形態素解析システム
『茶筌』version 2.4.3
使用説明書

松本裕治　高岡一馬　浅原正幸

平成 20 年 5 月

Morphological Analysis System ChaSen 2.4.0 Users Manual
Yuji Matsumoto, Kazuma Takaoka and Masayuki Asahara

JUMAN	
version 0.6	17 February 1992
version 0.8	14 April 1992
version 1.0	25 February 1993
version 2.0	11 July 1994
ChaSen	
version 1.0	19 February 1997
version 1.5	7 July 1997
version 2.0	15 December 1999
version 2.2.0	06 December 2000
version 2.3.0	16 February 2003
version 2.4.0	30 March 2007
ChaSen for Windows	
version 1.0	29 March 1997
version 2.0	15 December 1999
version 2.4.0	30 March 2007
version 2.4.1	3 July 2007
version 2.4.2	23 July 2007
version 2.4.3	30 May 2008
NAIST Technical Report	
1st edition(NAIST-IS-TR99008)	20 April 1999
2nd edition(NAIST-IS-TR99012)	15 December 1999

目 次

1 茶筌の使用法 **2**
1.1 インストール手順 2
1.2 実行方法 3
1.3 実行時のオプション 4
1.4 出力フォーマット 4
1.5 制約つき解析 7

2 chasenrc ファイル **9**

3 茶筌ライブラリ **13**

4 他のシステムからの利用 **14**
4.1 Perl からの使用 14

参考文献 **14**

付録 **17**

A 著作権および使用条件について **17**

B 更新履歴 **17**
B.1 茶筌 2.3.3 から 茶筌 2.4.0 への変更点 17
B.2 茶筌 2.3.2 から 茶筌 2.3.3 への変更点 17
B.3 茶筌 2.3.1 から 茶筌 2.3.2 への変更点 17
B.4 茶筌 2.3.0 から 茶筌 2.3.1 への変更点 17
B.5 茶筌 2.2 から 茶筌 2.3 への変更点 18
B.6 茶筌 2.0 から 茶筌 2.2 への拡張点 18
B.7 JUMAN 2.0 から 茶筌 2.0 への拡張点 18
B.8 茶筌 1.5 から 茶筌 2.0 への拡張点 19
B.9 茶筌 1.0 から 茶筌 1.5 への拡張点 20
B.10 JUMAN 2.0 から 茶筌 1.0 への拡張点 20

C JUMAN3.0 と 茶筌 との関係について **21**

D 形態素解析器の今後について **22**

はじめに

計算機による日本語の解析において，欧米の言語の解析と比べてまず問題になるのに次の 2 点があります．一つは形態素解析の問題です．ワードプロセッサの普及などによって日本語の入力には大きな問題がなくなりましたが，計算機による日本語解析では，まず入力文内の個々の形態素を認識する必要があります．これには実用に耐えられるだけの大きな辞書も必要であり，これを如何に整備するかという問題も同時に存在します．もう一つの問題として，日本語には広く認められ同意を得られた文法，ないし，文法用語がないという現実です．学校文法の単語分類および文法用語は一般には広く知られていますが，研究者の間ではあまり評判がよくありませんし，計算機向きではありません．

日本語の解析に真っ先に必要な形態素解析システムは，多くの研究グループによって既に開発され技術的な問題が洗い出されているにも係わらず，共通のツールとして世の中に流布しているものはありません．計算機可読な日本語辞書についても同様です．

本システムは，計算機による日本語の解析の研究を目指す多くの研究者に共通に使える形態素解析ツールを提供するために開発されました．その際，上の二つ目の問題を考慮し，使用者によって文法の定義，単語間の接続関係の定義などを容易に変更できるように配慮しました．

大学で小人数で開発したシステムであり，色々な点で不完全な部分があると思います．可能な限り順次改良を重ねる予定です．皆様の寛容な利用をお願いいたします．

本茶筌システムの原形は，京都大学長尾研究室および奈良先端科学技術大学院大学情報科学研究科において開発された日本語形態素解析システム JUMAN(version2.0) です．JUMAN は，京都大学および奈良先端科学技術大学院大学のスタッフおよび多くの学生の協力を得て作成したものです．また，辞書に関しては，Wnn かな漢字変換システムの辞書，および，ICOT から公開された日本語辞書を利用し，独自に修正を加えました．JUMAN 2.0 をともに開発した東京大学の黒橋禎夫氏，現在キヤノン勤務の妙木裕氏には特に感謝いたします．

JUMAN 開発のきっかけを作って下さった長尾真先生に感謝します．JUMAN 開発に関して様々な形で協力していただいた筑波大学宇津呂武仁氏に感謝します．奈良先端大在学時の知念賢一氏には，茶筌システムの開発に関して多くの助言をいただきました．奈良先端大在学時の今一修氏，今村友明氏，北内啓氏には茶筌 1.0 および茶筌 2.0 β版の開発の際に山下達雄氏，平野善隆氏，松田寛氏には茶筌 2.0 版および茶筌 2.2 版の開発の際に種々の助力をいただきました．両氏および茶筌の開発に協力いただいた松本研究室のメンバーに深く感謝します．奈良先端大の鹿野清宏教授を代表とする「日本語ディクテーション基本ソフトウェアの開発」グループの方々には，IPA 品詞体系辞書の大幅な整備を行っていただきました．特に，御尽力いただいた法政大学の伊藤克亘氏，ASTEM の山田篤氏に感謝いたします．話し言葉の解析を中心にして辞書の整備に様々な助言をいただいた千葉大の伝康晴氏に感謝します．奈良先端大在学時の高林哲氏，工藤拓氏には autoconf，automake 化および RPM パッケージ作成に多くの助言をいただきました．ゴーチュイリン氏，鄭育昌氏，呂嘉氏には中国語版辞書の整備に尽力していただきました．また，一人一人の名を挙げることはできませんが，茶筌システムに対して多くのコメントと質問をいただいた利用者の方々に感謝します．

平成 19 年 3 月 30 日

本システムに関するお問い合わせは以下にお願いします．
〒 630-0192
奈良県生駒市高山町 8916-5
奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座
Tel: (0743)72-5240, Fax: (0743)72-5249
E-mail: chasen@is.naist.jp
URL: http://chasen-legacy.sourceforge.jp/

1 茶筌の使用法

1.1 インストール手順

1. 必要なツールをインストールする.

   茶筌をコンパイルするには以下のツールが必要である.

