各列は，各属性名の部分をダブルクリックすることによってソートすることができる．また，属性名の部分を右クリックして「Compact Row」を選択すると，その列の情報は無視される．例えば，品詞だけで統計を取りたい場合は，品詞以外の属性を無視するようにすればよい．また，動詞等の活用語の場合は，morph, reading, cforms など出現形によって異なる値をもつ属性を無視することにより，それぞれの単語の出現頻度を正しく知ることができる．

各行の左端の数字部分を右クリックすると「List Occurrences」というポップアップが表示されるので，それを選択すると，その行に対応するすべての文の KWIC 表示が別ウィンドウに得られる．

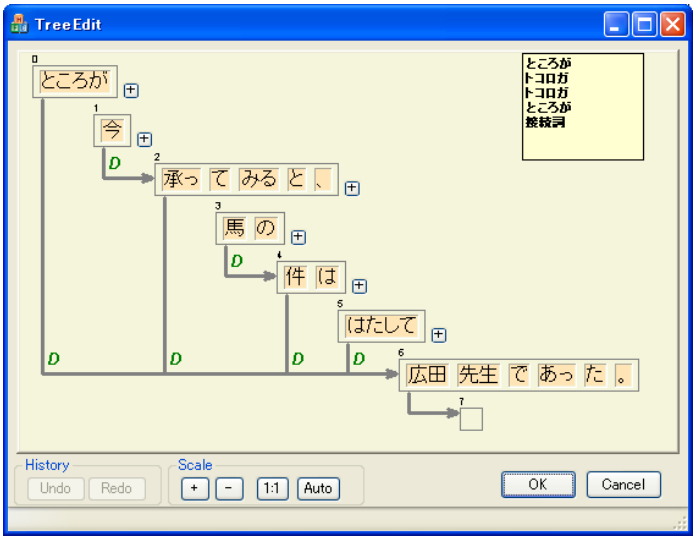



図 8: 係り受け木の表示

## 4.5 共起情報検索 (Collocation)

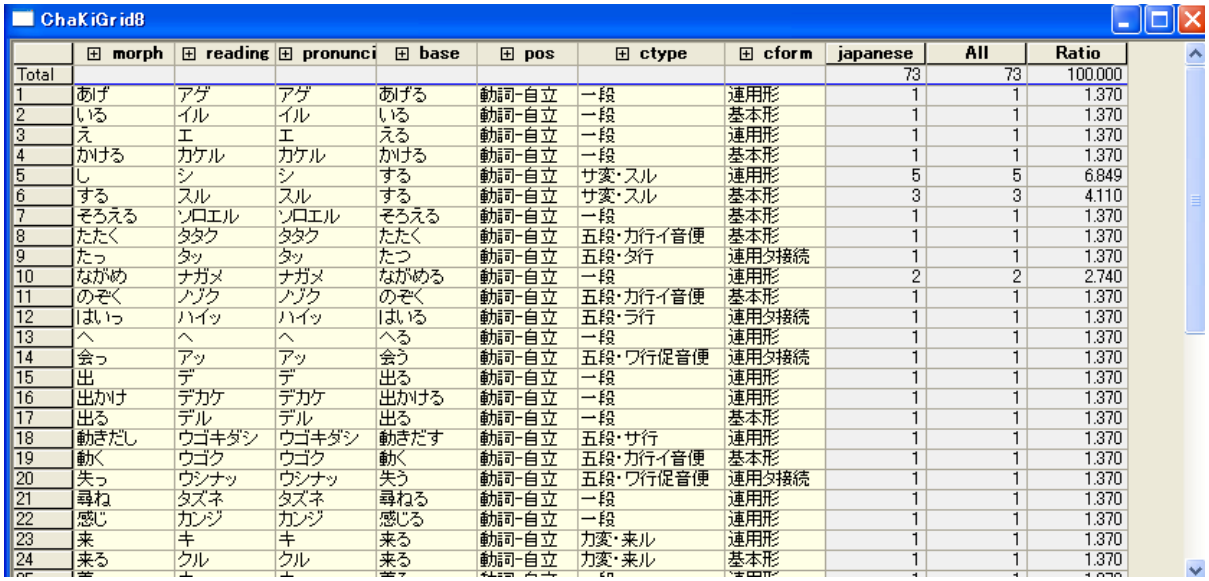
中心語の前後にどのような単語が出現するかをまとめて集計するのがこの機能である．単語検索を実行し，検索結果が得られた後，この機能を用いることができる．前述の緑色の矢印によって中心語を移動してもよい．ここでの集計は，KWIC 窓に表示されているデータに対して集計が行われる．

Collocation ボタンを押すと，この機能で指定可能なオプションが図 12 のように表示される．Type of Statistics については，Raw frequency, MI score, Bigram (right), Bigram(left), (Continuous) Frequent Sequence Mining のオプションがある．Target Attributes は，統計を取るべき要素の単位を表しており，複数の属性が選択可能である．Window Size は，中心語の左右何語までの共起の頻度を取るかを指定する．図の例では，語の出現形 (morph) について，左右 4 単語ずつの統計を取ろうとしている．検索条件をこのように指定した後で，右の Query ボックスの上にある「Collocation」ボタンを押すと，図 13 のようなウィンドウが表示される．一番左の列が，中心語と共起する単語の一覧であり，それぞれがどの位置に何回出現したかが，この表により表されている．この表も Excel 可読の csv ファイルとして出力することができる．また，各列は，列の最上位置をマウスでダブルクリックすることにより，頻度順に並べ替えることが可能である．

図 14 は，相互情報量などの共起尺度の表示を指定する様子を示している．図では，Type of Statistics として「MI Score」を選び，統計を取る単位の指定 (Target Attribute) として base および pos を選んでいる．つまり，基本形と品詞が異なる語のみ異なる要素として統計を取るように指定している．また，Window size として左右とも 2 としており，現在の語の左右 2 語以内に現れる語を対象に統計情報を取るよう指定している．共起情報の計算の実行は，上と同様，Query ボックス内の「Collocation」ボタンを押すことで行われる．結果の例を図 15 に示す．

Type of Statistics の他のオプションである Bigram (right), Bigram(left) は，現在の中心語から右あるいは左へ連接する単語の N-gram 統計を取る．このオプションでは，Minimum Frequency と Minimum Length の 2 つの条件が指定できる．検索したい N-gram に対して，前者は出現頻度の下限，後者は長さの下限を指定しており，これらによって指定された頻度や長さを下回る N-gram は検索対象にならない．これまでと同様，Query ボックスの「Collocation」ボタンを押すことにより，条件を満たす N-gram の一覧と出現頻度が別ウィンドウに表示される．

