Vol.2010-NL-199 No.18
2010/11/19

形態素・係り受け解析済みコーパス管理・検索ツール「茶器」

松本裕治[†1]　浅原正幸[†1]
岩立将和[†1]　森田敏生[†2]

科研費領域研究研究「日本語コーパス」の一環として開発してきたコーパス管理ツール「茶器」の機能と現状について報告する．茶器は，形態素解析 (品詞情報)，係り受け解析のアノテーション（注釈）が付与されたコーパスを格納し，様々な検索，検索結果や統計情報の表示，注釈誤りの修正などの機能をもつツールであり，注釈付きコーパスの格納，検索，作成，修正のための環境を提供する．主な機能は，文字列，形態素列，文節係り受け構造などを指定したコーパスの検索と，検索結果の KWIC 表示と係り受け木の表示，種々の統計情報の表示，注釈付けエラーの修正などである．現在は，茶筌/MeCab による形態素解析，南瓜による係り受け解析結果をデータベースに取り込む機能を提供するが，特に言語には依存せず，任意の言語の品詞/依存構造注釈付きコーパスを扱うことができる．

Morphological and Dependency Structure Annotated Corpus Management Tool: ChaKi

Yuji Matsumoto,[†1] Masayuki Asahara,[†1]
Masakazu Iwatate[†1] and Toshio Morita[†2]

This paper introduces a annotated corpus management system ChaKi that has been developed under the auspices of the Japanese Corpus Project (Grant-in-Aid for Scientific Research in Priority Areas). The system handles morphologican and dependency structure annotated corpora and facilitates various functions such as storing, retrieving, creating and error-correcting annotated corpora. String, word and dependency structure based corpus retrievals are possible, and the results are shown as KWIC format or as dependency trees. While the current system transfers corpora with the ChaSen/MeCab or CaboCha output format into databases, it is language independent and can be applied flexibly to any POS/dependency structure annotated corpora.

1. まえがき

言語研究におけるコーパス（電子的に扱うことのできるテキストデータを総称してコーパスと呼ぶ）の重要性は説明するまでもない．コーパスの利用は，その量と質の両方向に拡がりを見せている．大規模な未解析（あるいは自動解析された）コーパスの利用と，詳細な言語情報が付与された注釈付きコーパスの整備である．品詞情報と句構造情報が付与された Penn Treebank[1] の果たした役割は大きい．日本語でも京都大学テキストコーパス[2] が，単語，文節分かち書き，係り受け解析情報が付与された treebank として標準的に利用されてきた．

注釈付きコーパスを利用し，統計的機械学習を用いた言語解析法の提案が盛んに行われ，形態素解析や統語解析について実用的なシステムの構築が可能になった．精密な注釈付けは最終的には人手に頼るしかないが，上記のように多くの利用者をもつコーパスにも多くの注釈付け誤りが含まれることが知られている．他方，コーパスを用いた言語研究には，大規模な注釈付きコーパスを柔軟に検索するツールの使用が必須であるが，簡便に利用可能なツールは限定された機能しかもたないことが多い．本稿で紹介する「茶器」は，形態素，文節分かち書き，係り受け構造（同格・並列構造，括弧等に埋め込まれた表現への対応も含む）の付与されたコーパスを柔軟に検索し，また，発見された注釈付け誤りの修正，平文からの上記の注釈付け機能，などを提供するツールであり，精度の高い注釈付けコーパスの作成と利用を目的としたツールである．

コーパスの検索については，文字列あるいは単語列の検索と結果の表示については，例えば，WordSmith [*1] に代表される様々な KWIC concordancer が入手可能であるが，品詞等の文法情報や統語情報を検索の対象としているものは少ない．

我々は，特定領域研究「日本語コーパス」プロジェクト[*2] にツール班として参画しており，茶器は，その一環として作られたツールである．本コーパスのコアデータ（約 100 万

†1 奈良先端科学技術大学院大学 ({matsu,masayu-a,masakazu-i}@is.naist.jp)
Nara Institute of Science and Technology
†2 総和技研 (morita@sowa.com)
Sowa Giken Corp.
*1 http://www.lexically.net/wordsmith/version5/index.html
*2 正式名称，文部科学省科学研究費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」(代表者：前川喜久雄)(研究期間：2006 年度〜2010 年度)