   - Darts[1] バージョン 0.3 以降
   - （システムに標準装備されていなければ）libiconv

2. 'configure' を実行する.

```
% ./configure
```

   - Darts のヘッダファイルを指定する場合

```
% ./configure --with-darts=/usr/local/include
```

   - libiconv を使う場合

```
% ./configure --with-libiconv=yes
```

   - libiconv の場所を指定する場合

```
% ./configure --with-libiconv=/usr/local
```

   コンパイラやコンパイルオプションは自動的に設定される.

   configure の詳しい使用法については INSTALL あるいは './configure --help' の出力を参照のこと.

3. 'make' を実行する.

```
% make
```

   茶筌本体の実行ファイルは chasen/chasen に，ライブラリは lib/ に，辞書作成のプログラムは mkchadic/ 以下に作成される．OS 標準の make を使うとコンパイルに失敗することがある．その場合は GNU make を使用する.

4. 'make install' を実行する.

```
% make install
```

   バージョン 2.1 からインストール先ディレクトリが変更されており，デフォルトでは以下の場所にインストールされる．PREFIX は ./configure --prefix で指定することができる (デフォルトは /usr/local).

[1] http://cl.aist-nara.ac.jp/%7etaku-ku/software/darts/

PREFIX/bin/chasen	茶筌の実行ファイル
PREFIX/libexec/chasen/	辞書作成プログラム
PREFIX/lib/libchasen.*	茶筌ライブラリ
PREFIX/include/chasen.h	ヘッダファイル
PREFIX/share/chasen/doc/	マニュアル

ただし，以下のものはインストールされない．

perl/ChaSen.pm	Perl モジュール

chasenrc はシステムインストール時にはインストールされない．辞書 (ipadic-2.6.0 以降) インストール時に chasen-config から chasenrc のパスを受け取り，PREFIX/etc 以下に chasenrc がない場合に自動的にコピーされる．既に PREFIX/etc 以下に chasenrc がある場合コピーされないため管理者が手で変更する必要がある．

1.2 実行方法

システムの実行ファイルは，‘make install’ によって PREFIX/bin/chasen などにインストールされる．

- 形態素解析の実行

  茶筌は，以下のように chasen コマンドを実行することにより起動される．

```
% chasen [オプション] [ファイル名...]
```

  標準入力，または引数で指定されたファイルから一行ごとに文を読み込んで形態素解析処理を行なう．

- 処理内容

  コスト最小 (それぞれの形態素の区切りで最小コストとの差が許容されるコスト幅以内) の解を求め，結果をオプションに従って表示する．各オプションの意味は次節にまとめる．

- 使用例

  入力ファイルを引数として指定できる．以下に使用例を示す．

```
% cat temp
私は昨日学校へ行きました.
% chasen temp
私      ワタクシ    私      名詞-代名詞-一般
は      ハ          は      助詞-係助詞
昨日    キノウ      昨日    名詞-副詞可能
学校    ガッコウ    学校    名詞-一般
へ      へ          へ      助詞-格助詞-一般
行き    イキ        行く    動詞-自立          五段・カ行促音便    連用形
まし    マシ        ます    助動詞             特殊・マス          連用形
た      タ          た      助動詞             特殊・タ            基本形
.       .           .       記号-句点
EOS
```

1.3 実行時のオプション

形態素解析の実行については，いくつかのオプションが用意されている．以下にそれをまとめる．-r など引数をともなうオプションでは，オプションと引数の間には空白があってもなくてもかまわない．

- 解が曖昧性を含む場合の表示方法 (曖昧性がない場合はどの方法も同じ表示となる)

  -b　後方最長一致の解を一つだけ表示する (デフォルト)
  -m　曖昧性のある部分だけ，複数の形態素を表示する
  -p　曖昧性の組合せを展開し，すべての解を個別に表示する

- 各形態素の表示方法

  -f　カラムを整えて表示 (デフォルト)
  -e　完全な形態素情報を文字で表示
  -c　完全な形態素情報をコードで表示
  のを出力
  -v　VisualMorphs のための詳細表示
  -F format　形態素を format で指定された形式で出力
  -Fh　-F オプションの出力フォーマットのヘルプを表示

- その他

  -j　句点あるいは空行を文の区切りとして解析
  -o file　解析結果出力ファイルを指定
  -w width　コスト幅を指定
  -r rc_file　rc_file を chasenrc ファイルとして使用
  -R　デフォルトの chasenrc ファイル (PREFIX/etc/chasenrc) を読み込む
  -L lang　言語を指定
  -lp　品詞番号と品詞名のリストを表示
  -lt　活用型番号と活用型名のリストを表示
  -lf　活用型番号，活用形番号と活用形名のリストを表示
  -i　入力文の文字コードを選択 (e: EUC-JP, s:Shift_JIS, w:UTF-8, u:UTF-8, a:ISO-8859-1)
  -h　ヘルプメッセージを出力
  -V　茶筌のバージョンを出力
  -s　制約つき解析

-j オプションについて

茶筌では通常，改行をもって一つの入力文字列の終了とする．そのため，文の途中で改行が挿入されているファイルを解析した場合，正しい結果が得られなくなることが多い．

そのようなときは -j オプションをつけると，句読点など (デフォルトでは「.。!?」の 4 文字) あるいは空行を文の区切りとして解析を行うようになる．

また，chasenrc ファイルの「区切り文字」の項目を指定することにより，-j オプションをつけた時の文の区切り文字を設定することができる．

1.4 出力フォーマット

-F オプションや，chasenrc ファイルの「出力フォーマット」で出力フォーマットを指定することにより，解析結果の出力形式を変えることができる．

出力フォーマットの文字列の末尾に '\n' があれば，各形態素情報の表示ごとに改行を行い，文末の次に 'EOS' の 1 行を出力する．末尾に '\n' がなければ，1 文中の形態素情報を 1 行で出力し，行末に改行を表示する．

また，出力フォーマットに '-f', '-e', '-c' を指定すると，それぞれ -f, -e, -c と同じ出力形式になる．

出力フォーマットの使用例をいくつかあげる．

- デフォルト (-f オプション) と同様の出力

  "%m\t%y\t%M\t%U(%P-)\t%T␣\t%F␣\n" または "-f"

- 見出し，読み，品詞をタブで区切って表示

  "%m\t%y\t%P-\n"

- 見出し語のみ

  "%m\n"

- 分かち書き (見出し語を空白で区切って表示)

  "%m␣"

- 漢字かな変換

  "%y"

- ルビつき表示．"漢字 (かな)" の形式で表示する．

  "%r␣()"