Type of Statistics の新しいオプションとして，Continuous Frequent Sequence Mining と Frequent Sequence Mining の機能がある．前者は，検索された文集合の中心語に限定せず，すべての連結系列 (N-gram) を検索するオプションである．一方，Frequent Sequence Mining は，すべての (必ずしも連結して現れ



ChaKiGrid8

| | morph | reading | pronunci | base | pos | ctype | cform | japanese | All | Ratio |
|---|---|---|---|---|---|---|---|---|---|---|
| Total | | | | | | | | 73 | 73 | 100.000 |
| 1 | あげ | アゲ | アゲ | あげる | 動詞-自立 | 一段 | 連用形 | 1 | 1 | 1.370 |
| 2 | いる | イル | イル | いる | 動詞-自立 | 一段 | 基本形 | 1 | 1 | 1.370 |
| 3 | え | エ | エ | える | 動詞-自立 | 一段 | 連用形 | 1 | 1 | 1.370 |
| 4 | かける | カケル | カケル | かける | 動詞-自立 | 一段 | 基本形 | 1 | 1 | 1.370 |
| 5 | し | シ | シ | する | 動詞-自立 | サ変・スル | 連用形 | 5 | 5 | 6.849 |
| 6 | する | スル | スル | する | 動詞-自立 | サ変・スル | 基本形 | 3 | 3 | 4.110 |
| 7 | そろえる | ソロエル | ソロエル | そろえる | 動詞-自立 | 一段 | 基本形 | 1 | 1 | 1.370 |
| 8 | たたく | タタク | タタク | たたく | 動詞-自立 | 五段・カ行イ音便 | 基本形 | 1 | 1 | 1.370 |
| 9 | たっ | タッ | タッ | たつ | 動詞-自立 | 五段・タ行 | 連用タ接続 | 1 | 1 | 1.370 |
| 10 | ながめ | ナガメ | ナガメ | ながめる | 動詞-自立 | 一段 | 連用形 | 2 | 2 | 2.740 |
| 11 | のぞく | ノゾク | ノゾク | のぞく | 動詞-自立 | 五段・カ行イ音便 | 基本形 | 1 | 1 | 1.370 |
| 12 | はいっ | ハイッ | ハイッ | はいる | 動詞-自立 | 五段・ラ行 | 連用タ接続 | 1 | 1 | 1.370 |
| 13 | へ | ヘ | ヘ | へる | 動詞-自立 | 一段 | 連用形 | 1 | 1 | 1.370 |
| 14 | 会っ | アッ | アッ | 会う | 動詞-自立 | 五段・ワ行促音便 | 連用タ接続 | 1 | 1 | 1.370 |
| 15 | 出 | デ | デ | 出る | 動詞-自立 | 一段 | 連用形 | 1 | 1 | 1.370 |
| 16 | 出かけ | デカケ | デカケ | 出かける | 動詞-自立 | 一段 | 連用形 | 1 | 1 | 1.370 |
| 17 | 出る | デル | デル | 出る | 動詞-自立 | 一段 | 基本形 | 1 | 1 | 1.370 |
| 18 | 動きだし | ウゴキダシ | ウゴキダシ | 動きだす | 動詞-自立 | 五段・サ行 | 連用形 | 1 | 1 | 1.370 |
| 19 | 動く | ウゴク | ウゴク | 動く | 動詞-自立 | 五段・カ行イ音便 | 基本形 | 1 | 1 | 1.370 |
| 20 | 失っ | ウシナッ | ウシナッ | 失う | 動詞-自立 | 五段・ワ行促音便 | 連用タ接続 | 1 | 1 | 1.370 |
| 21 | 尋ね | タズネ | タズネ | 尋ねる | 動詞-自立 | 一段 | 連用形 | 1 | 1 | 1.370 |
| 22 | 感じ | カンジ | カンジ | 感じる | 動詞-自立 | 一段 | 連用形 | 1 | 1 | 1.370 |
| 23 | 来 | キ | キ | 来る | 動詞-自立 | カ変・来ル | 連用形 | 1 | 1 | 1.370 |
| 24 | 来る | クル | クル | 来る | 動詞-自立 | カ変・来ル | 基本形 | 1 | 1 | 1.370 |

図 9: Word Count の検索結果表示画面

るとは限らない，すなわち非連結) 系列を検索するオプションである．このような系列は，たとえ Minimum Frequency や Minimum Length のような条件を指定しても，一般に非常に多数存在する可能性がある．非常に多くの系列が表示されるのを防ぐため，前者の連結系列検索では，Minimum Frequency と Minimum Length を満たし，かつ，最大長の系列のみを表示するようになっている．また，後者の非連結系列検索では，右方向について最大長の系列のみを表示するようになっている．

図 16 に，連接系列検索の質問画面の例を示す．ここでは，出現形 (morph) を対象に，出現頻度の下限を 3，長さの下限を 2 と指定している．また，Stopwords のボックスに “EOS BOS 。 、” の 4 つが記入されているが，これは検索の対象としない語を指定するのに用いられる．なお，毎回 Stopwords を指定するのが面倒な場合は，コーパス定義ファイル（コーパス語とに用意される “.def” を拡張子にもつファイル）の中に例えば次の 1 行を記入しておけばよい．(各語は半角の空白で区切られていることに注意)

```
stopwords=EOS BOS 。 、
```