Vol.2010-NL-199 No.18
2010/11/19

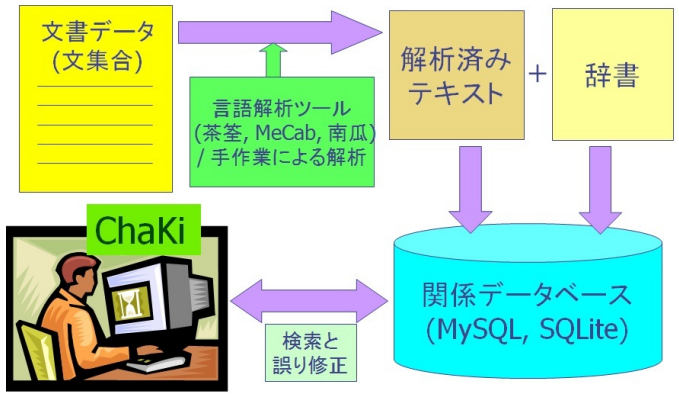

図 1 茶器の概念図
Fig. 1 Configuration of ChaKi

語）に文節係り受け情報までの注釈付与（の修正作業）に使われている．また，言語研究へのコーパス利用ツールとしても位置づけている．なお,「日本語コーパス」には係り受け以上の情報（固有表現，述語項構造，共参照）も付与する予定であり，このような広い範囲の注釈付けを対象とした汎用のコーパスアノテーションツール Slate[3] が同メンバーである徳永研究室で開発されている．

2. 茶器の概要

茶器利用の概念図を図 1 に示す． 文書データは，自動形態素解析，係り受け解析を施し，あるいは，人手によって注釈付けしたものを用意する．NAIST-jdic（茶筌の標準辞書）と UniDic[4] に対応した茶筌[⋆3]，MeCab[⋆4]，南瓜[⋆5] の出力をデータベースに変換するツールが備わっている．ただし，これらの解析器の出力のみに対応しているという意味ではなく，南瓜の出力形式（京都大学テキストコーパスと同様）に準じたフォーマットであれば，日本語に限定せずどのようなコーパスでも扱うことができる．

旧来の茶器[5] は Visual C++と Ruby により実装されていたが，現在の茶器は，Microsoft .NET Framework/C#上に移植されている（旧版とを区別する場合は，ChaKi.NET と表

⋆3 http://chasen-legacy.sourceforge.jp/
⋆4 http://sourceforge.net/projects/mecab/
⋆5 http://sourceforge.net/projects/cabocha/

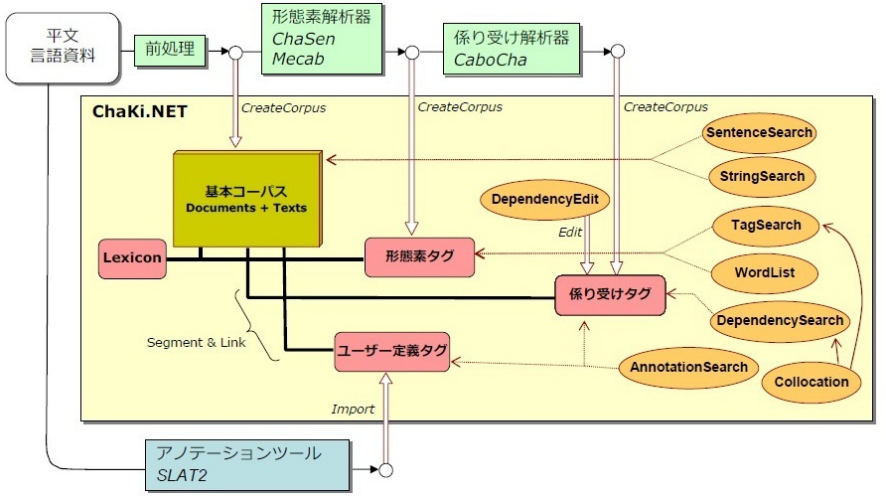

図 2 茶器の内部構成図
Fig. 2 Internal Structure of ChaKi

示）．データベースシステムには関係データベースを用いており，Client-Server でないファイルベースの DB である SQLite をはじめ各種 Client-Server 型の RDB（MySQL, SQL-Express, PostgreSQL）に対応している．茶器の GUI 部がこれらのデータベースに対する様々な検索や修正機能を提供する．