出力フォーマットの変換文字の一覧を以下に示す．

変換文字	機能
%m	見出し (出現形)
%M	見出し (基本形)
%y, %y1	読みの第一候補 (出現形)
%Y, %Y1	読みの第一候補 (基本形)
%y0	読み全体 (出現形)
%Y0	読み全体 (基本形)
%a	発音の第一候補 (出現形)
%A	発音の第一候補 (基本形)
%a0	発音全体 (出現形)
%A0	発音全体 (基本形)
%rABC	ルビつきの見出し (“A 漢字 B かな C” と表示)(※ 1)
%i, %i1	付加情報の第一候補
%i0	付加情報全体
%Ic	付加情報 (空文字列か “NIL” なら文字c)(※ 1)
%Pc	各階層の品詞を文字c で区切った文字列
%Pnc	1　n(n:1　9) 階層目までの品詞を文字c で区切った文字列
%h	品詞の番号
%H	品詞文字列
%Hn	n(n:1　9) 階層目の品詞 (なければ最も深い階層)
%b	0(旧版との互換性のみ)
%BB	品詞細分類 (なければ品詞)
%Bc	品詞細分類 (なければ文字c)(※ 1)
%t	活用型の番号
%Tc	活用型 (なければ文字c)(※ 1)
%f	活用形の番号
%Fc	活用形 (なければ文字c)(※ 1)
%c	形態素のコスト
%S	解析文全体
%pb	最適パスであれば “*”, そうでなければ “␣”
%pi	パスの番号
%ps	パスの形態素の開始位置
%pe	パスの形態素の終了位置 +1
%pc	パスのコスト
%ppiC	前に接続するパスの番号を文字C で区切り列挙
%ppcC	前に接続するパスのコストを文字C で区切り列挙
%?B/STR1/STR2/	品詞細分類があればSTR1，なければSTR2(※ 2)
%?I/STR1/STR2/	付加情報が “NIL” でも “”(空文字列) でもなければSTR1，そうでなければSTR2(※ 2)
%?T/STR1/STR2/	活用があればSTR1，なければSTR2(※ 2)
%?F/STR1/STR2/	%?T/STR1/STR2/ と同じ
%?U/STR1/STR2/	未知語ならSTR1\，そうでなければSTR2(※ 2)
%U/STR/	未知語なら”未知語”，そうでなければSTR(%?U/未知語/STR/と同じ)(※ 2)
%%	% そのもの

変換文字	機能
.	フィールド幅の指定
-	フィールド幅の指定
1-9	フィールド幅の指定
\n	改行文字
\t	タブ
\\	\ そのもの
\'	' そのもの
\"	" そのもの

※ **1** ipadic では,「行く (いく/ゆく)」のように形態素が複数の読みを持つ場合，その読みを「{ イ/ユ } ク」のように，半角のブレースとスラッシュを使って表している．通常の読みの出力 (出力フォーマットの %y) では，その第一候補である「イク」が出力され，%y0 を使うと読み全体である「{ イ/ユ } ク」が出力される．

※ **1** A,B,C,c が空白文字の時は何も表示しない．

※ **2** '/' には任意の文字が使える．また，括弧 "(){}[]<>" を用いることもできる．以下に例をあげる．

- %?T#STR1#STR2#
- %?B(STR1)(STR2)
- %?U{STR1}/STR2/
- %U[STR]

1.5 制約つき解析

「制約つき解析」とは，入力文の一部の形態素情報が既知である，あるいは境界がわかっているときに，それを満たすように解析することを云います。

たとえば,「にわにはにわにわとりがいる。」という文に対して,「はにわ」の部分が名詞であるとか,「にわとり」の部分が一つの形態素であるというように指定した上で解析することができます。このとき，制約に反する 4 文字目の「は」が単独で形態素となったり,「にわとり」が「にわ」と「とり」に分割されるような解析候補は排除されます。

入力書式 制約つき解析の入力は茶筌の標準の出力と同じようなフォーマットであたえます。(ĥはタブを表します)

ただし、読み、基本形の情報は無視されます。

```
にわ\t ニワ\t にわ\tUNSPEC
に
はにわ\t ハニワ\t はにわ\t 名詞-一般
にわとり\t ニワトリ\t にわとり\tUNSPEC
がいる。
EOS
```

各行をセグメントと呼び、一つのセグメントは「形態素指定」「文断片」「文末」「注釈」のいづれかになります。

- 形態素指定
  そのセグメントが (それ以上分割されない) 一つの形態素であることを示します。
  形態素指定のセグメントは 4 カラム目以降に品詞情報を持ちます。品詞情報の書式も茶筌の標準の出力と同じです。
  品詞情報の代わりに「UNSPEC」と書くと、セグメントの見出し語で辞書を検索し、該当する語が解析結果となります。辞書にない語はそのまま未知語となります。

- 文断片
  品詞情報がないセグメントは文断片を表します。
  このセグメント内では、制約のない場合と同様に解析されます。ただし、形態素がセグメントをまたぐような解析候補は生成されません。

- 文末
  「EOS」「BOS/EOS」「文末」で始まる行、あるいは改行しか含まない行は文末です。

- 注釈
  品詞情報のカラムを「ANNO」とすると、そのセグメントは注釈になります。
  注釈は出力には表示されますが、解析には使われません。表示は chasenrc に従います。

解析例　以下に解析例を示します。

入力：

```
$ chasen -s
にわ\t ニワ\t にわ\tUNSPEC
に
はにわ\t ハニワ\t はにわ\t 名詞-一般
にわとり\t ニワトリ\t にわとり\tUNSPEC
がいる。
EOS
```

出力：

```
にわ\t\t\t 未知語
に\t ニ\t に\t 助詞-格助詞-一般
はにわ\t ハニワ\はにわ\名詞-一般
にわとり\t ニワトリ\t にわとり\t 名詞-一般
が\t ガ\t が\t 助詞-格助詞-一般
いる\t イル\t いる\t 動詞-自立\t 一段\t 基本形
。\t。\t。\t 記号-句点
EOS
```

制約つき解析における注意点

- 制約つき解析時 “ANNO” を指定しても chasenrc に注釈の指定がなければ出力には何も表示されない.
- 制約つき解析時、空白品詞機能と空白読み飛ばしが無効になる（注釈機能で代用すること）.