図 10: base と ctype だけを指定した Word Count の検索結果表示画面


図 11: 一つの行に対応する事例の再検索要求

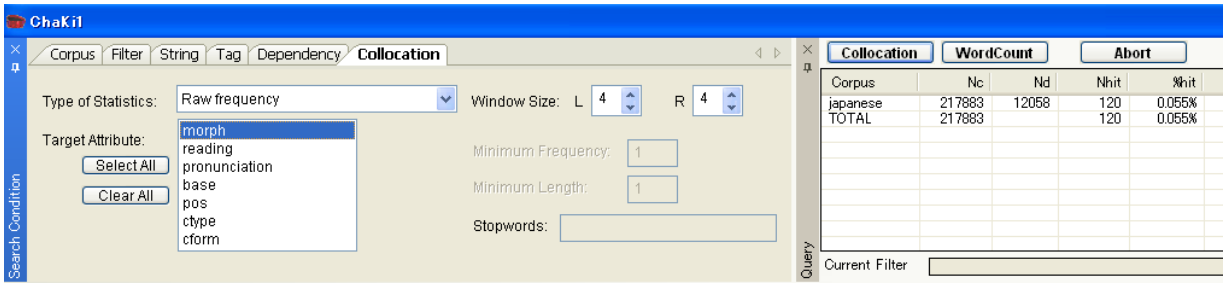
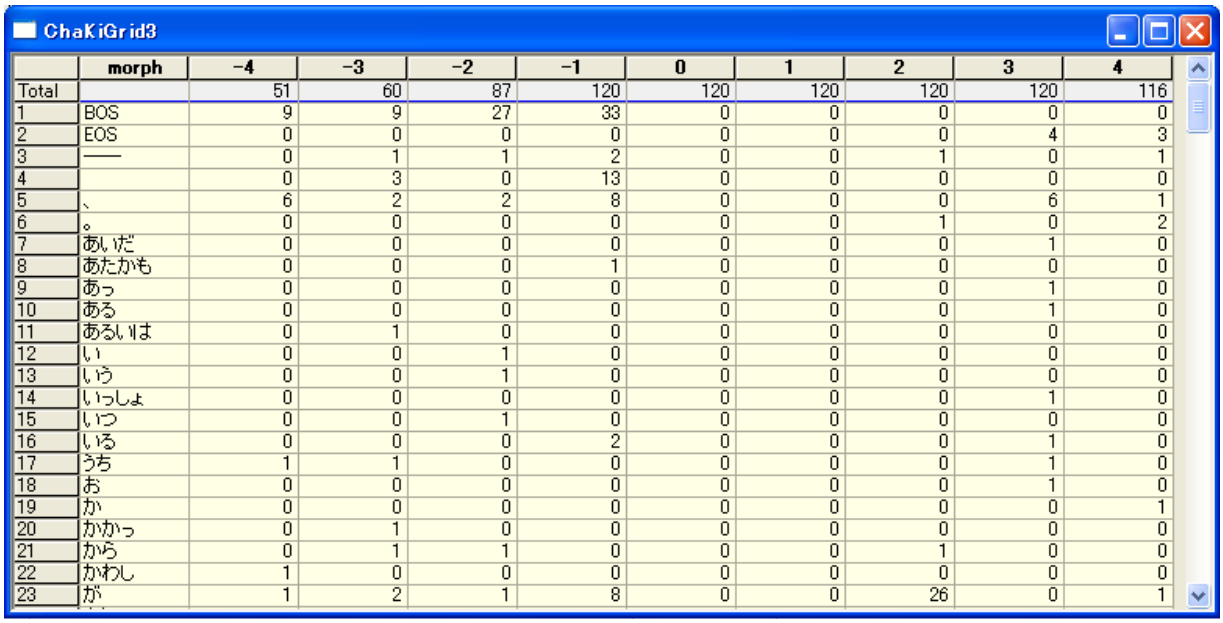



図 12: 共起頻度検索の質問指定画面



ChaKiGrid3

| | morph | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|
| Total | | 51 | 60 | 87 | 120 | 120 | 120 | 120 | 120 | 116 |
| 1 | BOS | 9 | 9 | 27 | 33 | 0 | 0 | 0 | 0 | 0 |
| 2 | EOS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 3 |
| 3 | —— | 0 | 1 | 1 | 2 | 0 | 0 | 1 | 0 | 1 |
| 4 | | 0 | 3 | 0 | 13 | 0 | 0 | 0 | 0 | 0 |
| 5 | 、 | 6 | 2 | 2 | 8 | 0 | 0 | 0 | 6 | 1 |
| 6 | 。 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 |
| 7 | あいだ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 8 | あたかも | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 9 | あっ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 10 | ある | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 11 | あるいは | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | い | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | いう | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | いっしょ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 15 | いつ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | いる | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 |
| 17 | うち | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 18 | お | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 19 | か | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 20 | かかっ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | から | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 22 | かわし | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23 | が | 1 | 2 | 1 | 8 | 0 | 0 | 26 | 0 | 1 |

図 13: 共起頻度検索結果

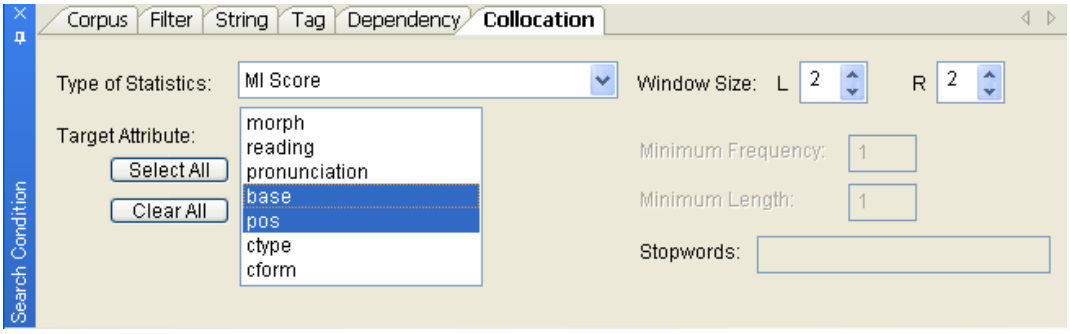
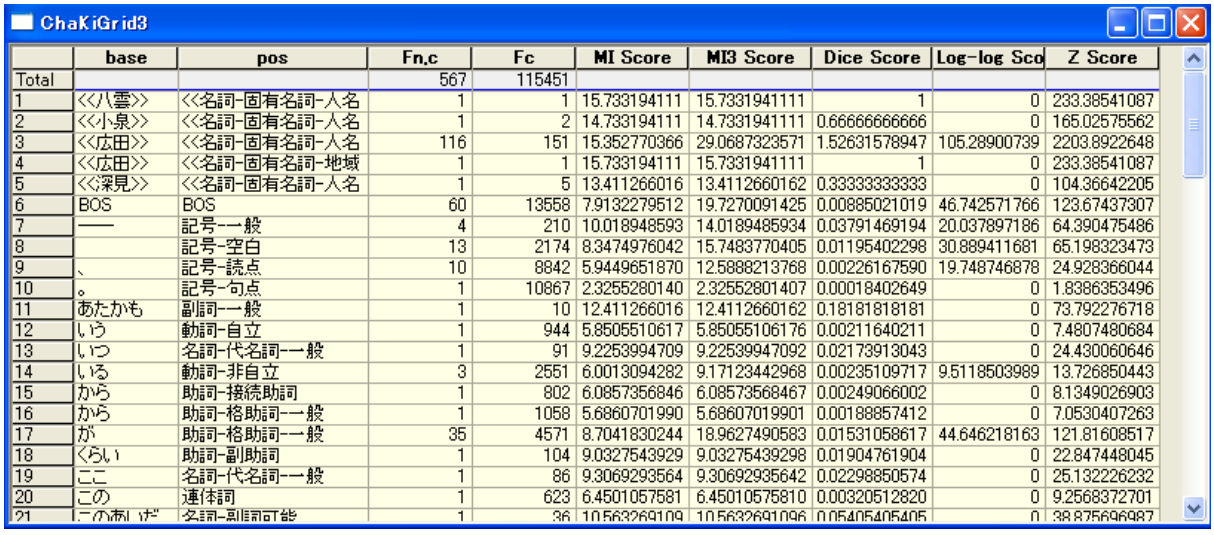