図 2 に茶器の内部構成図を示す． 茶器への入力テキストファイルは，生文（１文が１行に整形されている必要がある），形態素解析結果，あるいは，文節係り受け結果（文節は必須ではない）のいずれかからなる．例えば，係り受け解析済みコーパスが入力の場合は，そこから形態素情報や文情報も抽出される．辞書 (Lexicon) は必須ではないが，データベース内部では，コーパス中の形態素は辞書項目へのポインタとして定義されており，コーパス中の形態素の一覧が辞書となる．コーパスの形態素情報の修正を行う際には，辞書項目へのポインタの付け替えによって実現されるが，辞書に定義されていない形態素については，利用者が品詞や読みなどの情報を明示的に記述しなければならない．辞書が有用になるのは，形態素情報の修正作業時のみである．文字列，形態素列，係り受け木，それぞれに対して検索 (Search) 機能が実装されている．DependencyEdit は，係り受け木を表示し，文節の分かち書き，係り受け関係，形態素情報などすべての情報の修正作業を行うことができる．ま

Vol.2010-NL-199 No.18
2010/11/19

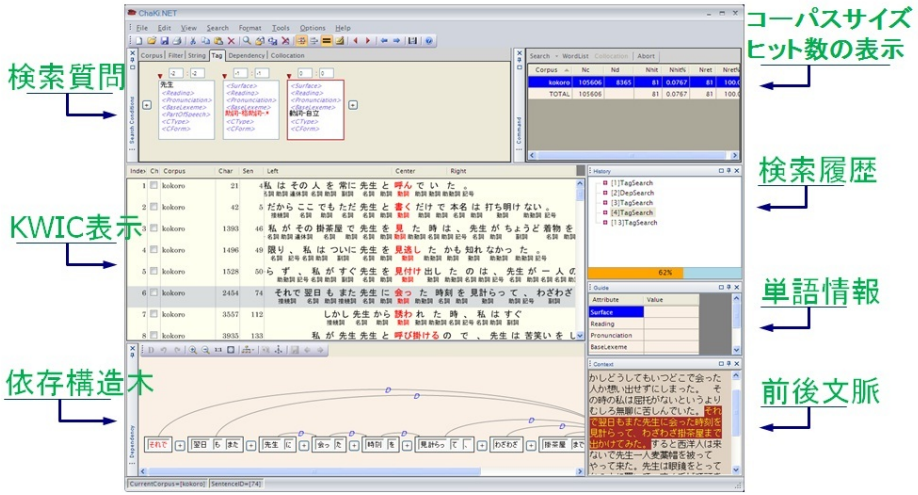

図 3 茶器の実行画面
Fig. 3 Snap Shot of ChaKi in Use

た，並列構造，同格構造，埋め込み要素の範囲指定などの注釈付け機能も提供する．汎用アノテーションツール Slate では，セグメントとリンクに基づく汎用の注釈付けを行うことができるが，茶器との間でデータベーススキーマの統一化を行っており，両ツールで注釈付けされたデータベースは相互参照可能になる予定である．

3. 茶器の諸機能

茶器でコーパスを扱うには，事前にコーパスから作成したデータベースが必要だが，SQLite の場合は，データベースが一つのファイルからなるため，そのファイルをコピーするれば茶器からアクセスが可能になる．図 3 に茶器の実行画面を示す．この画面は，「先生」という単語，「助詞-格助詞」を品詞名にもつ単語，さらに「動詞-自立」という品詞をもつ単語の 3 つが続けて出現する文を検索する質問が発せられた時の画面を示している．このパターンにマッチする文の一覧が KWIC(Keywords in Context) 形式で表示され，選択した文の係り受け木（依存構造木）や文章中で前後に現われる文が表示されている．このように茶器は多くのペインからなるが，各ペインの配置や表示の有無を利用者が指定することができる．本節では，茶器の諸機能について説明する．

図 4 検索質問の種類の一欄と検索質問例
Fig. 4 Types and examples of search queries

3.1 コーパスの指定

同じ品詞体系に基づくコーパスであれば，複数のコーパスを指定して同時に検索質問を発行することが可能である．データベースは SQLite の場合は，データベースファイルそのものを，その他の関係データベースについてはコーパス毎に作られるコーパス定義ファイルを指定する．