2 chasenrc ファイル

chasenrc ファイルは形態素解析プログラムに必要な様々な選択肢を定義するために用いられる．これらの定義は通常，PREFIX/etc/chasenrc に記述されるが，利用者のホームディレクトリの '.chasenrc' というファイルに記述することもできる．起動時オプションなどによって chasenrc ファイルを指定することもできる．具体的には次のような優先順位で chasenrc ファイルが読み込まれる．

1. (Unix, Windows) 起動時に -r オプションで指定されたファイル．
2. (Unix, Windows) 環境変数 CHASENRC で指定されたファイル．
3. (Windows) レジストリ HKEY_CURRENT_USER\Software\NAIST\ChaSen の chasenrc に設定してある chasenrc
4. (Unix) 利用者のホームディレクトリにある .chasen2rc．
5. (Unix) 利用者のホームディレクトリにある .chasenrc．
6. (Unix) PREFIX/etc/chasenrc(デフォルトではインストールされない)．

設定項目一覧を以下に示す．このうち，「DADIC」，「未知語品詞」，「品詞コスト」は必ず指定しなければならない．

1. **文法ファイルのディレクトリ**

   文法ファイル (grammar.cha，ctypes.cha，cforms.cha，connect.cha) が存在するディレクトリを指定する．

   ```
   (文法ファイル /usr/local/lib/chasen/ipadic/dic)
   ```

   「文法ファイル」は省略することができ，その場合 chasenrc ファイルがあるディレクトリと同じディレクトリを指定したとみなされる．茶筌に付属の辞書 ipadic1.01 以降の chasenrc ファイルでは「文法ファイル」は省略されている．

2. **システム辞書**

   ダブル配列辞書 (chadic.{da,lex,dat}) を，ファイル名から末尾の拡張子を除いたものを記述することによって指定する．複数組みの辞書を指定することもできる．また，相対パス，つまり "/" で始まらないパスを記述すると，文法ファイルと同じディレクトリを指定したとみなされる．例えば以下のように指定する．

   ```
   (DADIC chadic
          /home/rikyu/mydic/chadic)
   ```

   上の記述では，以下の二組の辞書ファイルが読み込まれる．

   (a) 文法ファイルと同じディレクトリにある chadic.{da,lex,dat}

   (b) /home/rikyu/mydic/ にある chadic.{da,lex,dat}

   辞書引きに際しては，これらの辞書の両方が用いられる[2]．

[2] 一組の辞書には同一の形態素の登録は行なわれないが，複数の辞書に同じ形態素が登録されている場合はあり得る．このような場合は，同じ形態素が複数得られることになる．

Darts によるダブル配列辞書を使うために「DADIC」を指定する．

```
(DADIC chadic)
```

上の記述では，文法ファイルと同じディレクトリにある chadic.da, chadic.lex, chadic.dat が読み込まれる．

使用する辞書の最大数は，32 個に設定されている．

3. 未知語の品詞

未知語が発見された時に，その語をどのような品詞として連接規則を適用するかを指示する．複数の品詞を指定した時は，それぞれの品詞について連接規則が適用される．

```
(未知語品詞 (名詞 サ変接続))                ; 1 個の品詞を指定
(未知語品詞 (名詞 サ変接続) (名詞 一般))    ; 複数の品詞を指定
```

4. 品詞のコスト

形態素解析プログラムでは，解析結果の優先情報をコストとして計算している．解析に曖昧性がある場合は，コストの総計が低いものを優先することになっている.「品詞コスト」では，各品詞のコストの倍率と,「未知語」についてのコストを定義する．コストは正の整数値をとる．

```
(品詞コスト
        ((*)                1)
        ((未知語)           500)
        ((名詞)             2)
        ((名詞 固有名詞)    3)
)
```

同じ品詞に対してコストの定義が複数回指定されている場合は，後のものが優先される．上の記述では,「名詞」の形態素のコストは基本的には 2 倍になるが,「名詞-固有名詞」以下に細分類される名詞だけは形態素のコストが 3 倍になる．また，先頭の '(*)' の指定により，ここで明示的に定義されていない形態素のコストはすべて 1 倍 (そのままのコスト値) となる．未知語の形態素のコスト値はすべて 500 になる．

5. 連接コストと形態素コストの相対的な重みの定義

形態素解析におけるコストの計算は形態素のコストと連接のコストの総計として計算される．これら二種類のコストに異なる重みを掛けたい場合には，それを指定することができる．解析結果のコストはそれぞれのコストにここで指定された重みを乗じた値の総計として計算される．省略した場合の重みは 1 である．

```
(連接コスト重み 1)      ; デフォルト値
(形態素コスト重み 1)    ; デフォルト値
```

6. コスト幅

形態素解析の過程において，常にコストが最低の結果を出すのではなく，ある程度のコスト幅を許容したい場合がある．この許容幅を指定することができる．コスト幅におさまるすべての解を出力するには -m オプションや-p オプションを使う．

```
(コスト幅 0)    ; デフォルト値
```

コスト幅は-w オプションでも指定することができる．その場合，-w オプションで指定したものが優先される．

7. 未定義連接コストの定義

連接規則ファイルに連接規則が定義されていない形態素間の連接コストを指定する．未定義連接コストを設定しないか，あるいは 0 を指定すると，連接規則が定義されていない形態素どうしは決して連接しないという意味になる．デフォルトは 0.

```
(未定義連接コスト 500)
```

8. 出力フォーマット

出力フォーマットを指定することにより，解析結果の出力形式を変えることができる．

```
(出力フォーマット "%m\t%y\t%P-\n")
```

出力フォーマットは-F オプションでも指定することができる．その場合，-F オプションで指定したものが優先される．詳しくは 1.4 節を参照のこと．

9. **BOS 文字列**

解析結果の文頭に表示する文字列を指定する．“%S” を使うと解析文全体を表示できる．デフォルトは空文字列 (何も表示しない).

```
(BOS 文字列 "解析文: [%S]\n")
```

10. **EOS 文字列**

解析結果の文末に表示する文字列を指定する．“%S” を使うと解析文全体を表示できる．デフォルトは “EOS\n”.

```
(EOS 文字列 "文末\n")
```

11. 空白品詞

茶筌は，半角の空白文字 (ASCII コード 32) とタブ (ASCII コード 9) を空白とみなし，これらを無視して解析する．通常は，解析結果に空白の情報を出力しないが,「空白品詞」を設定することにより，空白についての情報を出力するようになる．例えば，以下のように設定すると，空白を「記号-空白」として出力する．

```
(空白品詞 (記号 空白))
```

なお，出力フォーマットを “%m” に設定して，空白品詞を指定する (品詞は何でもよい) と，解析文と全く同じ出力が得られることになる．

12. 注釈

ある文字列で始まりある文字列で終わる文字列を注釈のように扱い，その文字列の部分を無視して解析させることができる．解析結果には，その文字列が一つの形態素として出力される．

chasenrc ファイルには，開始文字列，終了文字列からなるリストと出力時の品詞名あるいはフォーマット文字列を記述する．終了文字列は省略することができ，その場合，開始文字列と一致する文字列自身を注釈として扱う．また，出力時の品詞名あるいはフォーマット文字列を省略するとその形態素についての情報を全く出力しなくなる．

```
(注釈 (("<" ">") "%m\n")
      (("「") (記号 一般))
      (("」") (記号 一般))
      (("\"" "\"") (名詞 引用文字列))
      (("［" "］"))
      )
```