図 14: 共起情報検索の質問指定画面



ChaKiGrid3

| | base | pos | Fn.c | Fc | MI Score | MI3 Score | Dice Score | Log-log Sco | Z Score |
|---|---|---|---|---|---|---|---|---|---|
| Total | | | 567 | 115451 | | | | | |
| 1 | <<八雲>> | <<名詞-固有名詞-人名 | 1 | 1 | 15.733194111 | 15.7331941111 | 1 | 0 | 233.38541087 |
| 2 | <<小泉>> | <<名詞-固有名詞-人名 | 1 | 2 | 14.733194111 | 14.7331941111 | 0.66666666666 | 0 | 165.02575562 |
| 3 | <<広田>> | <<名詞-固有名詞-人名 | 116 | 151 | 15.352770366 | 29.0687323571 | 1.52631578947 | 105.28900739 | 2203.8922648 |
| 4 | <<広田>> | <<名詞-固有名詞-地域 | 1 | 1 | 15.733194111 | 15.7331941111 | 1 | 0 | 233.38541087 |
| 5 | <<深見>> | <<名詞-固有名詞-人名 | 1 | 5 | 13.411266016 | 13.4112660162 | 0.33333333333 | 0 | 104.36642205 |
| 6 | BOS | BOS | 60 | 13558 | 7.9132279512 | 19.7270091425 | 0.00885021019 | 46.742571766 | 123.67437307 |
| 7 | —— | 記号-一般 | 4 | 210 | 10.018948593 | 14.0189485934 | 0.03791469194 | 20.037897186 | 64.390475486 |
| 8 | | 記号-空白 | 13 | 2174 | 8.3474976042 | 15.7483770405 | 0.01195402298 | 30.889411681 | 65.198323473 |
| 9 | 、 | 記号-読点 | 10 | 8842 | 5.9449651870 | 12.5888213768 | 0.00226167590 | 19.748746878 | 24.928366044 |
| 10 | 。 | 記号-句点 | 1 | 10867 | 2.3255280140 | 2.32552801407 | 0.00018402649 | 0 | 1.8386353496 |
| 11 | あたかも | 副詞-一般 | 1 | 10 | 12.411266016 | 12.4112660162 | 0.18181818181 | 0 | 73.792276718 |
| 12 | いう | 動詞-自立 | 1 | 944 | 5.8505510617 | 5.85055106176 | 0.00211640211 | 0 | 7.4807480684 |
| 13 | いつ | 名詞-代名詞-一般 | 1 | 91 | 9.2253994709 | 9.22539947092 | 0.02173913043 | 0 | 24.430060646 |
| 14 | いる | 動詞-非自立 | 3 | 2551 | 6.0013094282 | 9.17123442968 | 0.00235109717 | 9.5118503989 | 13.726850443 |
| 15 | から | 助詞-接続助詞 | 1 | 802 | 6.0857356846 | 6.08573568467 | 0.00249066002 | 0 | 8.1349026903 |
| 16 | から | 助詞-格助詞-一般 | 1 | 1058 | 5.6860701990 | 5.68607019901 | 0.00188857412 | 0 | 7.0530407263 |
| 17 | が | 助詞-格助詞-一般 | 35 | 4571 | 8.7041830244 | 18.9627490583 | 0.01531058617 | 44.646218163 | 121.81608517 |
| 18 | くらい | 助詞-副助詞 | 1 | 104 | 9.0327543929 | 9.03275439298 | 0.01904761904 | 0 | 22.847448045 |
| 19 | ここ | 名詞-代名詞-一般 | 1 | 86 | 9.3069293564 | 9.30692935642 | 0.02298850574 | 0 | 25.132226232 |
| 20 | この | 連体詞 | 1 | 623 | 6.4501057581 | 6.45010575810 | 0.00320512820 | 0 | 9.2568372701 |
| 21 | このあいだ | 名詞-副詞可能 | 1 | 36 | 10.563269109 | 10.5632691096 | 0.05405405405 | 0 | 38.875696987 |

図 15: 共起情報検索の質問指定画面

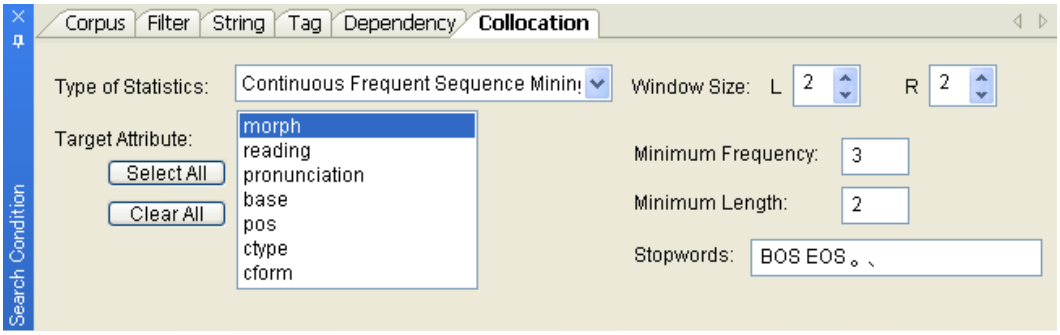
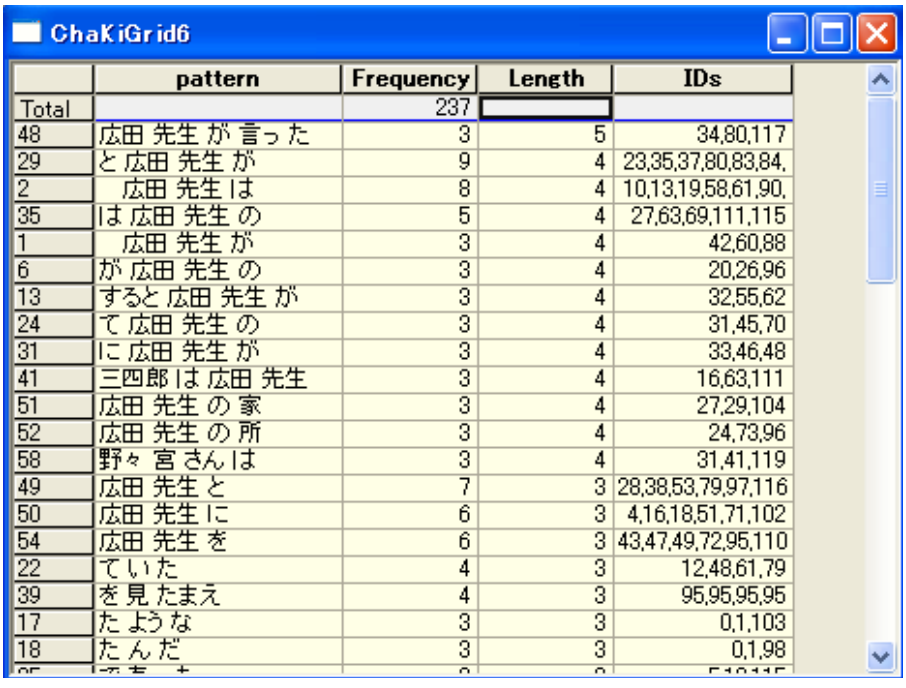