3.2 コーパス検索機能

コーパスに出現する文字列，単語列，係り受け構造の 3 種類の検索が可能である．図 4 にそれぞれの検索質問の例を示す．

文字列検索: 任意の文字列での検索が可能．日本語では意味がないが，英語等のアルファベットで大文字と小文字を区別するかどうか指定できる．また，文字列の指定には，正規表現を用いることができる．他の検索機能でも，文字列入力部では正規表現を利用することができる．

単語列検索: 各単語は，表層文字列以外に，読み，発音，品詞（活用する語は，それに加えて活用型，活用形，基本形）をもち，どの情報を与えても検索を行うことができる．図 4 の「単語列検索質問」の例の各単語を表わすボックスの上の小さな 2 つのボックスは，単語の相対位置を表わす．いずれかの単語が (0,0) の値をもち，その単語から相対的に何個から何個の範囲にあるかを 2 つの数値を用いて指定すればよい．

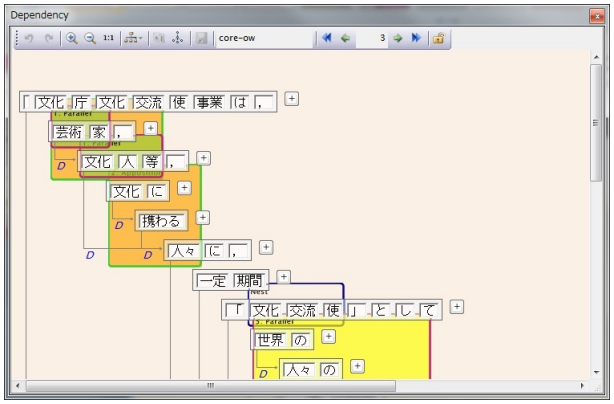

図 5 係り受け，同格，並列構造のアノテーション例
Fig. 5 Sample of annotation with dependency, apposition and coordination

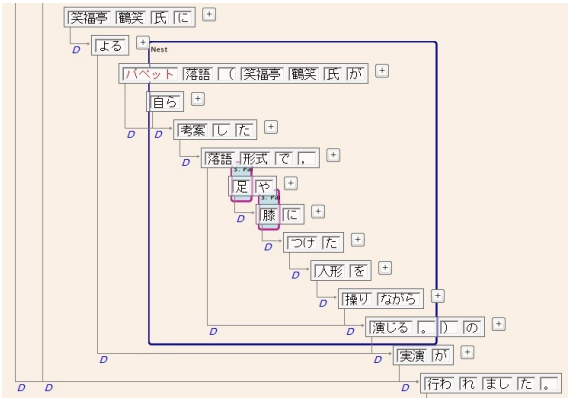

図 6 埋め込み構造のアノテーション例
Fig. 6 Sample of embedded structure

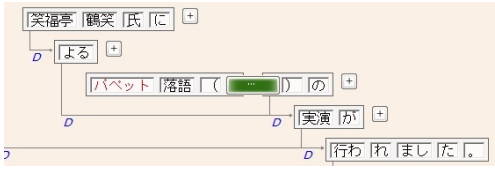

図 7 埋め込み構造の縮退例
Fig. 7 Sample of truncated embedded structure

係り受け構造検索: 文節間の係り受け関係を指定した検索質問である．文節はその中に単語（列）を含み，どのような単語を含むかを指定することによって定義する．図 4 の「係り受け構造検索質問」では,「先生」という語と「助詞-格助詞」を品詞としてもつ語が連続して現われる文節と,「動詞-自立」を品詞としてもつ語を含む文節が係り受け関係にあるようなそのような構造を含む文を検索している．各文節にはこれら以外の単語が含まれていても問題ない．

どの質問についても，検索対象は質問と一致する文であり，その一覧が図 3 のように KWIC 表示される．例外として，WordList という検索があり，検索にマッチした単語（検索が複数の単語をもてば，それら組）の一覧が頻度と共に表示される．

3.3 検索結果の表示

検索質問が発せられると，図 3 の右上のペインに，各コーパスに含まれる総単語数，総異なり単語数，検索質問のヒット数などが表示され，ヒットした文が KWIC 表示される．各単語は様々な情報をもつが，KWIC 画面には 2 つまでの情報が表示できる．また，単語の様々な属性（品詞，活用形，頻度など）を指定して，単語の背景色とフォント色を指定できる．この機能のより，特定の品詞のみをハイライトしたり，頻度が低い単語のみをハイライトすることが可能である．