例えば，上のように記述すると，以下のように解析，出力される．

- <img src="cha.gif"> のように “<” で始まり “>” で終わる文字列をそのまま出力する．
- “「” あるいは “」” を「記号-一般」として出力する．
- “"hello(again)"” のようにダブルクォーテーションで囲まれた文字列を「名詞-引用文字列」として出力する．
- “［ちゃせん］” のように “［” で始まり “］” で終わる文字列を無視して解析し，解析結果にはその文字列の情報を出力しない．

13. 連結品詞

ある品詞の形態素が連続して出現したときに，一つの形態素として連結して出力させるときに使用する．

```
(連結品詞 ((複合名詞) (名詞) (接頭詞 名詞接続) (接頭詞 数接続))
  ((記号)))
```

例えば，上の記述では以下のように品詞を連結する．

(a) 連続した「名詞」「接頭詞-名詞接続」「接頭詞-数接続」を連結し「複合名詞」として表示する．なお,「複合名詞」は品詞定義ファイル grammar.cha に記述しておく必要がある．

(b) 連続した「記号」を連結し,「記号」として表示する．

14. 複合語出力

形態素辞書ファイル (.dic) 内で定義した複合語について，複合語全体の形態素情報を出力する (“複合語”) か，複合語を構成する各単語の形態素情報を出力する (“構成語”) かを選択することができる．デフォルトは “複合語”．

```
(複合語出力 "複合語")
```

なお，複合語出力については -Oc, -Os オプションによっても制御することができる．

15. 区切り文字

    -j オプションをつけた時の文の区切り文字を並べ，一つの文字列にしたものを指定する (1.3 節参照)．区切り文字には全角文字，半角文字の両方を使用することができる．例えば

```
(区切り文字 "。.、，!? .,!? ")
```

    と定義すると，全角文字の「。.、，!?」のいずれか，または半角文字の “.,!?␣” (空白文字が入っていることに注意) のいずれかの文字が文の区切りとなる．

16. 文字コード指定

    あらかじめ形態素辞書ファイルなどの文字コードを変更してコンパイルしておくことにより，他の文字コードのファイルも解析することができる．その際、chasenrc に以下のように記述して文字コードを指定することができる．例えば

```
(ENCODE "w")
```

    と定義すると，文字コードが UTF-8 であるファイルを入力とする．指定できる文字コードは e: EUC-JP, s:Shift_JIS, w:UTF-8, u:UTF-8, a:ISO-8859-1.

3 茶筌ライブラリ

茶筌ライブラリ libchasen.a, libchasen.so を利用することで，茶筌のモジュールを他のプログラムに組み込むことができる．ヘッダファイルとして chasen.h をインクルードする．利用できるライブラリ関数・変数は以下の通りである．

```
#include <chasen.h>

int chasen_getopt_argv(char **argv, FILE *fp);

extern int Cha_optind;
```

茶筌にオプションを渡す．もし茶筌の初期化が行われていなければ，初期化を行ってからオプションの設定を行う．デフォルトのオプションのままでよければ，この関数を呼び出さずに以下の解析関数を呼び出してもかまわない．

argv にはコマンドラインオプションとして NULL で終わる文字列の配列を指定する．ただし argv[0] はプログラムのファイル名である．オプション指定に誤りがあった場合，ファイル・ポインタ fp にエラーメッセージを出力する．fp が NULL のときは何も出力しない．

オプション指定に誤りがなければ 0 を，誤りがあれば 1 を返す．

外部変数 Cha_optind には処理したオプション (argv[0] を含む) の数が格納される．

以下に使用例を示す．chawan というプログラムにおいて，‘-r /home/rikyu/chasenrc.proj -j’ というオプションを茶筌に渡している．この関数の実行後 Cha_optind には 4 が代入される．

```
char *option[] = {"chawan", "-r", "/home/rikyu/.chasenrc.proj", "-j", NULL};
chasen_getopt_argv(option, stderr);
```

```
#include <chasen.h>
```

```
int chasen_fparse(FILE *fp_in, *fp_out);

int chasen_sparse(char *str_in, FILE *fp_out);

char *chasen_fparse_tostr(FILE *fp_in);

char *chasen_sparse_tostr(char *str_in);
```

もし茶筌の初期化が行われていなければ，初期化を行ってから形態素解析を行う．入力と出力がファイルであるか文字列であるかによって，4 つの関数がある．

chasen_fparse(), chasen_fparse_tostr() はファイル・ポインタ fp_in から読み込んだ文字列を解析する．chasen_getopt_argv() で -j オプションを指定したときは，句点などを文の区切りとして解析を行う．

chasen_sparse(), chasen_sparse_tostr() は 文字列 str_in を解析する．

chasen_fparse(), chasen_sparse() は解析結果をファイル・ポインタ fp_out に出力する．返り値は 0 を返す．

chasen_fparse_tostr(), chasen_sparse_tostr() は解析結果を茶筌内部で確保したメモリ領域に格納し，そのポインタを返す．この領域は，次に chasen_fparse_tostr(), chasen_sparse_tostr() を呼び出すまで有効である．

4 他のシステムからの利用

4.1 Perl からの使用

perl/ChaSen.pm を使うことにより，perl から茶筌を利用できる．インストール方法，使用方法については perl/README を参照のこと．

参考文献

[1] 益岡隆志，田窪行則：『基礎日本語文法 –改訂版–』くろしお出版，1992.

[2] 妙木裕，松本裕治，長尾眞：「汎用日本語辞書および形態素解析システム」情報処理学会第 42 回全国大会予稿集，1991.

[3] 松本裕治，黒橋禎夫，宇津呂武仁，妙木裕，長尾真：「日本語形態素解析システム JUMAN 使用説明書 version 2.0」，NAIST Technical Report, NAIST-IS-TR94025，1994.

[4] 山下達雄，松本裕治:「形態素解析視覚化システム ViJUMAN version 1.0 使用説明書」，NAIST Technical Report, NAIST-IS-TR96005，1996.

[5] 山下達雄，松本裕治：「形態素解析結果の視覚化システム ViJUMAN とその学習機能」，情報処理学会研究報告 96-NL-115, pp.29-34，September 1996.

[6] 平野 善隆：「用言の活用を考慮した韓国語品詞体系の提案とそれを用いた韓国語形態素解析」，奈良先端科学技術大学院大学修士論文, NAIST-IS-MT9551092，March 1997.

[7] 山下達雄：「規則と確率モデルの統合による形態素解析」，奈良先端科学技術大学院大学修士論文, NAIST-IS-MT9551119，March 1997.

[8] 山下達雄，松本裕治：「コスト最小法と確率モデルの統合による形態素解析」，情報処理学会研究報告 96-NL-119，May 1997.