図 16: 連接系列検索の質問指定画面




| | pattern | Frequency | Length | IDs |
|---|---|---|---|---|
| Total | | 237 | | |
| 48 | 広田 先生 が 言っ た | 3 | 5 | 34,80,117 |
| 29 | と 広田 先生 が | 9 | 4 | 23,35,37,80,83,84, |
| 2 | 広田 先生 は | 8 | 4 | 10,13,19,58,61,90, |
| 35 | は 広田 先生 の | 5 | 4 | 27,63,69,111,115 |
| 1 | 広田 先生 が | 3 | 4 | 42,60,88 |
| 6 | が 広田 先生 の | 3 | 4 | 20,26,96 |
| 13 | すると 広田 先生 が | 3 | 4 | 32,55,62 |
| 24 | て 広田 先生 の | 3 | 4 | 31,45,70 |
| 31 | に 広田 先生 が | 3 | 4 | 33,46,48 |
| 41 | 三四郎 は 広田 先生 | 3 | 4 | 16,63,111 |
| 51 | 広田 先生 の 家 | 3 | 4 | 27,29,104 |
| 52 | 広田 先生 の 所 | 3 | 4 | 24,73,96 |
| 58 | 野々 宮 さん は | 3 | 4 | 31,41,119 |
| 49 | 広田 先生 と | 7 | 3 | 28,38,53,79,97,116 |
| 50 | 広田 先生 に | 6 | 3 | 4,16,18,51,71,102 |
| 54 | 広田 先生 を | 6 | 3 | 43,47,49,72,95,110 |
| 22 | て い た | 4 | 3 | 12,48,61,79 |
| 39 | を 見 たまえ | 4 | 3 | 95,95,95,95 |
| 17 | た よう な | 3 | 3 | 0,1,103 |
| 18 | た ん だ | 3 | 3 | 0,1,98 |

図 17: 連接系列検索の質問指定画面

## 5　タグ付け誤りの修正機能

茶器は，コーパス中のタグ付け誤りを修正する機能を持つ．形態素誤りと係り受け誤りは別のインタフェースを利用して修正する．誤りを含む文を選択し，「Tools」オプション，あるいは，修正を行いたい文にマウスを合わせて右ボタンを押し（図 18），「TagEdit」あるいは「TreeEdit」を選択する．以下，それぞれについて説明する．

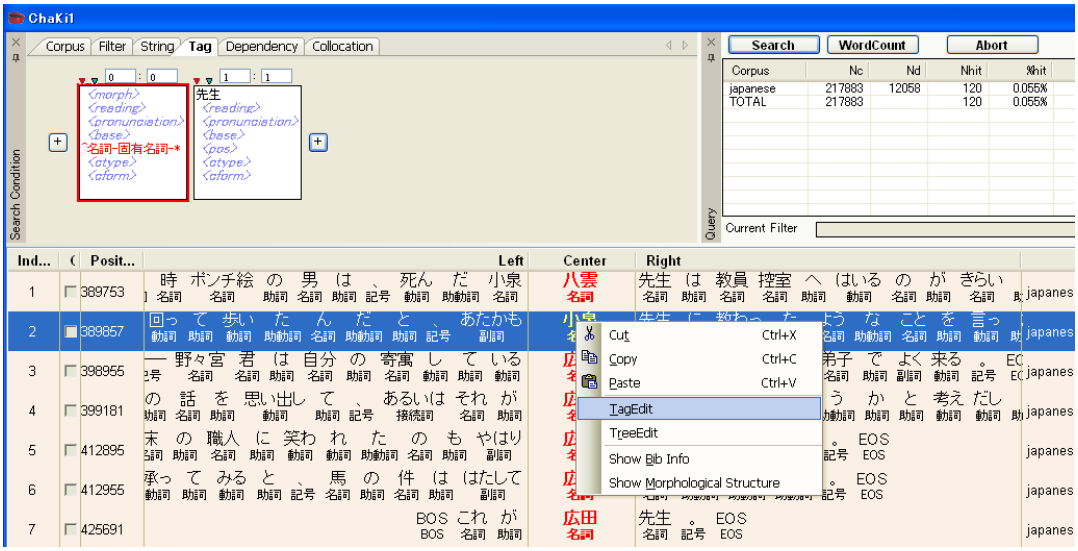



図 18: タグ付け誤り修正機能の起動画面

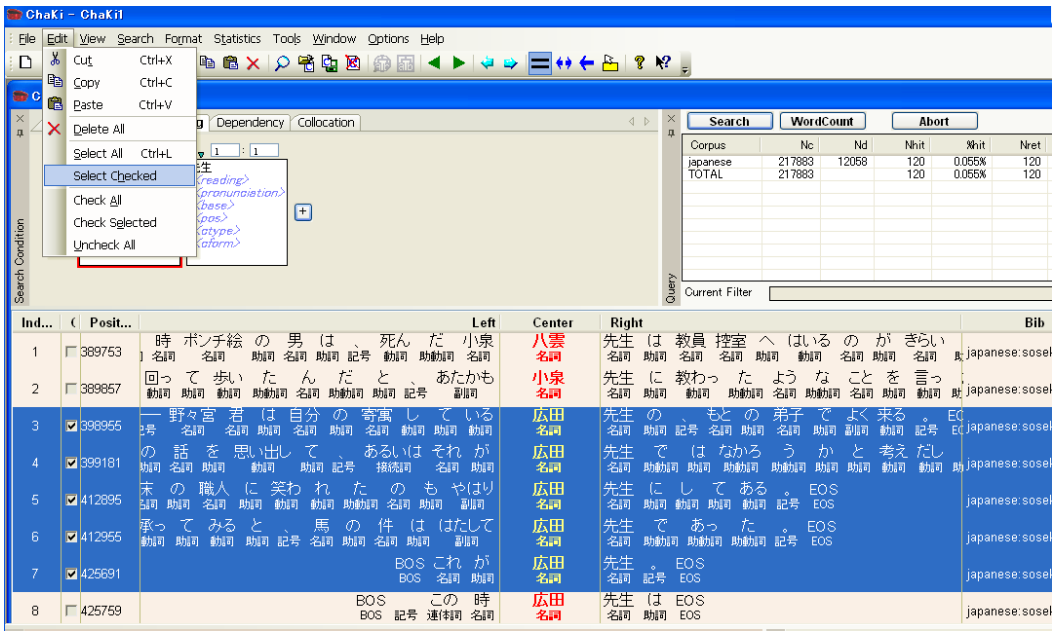



図 19: 誤り修正を行う文の選択

## 5.1 形態素誤りの修正

形態素に関するタグ付け誤り（分かち書き誤り，品詞誤り等）は，同じ誤りが複数箇所で生じることがある.「TagEdit」は，１文に限らず，複数の文を選択した状態でも起動することができる．複数文を選択した場合には，TagEdit は，それらの文に含まれる共通部分のみを修正の対象とする．