検索結果の文を選択してダブルクリックすると係り受け構造が表示される．図 5 は，係り受け構造ペインの一部を示している．係り受け関係が D というラベルを持つ枝で表され，並列構造（「芸術家」と「文化人等」）と同格構造（「芸術家，文化人等」と「文化に携わる人々」）が範囲を示すボックスによって注釈付けされている．このように，現在は，係り受け関係と並列・同格は独立に注釈付けが行われている．

範囲指定が必要な別の例として，括弧などにより埋め込まれた構造がある．図 6 は,「パペット落語」の説明文が括弧の中に記述された例のアノテーションを示す．Nest というラベルのついたボックスは，その外側と内側のアノテーションが独立に行われることを示している． 図 7 に示すように，このボックス表示は（並列，同格とも同様）縮退して表示からは隠してしまうことができ，その場合は，埋め込み構造を省いた外側の文を表示することができる．

3.4 付加的な情報の表示

上記以外に，検索結果に関連する様々な情報を表示する機能がある．図 3 の右側にいくつ

かの小さなペインが表示されているが，ここには表示されていないペインも存在する．主なものは次の通りである．

単語情報: KWIC 画面には，上で述べたように単語がもつ様々な属性のうち 2 つまでを表示することができる．図 3 では，表層形と品詞のみが表示されている．その他の属性情報を見たい場合には，マウスを単語の上に持って行くと，単語情報ペインにその単語がもつすべての情報が表示される．UniDic の場合，各単語は多数の属性をもつため，茶器では 9 つの属性を特に表示する仕様になっている．どの属性を単語ペインに表示するかは，利用者がカスタマイズすることができる．

文の書誌情報: 各文には出典や著者などの書誌情報を付与することができる．図 3 ではそのペインが表示されていないが，書誌情報としては，出典のタイトル（書名），著者，出版社，出版年などがあり，元のコーパスとは別ファイルで指定する必要がある．書誌情報に関するファイルは必須ではない．

文脈情報: 検索結果の文の前後に現われる文を，文数を指定して文脈ペインに表示することができる．図 3 では「前後文脈」として示されている．

検索履歴: このペインでは，茶器を立ち上げてからの検索の履歴が表示されるので，以前に行った検索の結果をいつでも表示することができる．

3.5 統計情報の表示

言語研究でコーパスを利用する際には，検索あるいはコーパス中の文全体に対して統計情報を取得する必要がある場合が多い．茶器では，検索結果の KWIC 表示に対して，次のような統計情報の計算を行う．

前後共起: KWIC の中心語の前後の指定された window 幅内にどの単語が何回出現しているかの頻度情報を表にして表示する

N-gram 統計: KWIC の中心語から右あるいは左に連続して続く単語 N-gram の頻度統計．最低出現回数と N の最小値を指定して，候補を制限することができる．例えば，N-gram(Right) を選択し，Minimum Frequency を 5，Minimum Length を 4 に指定すると，KWIC の中心から右へ向かって長さが 4 以上でかつ出現頻度が 5 回以上の単語 N-gram の一覧が，頻度と長さとともに表示される．

頻出系列マイニング: KWIC の中心語とは関係なく，ヒットした文集合を対象にして，頻度の高い単語列を抽出する．一般に，対象とする単語列はギャップを含むことを許す．頻度と長さ（単語数）以外に，ギャップの数の上限，およびギャップの長さ（ギャップに含まれる単語数）の上限を指定して対象を絞ることができる．

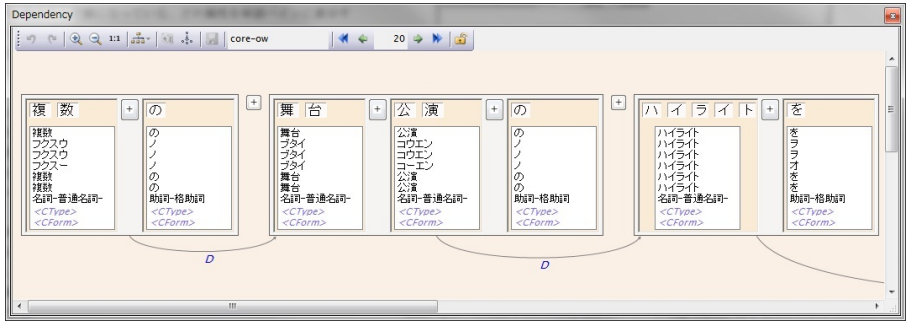

図 8 係り受け構造の水平表示と単語の詳細表示
Fig. 8 Flat display of dependency structure and full description of lexical entries