[9] 北内 啓，山下 達雄，松本 裕治：「日本語形態素解析システムへの可変長連接規則の実装」, 言語処理学会第三回年次大会論文集, pp.437-440, 1997.

[10] 「研究開発用知的資源タグ付きテキストコーパス報告書」平成９年度, テキストサブワーキンググループ, 技術研究組合 新情報処理開発機構, 1998.

[11] 松田 寛：「品詞タグ付きコーパス作成支援環境の構築」，奈良先端科学技術大学院大学修士論文, NAIST-IS-MT9851103，March 1999.

[12] 北内 啓, 宇津呂 武仁, 松本 裕治：「誤り駆動型の素性選択による日本語形態素解析の確率モデル学習」, 情報処理学会論文誌 Vol. 40, No. 5, p.p.2325-2337, May 1999.

[13] 松田 寛, 桐山 和久, 山田 悟史, 吉野 圭一, 松本裕治：「部分形態素解析を用いたコーパスの品詞体系変換」, 情報処理学会研究報告 99-NL-134, p.p.23-30, Nov. 1999.

[14] Masayuki Asahara：Extended Statistical Model for Morphological Analysis, 奈良先端科学技術大学院大学修士論文, NAIST-IS-MT9851001，March 2000.

[15] 松田 寛, 松本 裕治：「品詞タグ付きコーパス作成支援 GUI ツール VisualMorphs」, 情報処理学会研究報告 2000-NL-137, p.98, June, 2000.

[16] 浅原 正幸, 松本 裕治：「統計的日本語形態素解析に対する拡張 HMM モデル」, 情報処理学会研究報告 2000-NL-137, p.p.39-46, June, 2000.

[17] Masayuki Asahara, Yuji Matsumoto：Extended Models and Tools for High-performance Part-of-Speech Tagger, Proceedings of COLING 2000, July, 2000.

[18] 浅原 正幸, 松本 裕治：「誤り駆動による統計的品詞タグづけモデルの拡張」, 情報処理学会研究報告 2000-NL-139, p.p.25-32, Sep. 2000.

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

[20] 伝 康晴，浅原 正幸：「リレーショナル・データベースによる統合的言語資源管理環境」, 第１回「話し言葉の科学と工学」ワークショップ, Feb. 2001.

[21] 伝 康晴，宇津呂 武仁，山田 篤，浅原 正幸，松本 裕治：「話し言葉研究に適した電子化辞書の設計」, 第２回「話し言葉の科学と工学」ワークショップ, pp. 39-46, Feb. 2002.

[22] 浅原 正幸，松本 裕治：「形態素解析とチャンキングの組み合わせによる日本語テキスト中の未知語出現箇所同定」, 情報処理学会研究報告, 自然言語処理研究会, SIGNL-154, pp.47-54, 2003

[23] 中川 哲治，工藤 拓，松本 裕治 ：「Support Vector Machine を用いた形態素解析と修正学習法の提案」, 情報処理学会論文誌, Vol.44, No.5, pp.1354-1367, May 2003.

[24] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto：”Appliying Conditional Random Fields to Japanese Morphological Analysis”, EMNLP-2004, 2004.

[25] 松本裕治，高岡一馬，浅原正幸，工藤拓 ：「茶筌と南瓜による日本語解析–構文情報を用いた文の役割分類」 人工知能学会誌, Vol.19, No.3, pp.334-339, 2004.

[26] Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto： ”Chinese Word Segmentation by Classification of Characters”, International Journal of Computational Linguistics and Chinese Language Processing , Vol.10, No.3, pp.381-396, September, 2005.

[27] Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto： ”Training Multi-Classifiers for Chinese Unknown Word Detection” , Journal of Chinese Language and Computing, Vol.15, No.1, pp.1-12, 2005.

[28] 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」, 言語処理学会第 11 回年次大会発表論文集, pp.245-248, 2005.

[29] 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」, 言語処理学会第 11 回年次大会発表論文集, pp.604-607, 2005.

[30] 工藤 拓：「形態素周辺確率を用いた分かち書きの一般化とその応用」, 言語処理学会第 11 回年次大会発表論文集, 2005.

[31] Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto：”Machine Learning-based Methods to Chinese Unknown Word Detection and POS Tag Guessing”, Journal of Chinese Language and Computing, Vol.16, No.4, pp.185-206, 2006.

[32] 東藍，浅原正幸，松本裕治：「条件付確率場による日本語未知語処理」 情報処理学会研究報告, 自然言語処理研究会, SIGNL-173, pp.67-74, 2006.

[33] 東藍，工藤拓，浅原正幸，松本裕治：「日本語未知語処理のための大規模未解析データの利用法」情報処理学会研究報告, 自然言語処理研究会, SIGNL-179, 2007.

付録

A　著作権および使用条件について

茶筌システムは，広く自然言語処理研究に資するため無償のソフトウェアとして開発されたものである．茶筌の著作権は，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座 (松本研究室) が保持する．本ソフトウェアの使用，改変，再配布については，特に制限を課すことはしないが，再配布については，次の事項を条件とする．

- 再配布されるソフトウェアに，著作権に関する本節の記述と使用説明書の表紙裏のページの著作権に関する但し書きを必ず含むこと．

なお，本ソフトウェアの著作権者である奈良先端科学技術大学院大学は，原形あるいは改変された形で配布された本ソフトウェアに関連して生じる一切の損失に対して保証の責を負わないこととする．

また，上に述べた著作権は茶筌システム本体についてのものであり，ipadic をはじめとする他の辞書については，各辞書についての著作権条項があるためそちらを参照すること．

B　更新履歴

B.1　茶筌 2.3.3 から 茶筌 2.4.0 への変更点

- 制約つき解析の実装
- Windows 版パッケージの再構成
- chasenrc による文字コード指定
- UTF-8 の指定文字に 'u' を追加

B.2　茶筌 2.3.2 から 茶筌 2.3.3 への変更点

- 辞書に読み，発音の情報がないとき，%y，%a などは空白を表示するよう変更．
- (MinGW 版のみ) chasenrc と文法ファイルのパスをレジストリで指定できるようにした．

B.3　茶筌 2.3.1 から 茶筌 2.3.2 への変更点

- 辞書検索の高速化
- 活用形指定辞書のサポート

B.4　茶筌 2.3.0 から 茶筌 2.3.1 への変更点

- PATDIC, SUFDIC の廃止
- -i オプション (文字コード選択) の導入
- UTF-8 対応

B.5 茶筌 2.2 から 茶筌 2.3 への変更点

- ダブル配列ライブラリ Darts を用いた辞書の実装
- サーバ・クライアントモードの廃止
- コマンドインタプリタの廃止
- `cforms.cha` で，基本形として定義する文字列を変更できるようにした．

```
(基本形　基本形-一般)
```