1 文の選択は，該当文をマウスでクリックすることにより選択できるが，複数の文の選択は，shift キーを用いた範囲選択，あるいは，Control キーを用いた選択を行うことも可能である．また，図 19 に示すように各文にはチェックボックスが用意されているので，修正が必要な文のチェックボックスを選択し，図のように　Edit → Select Checked により，チェックした文の選択を行うことができる．

形態素情報の修正として考えら得るのは，次の 2 通りの場合がある．

1. 品詞の誤り：単語は適切に区切られているが，誤った品詞がタグ付けされている場合
2. 形態素境界の区切り誤り：区切り誤りには，不必要な箇所での形態素境界の挿入，すなわち，区切り過ぎ，および，形態素境界であるべき箇所に区切りが入っていない２通りの誤りが考えられる．

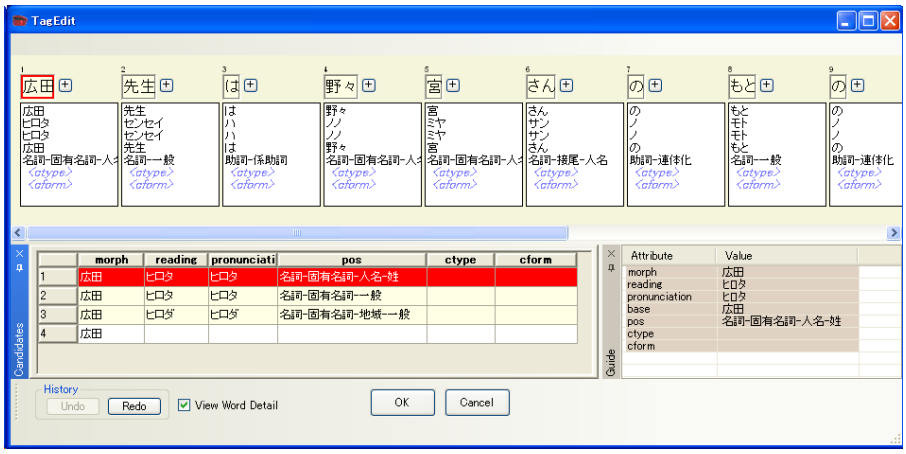



図 20: TagEdit の実行

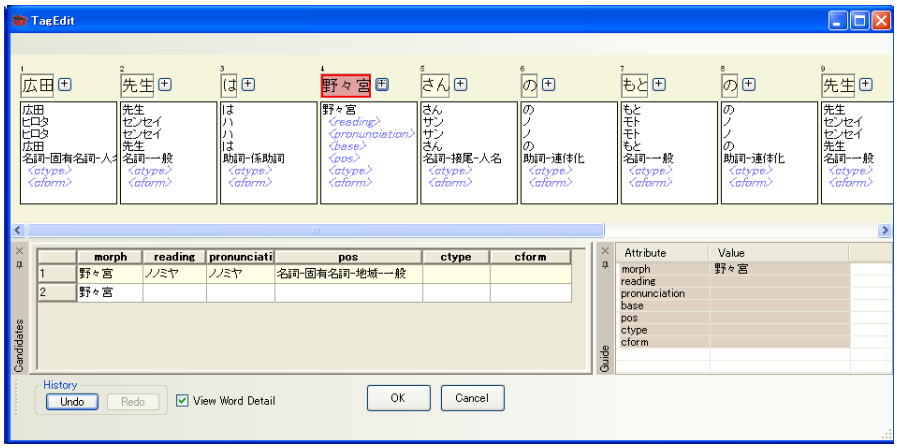



図 21: TagEdit による単語の結合

文を選択して，図 18 で示したように TagEdit を起動すると，図 20 に示すような形態素解析誤り修正ツールが立ち上がる．この例では,「広田」という単語がマウスで選択され，下方の画面に，辞書内に「広田」と

して登録されている語の一覧が表示されている．現在の辞書では,「広田」として３通りの品詞が可能であることがわかる．

品詞誤りの修正は，このように示された一覧の中から正しい語を選択することによって行うことができる．図の例に見られるように，単語一覧の最後の行が空欄になっている．辞書一覧の中に正しい品詞がない場合には，この空欄に正しい情報を記入して，それを選択すればよい．

図 21 に，過度に分かち書きされた形態素の結合作業の例を示す．図 20 では，本来「野々宮」という人名であるべき語が「野々」と「宮」に分かち書きされてしまっていたが,「野々」のすぐ右にある「＋」と書かれた小さなボックスをマウスで洗濯することにより，図 21 のように，その前後の語が「野々宮」のように一語にまとめられている．下部の辞書内の当該単語一覧に正しい品詞があれば，それを選択することで修正が完了するが，この例のように「野々宮」として地名しか登録されていない場合は，2 行目の空欄に人名に必要な情報を書き込んで，それを選択することにより修正を行うことができる．

本来分割すべき複数の語が一語として解析されている場合は，分かち書きをおこなうべき文字の位置にマウスを移動し，左クリックすることにより 2 つの語に分割することができる．分割された語の品詞等の情報は，上記と同様に行うことができる．

## 5.2　係り受け誤りの修正

係り受け解析済みのコーパスを対象にしている場合は，4.3 節の図 8 で示したように，文を選択して「TreeEdit」を Tools メニューあるいは右クリックで選択すれば，係り受け木が表示される．

係り受け解析の誤りは，文節区切り誤り，係り先誤り，係り受け関係誤りがありうる．係り受け関係名の誤りについては，各枝に示されたラベル (以下の図では D という関係名) の位置でマウスの右ボタンをクリックすれば，可能なラベルの一覧が得られるので，そこから選択すればよい．ただし，可能なラベル (係り受け関係名) の一覧は，“.dps” という拡張子を持つファイルで定義されている必要があり，そのファイルをコーパス定義ファイルの中で指定する必要がある．