4. 注釈付きコーパスの修正，作成機能

茶器は，茶筌，MeCab などの形態素解析，または，南瓜などの係り受け解析済みのコーパスを格納することを想定しているが，このような自動解析されたコーパスには様々な注釈付け誤りが含まれている．また，茶器は未解析のコーパスを読み込むこともできる．これらのコーパスの注釈付け情報の修正，あるいは，新しく注釈情報を付与する機能が重要だと考えている．図 5 で示した係り受け構造の表示ペインには，図 8 のように係り受け木を水平表示し，かつ，単語のもつ情報をすべて表示するモードがある． この表示画面を用いて，誤った単語の選択し直し，単語の分かち書き誤りの修正，文節分かち書き誤りの修正，係り受け関係の修正，および，これらすべての情報の新規の付与が可能である．図 5, 図 6 に示した同格，並列，埋め込みに関する範囲の付与もこの画面上でマウス操作によって行うことができる．

5. あ と が き

特定領域研究「日本語コーパス」のツール班の活動として構築してきたコーパス管理ツール「茶器」の概要について説明した．ここでは記述しなかった様々な機能があるが，開発中のツールであり，本稿で紹介した機能の中にもまだ不完全な点が残っている．

言語学研究のためにコーパスを簡単に利用できる環境を，そして，自然言語処理研究のために注釈付け誤りの修正を簡単に施して精度の高いコーパス作成環境を実現することを目指して本ツールを開発してきたが，当初予定の機能はほぼ完成することができた．

Vol.2010-NL-199 No.18
2010/11/19

茶器は，内部文字コードとして UTF-16 を使用しており，多言語に対応できる．研究室内では，日本語以外に中国語と Tagalog 語のコーパスの注釈付けに利用している．英語でも利用可能だが，単語間の空白の扱いについての仕様がまだ不確定である．現在想定している解析済みコーパスのフォーマットでは，形態素列の形態素間に空白が存在したかどうかに関する情報がないため，すべての単語の間に空白があるかないかの解釈をすることになり，例えば，punctuation mark の前後で原文に空白がなかった場合の情報を記録する方法が固まっていない．

一般的なデータベースシステムを利用していることで，柔軟な検索質問を受け付けることができる反面，大規模なコーパスの検索には効率面で問題がある．数百万語規模のコーパスの検索には大きな支障はないが，数千万語規模のコーパスの検索には通常のデスクトップやノートパソコンでは効率に問題があると言わざるを得ない．現時点では，コーパスを数百万語単位の複数のコーパスに分割し，これらすべてを対象に検索することにより，最初の解を早く得ることができる．今後，並列処理を利用した高速化を今後目指したいと考えている．

なお，本ツールは以下のページからダウンロード可能であり，簡単なオンラインマニュアルも用意されている．利用いただき，忌憚のない意見をいただければありがたい．

`http://sourceforge.jp/projects/chaki/`

謝辞 本ツールの開発に協力いただいた奈良先端大自然言語処理学講座のスタッフと学生諸君に感謝します．なお，本ツールは，文部科学省科学研究費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」の支援を得て構築されました．本プロジェクトの前川喜久雄代表を始め，ツール班の班員の皆さん，本ツールを利用し様々なコメント，要望をお寄せいただいた利用者の皆さんに感謝します．

参 考 文 献

1) Marcus, M.P.Santorini, B.and Marcinkiewicz, M.A.: Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, Vol. 2, No. 2, pp.313–330, (1993).
2) 京都大学テキストコーパス Version 4.0:
`http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html`
3) 徳永健伸，Dain Kaplan，飯田龍：汎用アノテーションツール Slate, 本研究会資料, (2010).
4) 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: 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, 22 号, pp.101–122, (2007).
5) Yuji Matsumoto, et al: An Annotated Corpus Management Tool: ChaKi, Proceedings of the 5th International Conference on Language Resources and Evaluation, (2006).