B.6 茶筌 2.0 から 茶筌 2.2 への拡張点

- 辞書とシステムの分離
  他言語の辞書整備により，辞書とシステムを分離した．chasenrc は辞書側が持ち，システムインストール時にはインストールさい．辞書インストール時に `PREFIX/etc` 以下に `chasenrc` をインストールする必要がある．
- autoconf, automake, libtool 化
  `./configure` で，自動的に環境を読み込み設定できるようにした．これに伴い，各辞書をコンパイルする際に必要になる情報を出力するプログラム `chasen-config` を導入した．

B.7 JUMAN 2.0 から 茶筌 2.0 への拡張点

茶筌 2.0 では品詞体系や連接規則の機能などを拡張した．この機能拡張版を v-gram 版，従来のバージョンを bi-gram 版と呼ぶ．v-gram 版は bi-gram 版と文法ファイルの形式が異なっているため，辞書に互換性がない．ただし，`mkchadic/convdic` を実行することにより，bi-gram 版の辞書を v-gram 版の辞書に変換することができる．

`convdic` は bi-gram 版の辞書があるディレクトリ上で，v-gram 版の辞書を格納するディレクトリを引数として実行する．例えば以下のように実行すると，bi-gram 版の辞書がある `dic` というディレクトリと同じ階層に `dic2` というディレクトリが作成され，その中に v-gram 版の辞書が格納される．なお，`convdic` 実行後，茶筌に付属の `dic/Makefile` を v-gram 版の辞書があるディレクトリ (下の例では `dic2`) にコピーする必要がある．また，chasenrc ファイルも用意する．

```
% cd dic
% ../mkchadic/convdic ../dic2
% cp Makefile ../dic2
```

茶筌 2.0 ではデフォルトで v-gram 版がコンパイルされる．‘`make bigram`’ を実行すれば bi-gram 版の実行ファイルが作成され，bi-gram 版の辞書を利用することができる．

v-gram 版は bi-gram 版と比べ，以下のような拡張機能や変更点がある．

1. 品詞を 2 階層から多階層に拡張した．
2. 連接規則を bi-gram の固定長から variable-gram(可変長) に拡張した．すなわち，連接する 2 個の単語 (あるいは品詞) の連接コストだけではなく，3 個以上の任意の長さの単語 (品詞) 列に対して単語 (品詞) の連接コストを記述できる．

3. *.dic で「発音」という属性を使える．出力フォーマットの%a, %A で表示できる．また，cforms.cha で発音の語尾を定義できる．

4. *.dic で「base」という属性を使える．見出し語の基本形などを表示する際，活用を持っていればその基本形を，活用がなく base を持っていれば base を表示する．英語の辞書などで使用する．

5. chasenrc ファイルの「連結品詞」の機能を拡張し，複数の種類の品詞を別々に連結できるようにした．

6. 空行に対しても “EOS”(正確には BOS 文字列と EOS 文字列) を表示する．つまり，“EOS” の個数が入力文の行数と一致する．

7. 解析結果のデフォルトの出力形式 (-f) で，見出し語などの直後の区切りがスペースではなくタブになった．

8. 辞書に登録されていない単語の品詞表示を「未定義語」から「未知語」に変更した．

9. 形態素辞書ファイル *.dic で単語のコスト値が省略されている場合，bi-gram 版ではコスト値が 10 となるのに対し，v-gram 版では*.dic 中の「デフォルト品詞コスト」で指定されたコスト値 (指定されていない場合は 65535) が用いられる．

10. bi-gram 版では形態素コストと連接コストを内部で 10 倍しているが，v-gram 版ではそのままの値を用いる．また，bi-gram 版では形態素コストの範囲が 0　6553.5(茶筌 1.51 以前は 0　25.5) であるが，v-gram 版では 0　65535 である．

11. 連接コスト 0 を「確率 1 で連接する」という意味に，-1 を「連接しない」という意味に変更した．また，連接コストの範囲を -1　32767 に変更した．

12. 文節区切りの機能を持つ，長さ 0 の品詞が使える．品詞定義ファイルで品詞名の後ろに ‘/’ をつけると文節区切りとして機能する．

B.8　茶筌 1.5 から 茶筌 2.0 への拡張点

ここでは v-gram 版，bi-gram 版に共通する拡張点をあげる．

1. chasenrc の「文法ファイル」を省略できるようにした.「PATDIC」「SUFDIC」が ’/’ で始まっていない場合は,「文法ファイル」のディレクトリからの相対パスとみなすようにした．

2. 辞書引きに SUFARY を使えるようにすることにより，半角文字も検索できるようにした．

3. SUFARY を使って英語を解析できるようにした．

4. -D なしで -R を指定した場合は Makefile で指定した chasenrc (/usr/local/share/chasen/dic/chasenrc など) を読み込むようにした．

5. 文頭・文末で出力する文字列を設定できるようにした．

6. 未知語品詞とそのコストを複数指定できるようにした．

7. chasenrc ファイルで「空白品詞」を指定することにより，空白も解析結果に出力できるようにした．

8. chasenrc ファイルで「注釈」を指定することにより，SGML タグのような特定の文字列を空白と同様に無視して解析できるようにした．

9. -lp, -lt, -lf オプションで品詞や活用のリストを表示できるようにした．

10. -o オプションで出力ファイルを指定できるようにした．

11. 出力フォーマット "%?T/STR1/STR2/" を使えるようにした．活用があれば STR1, なければ STR2 を出力する．そのほかに %?I, %?B, %?F, %?U も使えるようにした．

12. 出力フォーマット "%rABC" を導入し，ルビを表示できるようにした．

13. chasenrc ファイルで「BOS 文字列」「EOS 文字列」を指定することにより，文頭・文末で出力する文字列を設定できるようにした．

14. BOS 文字列，EOS 文字列，出力フォーマットで，解析文全体を表示する"%S" を使えるようにした．

15. 辞書ファイルの形態素コストの範囲を今までの 0　25.5 から，bi-gram 版は 0　6553.5 に，v-gram 版は 0　65535 に変更した．

16. 連接ファイルの連接コストの範囲を 0　255 から 0　32767 に変更した．

B.9　茶筌 1.0 から 茶筌 1.5 への拡張点

1. ライブラリ化を行い，茶筌のモジュールを他のプログラムに簡単に組み込めるようにした．

2. サーバ化を行い，クライアントを用いて他のマシンから解析を行うことができるようにした．また，クライアントの Emacs Lisp 版インタフェースを作成した．

3. -w オプションでコスト幅を指定できるようにした．

4. chasenrc ファイルに「区切り文字」を指定することにより，jfgets() の区切り文字を設定できるようにした．半角文字を指定することも可能．また，区切り文字のデフォルトを ”.。!?” に変更した．