具体的には，例えばコーパス定義ファイルが “kokoro.def” であるとき，その中に次のような１行を入れる．

```
deplist=cabocha.dps
```

そして，係り関係名定義ファイル (この例では，“cabocha.dps”) の内容を次のように定義する．つまり，係り関係名としては，A, D, P の３通りが定義される．これらの関係名は，係り受け検索の質問画面でも利用され，係り関係を表す矢印のラベルなを，マウスボタンを押して選択するためにも用いられる．最後の行のアスタリスクは，検索質問でのみ用いられるラベルで，任意の関係名とマッチする．

```
A
D
P
*
```

文節区切り誤りには，本来２つ以上の文節に分かれているべき文節が１つにまとまっている場合と，１つの文節を２つ以上に分けなければならない場合がある．前者の場合の誤り修正を行うには，文節の直後にある小さな「＋」をもつボックスをマウスでクリックすると，その文節と直後の文節が一つにまとめられる．ただし，これらの間に直接の係り受け関係がなければならない．後者の場合は，文節区切りを行いたい部分にマウスを近づけるとマウスがハサミの形になるので，そこで左クリックすると，その場所で文節が２つに分割される．

例えば，図 22 において，二つの文節に分かれている「その」と「時」を一つの文節としてまとめたい場合，まず，これらを含む２つの文節を結合するために，前の文節「その」の直後にある小さな「＋」をもつボックスをマウスでクリックする．その結果，図 23 のように，これらの文節が１つに結合される．

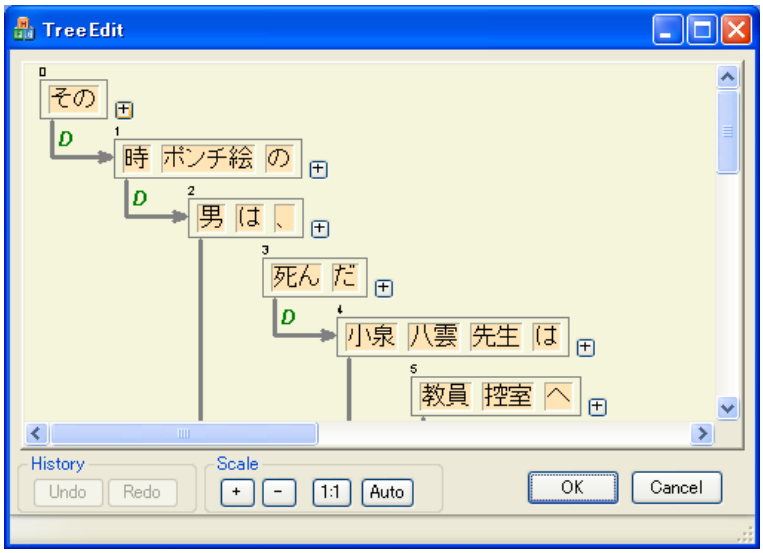



図 22: 係り受け解析誤り修正画面

この文節をあらためて正しく分割するためには,「その時」と「ポンチ絵の」の間にマウスを持っていくと，マウスがハサミの形に変わるので，その場で左ボタンをクリックする．図 24 がその結果を示しており，元の文節が「その時」と「ポンチ絵の」の 2 つの文節に分割されている．

文節の係り先の修正を行う場合は，該当する矢印の先頭をマウスの左ボタンで選択し，そのまま正しい係り先の文節の先頭の位置でドロップすることにより係り先文節の修正を行うことができる．

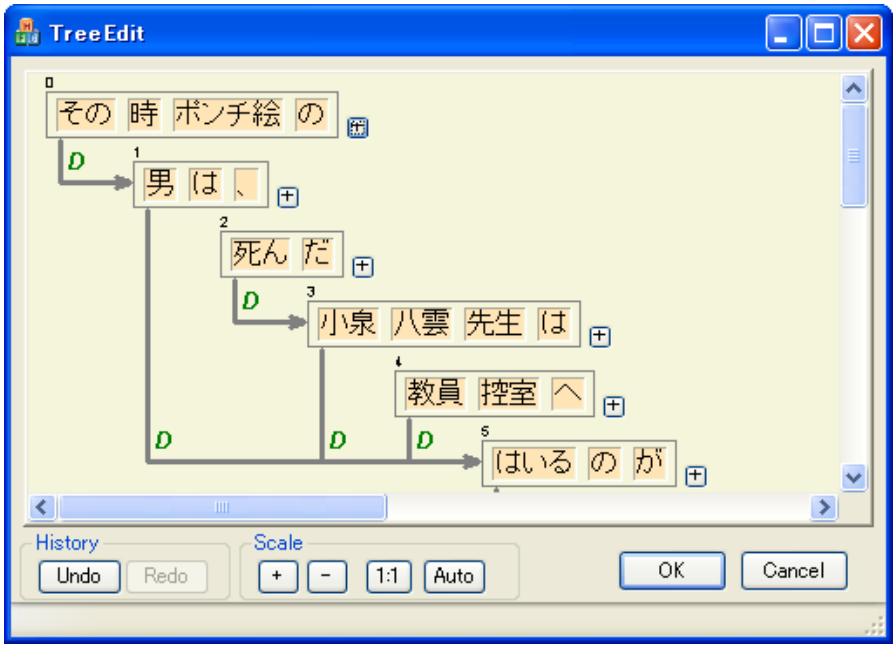



図 23: 文節の結合結果

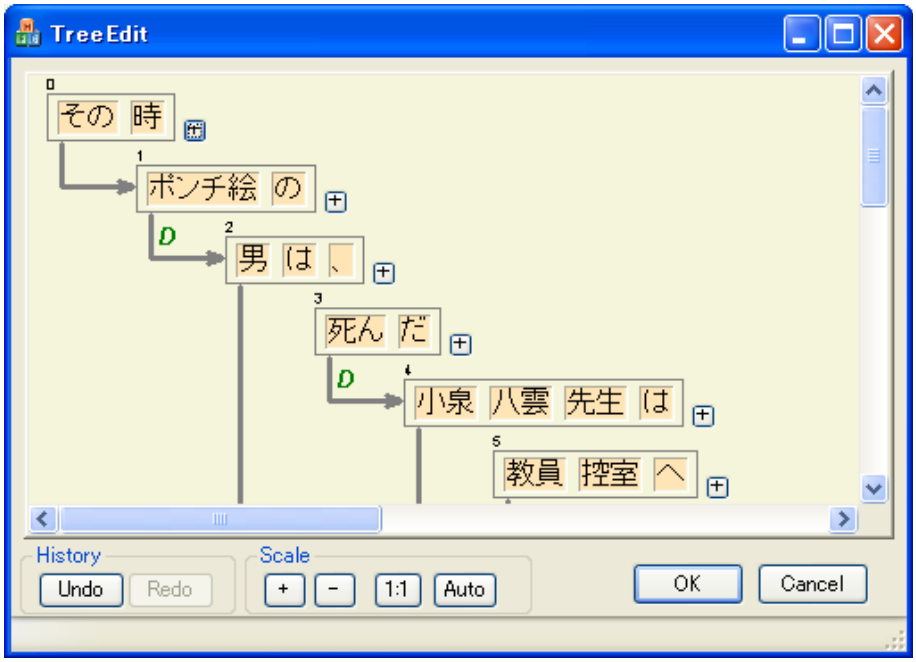



図 24: 文節の再分割の結果

## 謝辞

本システムの構築に対して得られた資金提供，および，構築に協力いただいた諸氏に感謝します．本システムの version1.0 は，科学研究費補助金基盤研究 (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 (課題番号:15300046)」(平成 15〜17 年度) の支援を得ました．その後の機能拡張については，科学研究補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築」の一環として「書き言葉コーパスの自動アノテーションの研究 (課題番号:18061005)」による支援を得ています．

初期のシステム構築について協力をいただいた高岡一馬氏，高橋由梨加さんに感謝します．また，その他様々な協力をいただいた奈良先端大の学生諸君に感謝します．

# A　コーパス定義ファイル

コーパス定義ファイルは，拡張子.def をもつファイルで，コーパス名 (corpusname)，データベースサーバー名 (server)，ユーザ名 (user)，パスワード (password) などの基本情報を記述する．また，単語検索時等に品詞を指定する場合，品詞一覧を記述した「品詞定義ファイル」がある場合，その名前を記述する (poslist)．品詞定義ファイルがコーパス定義ファイルと同じフォルダにある場合は，その名称だけを指定すればよい (attrs)．また，単語検索時等に検索対象の属性を明示的にこのファイルで指定することができる．

品詞定義ファイルの詳細は付録Bに示すが，階層的な品詞定義のうち，最後の名称だけを品詞名として用いる場合は，”postype=1” という１行を（下記の例のように）入れること．

以下が，Penn Treebank（コーパス名が ptb）に対するコーパス定義ファイルの例である．例えば，この内容を ptb.def という名前のファイルに格納すればよい．

```
corpusname=ptb
server=localhost
user=root
password=okage
poslist=PTB.pos
postype=1
attrs=morph
attrs=base
attrs=pos
stopwords=EOS BOS ， ．
```

以下が，RWCP コーパス（コーパス名 rwcp）に対するコーパス定義ファイルの例である．この内容を，例えば rwcp.def という名前のファイルに格納すればよい．このように，attrs を省略すると，品詞タグ付きデータが標準にもつ７つの属性すべて（morph, reading, pronunciation, base, pos, ctype, cform）が表示対象になる．

```
corpusname=rwcp
server=localhost
user=root
password=okage
poslist=rwcp.pos
stopwords=EOS BOS 、 。
```

# B　品詞定義ファイル

品詞定義ファイルは，コーパスで用いている品詞のリストを格納するテキスト形式のファイルである．

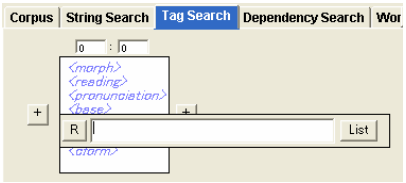



図 25: 品詞名ボックス

単語検索時に品詞属性を指定する際に，以下のような品詞名を記入するボックスが現れるが，品詞定義ファイルを「コーパス定義ファイル」内で指定しておくと，図 25 の「List」ボタンをクリックすることで，図 26 のように品詞一覧が表示され，その中から品詞名を選択することができるようになる．

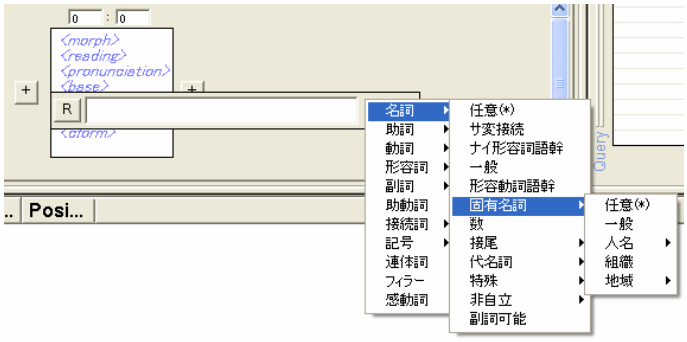



図 26: 品詞名の選択

品詞定義ファイルは，付録Aで述べたとおり，コーパス定義ファイル内で指定する．品詞定義ファイルの内容は，以下に示すように階層構造をカンマによって記述する（以下の定義は，図 26 で示されている階層構造に対応している）．

名詞
, 任意 (*)
, サ変接続
, ナイ形容詞語幹
, 一般
, 形容動詞語幹
, 固有名詞
, , 任意 (*)
, , 一般
, , 人名
, , , 任意 (*)
, , , 名
, , , 姓

なお，Penn Treebank や British National Corpus では，品詞は階層的には定義されていないが，品詞数は，50 あるいは 70 種類に及び，これらすべてを一階層で表示するのは煩雑である．品詞定義ファイル内で，名詞類や動詞類を階層的に定義することにより，図 27 のように品詞定義を階層的に行うことができるが，品詞名自体は，図中の「NNS」のように単一名にする必要がある．コーパス定義ファイルに次の一行を記入することにより，品詞定義ファイルでは階層的な品詞定義を行いつつ，品詞名の指定は，階層の最下部（木の葉の位置）の品詞名のみを入力するようにできる．

```
postype=1
```

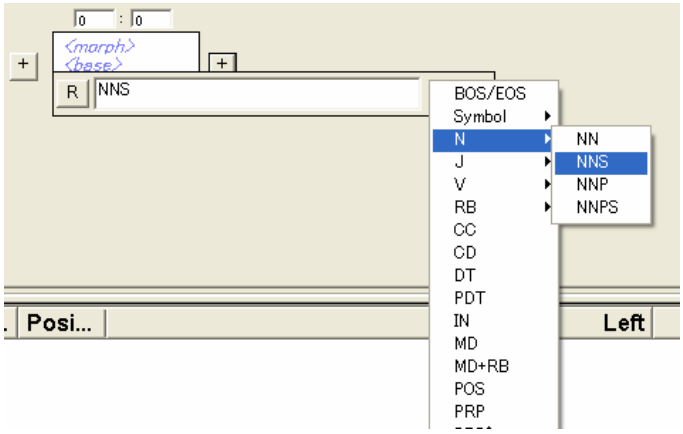



図 27: 英語の品詞名の選択