5. バッファを動的に確保することにより，文字列が長いときでも “Too many morphs” の警告が出ないようにした．

6. 美茶 (ViCha) 用出力オプション -v を新設した．

7. -d オプションと -b を同時に指定したときに-d の出力形式で最適解パスだけ表示できるようにした．

B.10　JUMAN 2.0 から 茶筌 1.0 への拡張点

1. 辞書検索の方法を従来の NDBM を用いて疑似的に TRIE 構造を実現する方法から，独自開発のパトリシア木を用いたものに変更した．解析に必要な辞書のサイズが約 4 分の 1 に縮小した．また，辞書のコンパイル時間が 3　40 分の 1 になった．

2. 解析システムの見直しを行ない，高速化を図った．解析速度が約 8　11 倍になった (JUMAN 2.0 との比較)．

3. 多くのプラットフォームでインストール可能になるようにコードを書き直した．また，GNU C コンパイラ (gcc) だけでなく OS 付属の C コンパイラなどでもコンパイルできるようにした．

4. 日本語 EUC だけでなく，JIS(ISO-2022-JP) の文字列も解析できるようにした．

5. 未定義連接コストの導入により，未定義語の出力を減らすことができるようになった．

6. 連結品詞を定義できるようにし，最適パスを出力する時に，その品詞の単語を一単語に連結して表示するようにした．

7. 活用語尾の読みを定義できるようにすることにより,「来る」「得る」などの読みがひらがなで表示されるようになった.

8. 入力文を改行コードで区切るのではなく，句点により区切るオプション (-j) を追加した.

9. -r オプションや環境変数 CHASENRC で chasenrc ファイルを指定できるようにした.

10. -F オプションや chasenrc ファイルの「出力フォーマット」で解析結果の出力形式を変更できるようにした.

11. 文法の見直しを行ない，品詞分類「特殊」の下の「括弧」を「括弧開」と「括弧閉」に分離した．また，同じく「特殊」の下に「空白」を定義した.「空白」は具体的には全角の空白を表す.

12. 助動詞の活用型に「助動詞べきだ型」を追加した．助動詞「べきだ」の活用を従来の「ナ形容詞」型から「助動詞べきだ型」に変更した.

13. 辞書登録語について見直し，追加削除等の修正を行なった.

C　JUMAN3.0 と 茶筌 との関係について

JUMAN 2.0 が 1994 年 7 月にリリースされて以降，京都大学長尾研究室と奈良先端大松本研究室では，それぞれ異なる方向での拡張を試みていました．京都大学では，従来の bi-gram モデルでは記述できない連接関係を記述するために連語処理や括弧の透過処理などの機能を追加し，文法ファイル，形態素辞書に大幅な修正を行なった拡張版を作成していました．奈良先端大では，今後大量の蓄積が始まると思われる日本語タグ付きコーパスから bi-gram 以上の連接規則 (単語レベルや品詞レベルの設定も含む) を自動的に学習する機能を追加するための拡張と，UNIX のハッシュデータベース NDBM に依存しない辞書の構築を考えていました．後者の拡張は UNIX 以外の OS での稼働を要求する声に対応することと辞書のコンパイル時間と検索速度の改善を目指したことによります．bi-gram 以上の連接規則に対する両者の考え方がかなり異なるため，両者の融合は見合わせることにし，いち早く完成した京都大学の拡張版が 1996 年 6 月に JUMAN3.0beta として公開されました.

奈良先端大で拡張を予定していた機能には下に示すような項目があり，茶筌 1.0 を 1997 年 2 月に公開し，以後，茶筌 1.5, 1.51, 2.0, 2.2, 2.3 を経て，茶筌 2.4 においてそのほとんどが実現されました.

1. (茶筌 1.0) 辞書システムの独自開発 (NDBM の棄却, パトリシア木の採用)

2. (茶筌 1.0) 解析システムの見直しと高速化

3. (茶筌 1.0) 未定義連接コスト，連接品詞，解析結果出力フォーマットの導入

4. (茶筌 1.0) JIS 文字列の解析

5. (茶筌 1.0) 活用語尾の読みの定義

6. (WinCha1.0) Windows への対応

7. (茶筌 1.5) ライブラリ化

8. (茶筌 1.5) サーバ化

9. (茶筌 2.0) 品詞定義の多階層化

10. (茶筌 2.0) 接続規則の可変長化

11. (茶筌 2.0) 半角文字を含む単語の辞書登録 (SUFARY を利用した辞書)

12. (茶筌 2.0) 出力フォーマットの拡充

13. (茶筌 2.0) 解析済みデータからの可変長接続コストの学習

14. (茶筌 2.4) 制約つき解析

D　形態素解析器の今後について

工藤拓氏により MeCab という形態素解析器が公開されています[3] 。茶筌は形態素解析モデルとして隠れマルコフモデルという生成モデルを用いているのに対し，MeCab は条件付確率場という識別モデルを用いています。工藤氏の論文 [24] では，新しいモデルの方が解析精度が向上したことを示しています。MeCab のその他の特色として「ソフトわかち書き」[30][4] を出力することがあげられます。現在の茶筌の枠組では，MeCab のように素性（特徴量）を自由に設計できず，このような新しい解析モデルには対応できないでおります。

辞書関連でも近年さまざまな改良が行われてきました。新しい JUMAN 辞書[5] では，日本語の基本的語彙を選別するとともに，表記ゆれ情報の整備が行われています。千葉大の伝氏のグループにより UniDic [21] と呼ばれる，自然言語処理研究者のみならず，人文系研究者にも音声処理研究者にも使いやすい辞書が近く公開されるようです。奈良先端大では，IPADIC の辞書項目を選別し，表記ゆれ情報，複合語情報を付与した日本語辞書を公開する予定です。新しい辞書では，名前を変更し，IPADIC で懸案となっていた ICOT 条項を廃止する予定です。また，奈良先端大では，権利関係の処理が終わり次第 Penn Chinese Treebank 体系の品詞が付与された中国語形態素解析用辞書を公開する予定です。中国語形態素解析用辞書公開時には，MeCab 作者の工藤さんとも協議して，ChaSen 用モデルだけでなく，MeCab 用モデルも同時に公開したいと考えております。

JUMAN, 茶筌, MeCab が解決していない問題として，未知語（辞書にない語）の問題があります。現在奈良先端大におきまして未知語の問題を解決する機械学習モデルを開発しております [32, 33]。いずれ、現在の茶筌とは異なる枠組で構成された、未知語解析モデルつきの形態素解析器を公開できればと考えております。

[3] http://mecab.sourceforge.net/
[4] http://mecab.sourceforge.net/soft.html
[5] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html