

中古和文 UniDic Windows 版パッケージ

「和文茶まめ」使用説明書

このマニュアルは現代語用 UniDic 付属のマニュアルを元にした暫定版です。
一部未整備な部分がありますが、ご了承ください。

1. はじめに	2
2. 茶まめの使い方	2
2.1. 起動	2
2.2. 解析対象テキストの設定	3
2.3. 解析前処理の設定	3
2.4. 解析器と解析オプションの設定	4
2.5. 解析後処理の設定	4
2.6. 解析結果の出力	5
3. 茶まめの出力形式	6
3.1. ファイルの文字コード	6
3.2. 表形式テキストのフィールド	6
表形式テキスト出力の例	6
3.3. 出力される XML のタグ	7
4. 茶まめが行う処理	8
4.1. 茶まめの処理の流れ	8
4.2. 解析対象の XML 文書化処理	8
5. コマンドプロンプトでの利用 (ChaSen)	9
6. FAQ (よくある質問)	10
7. 著作権・問い合わせ先	11

2012 年 4 月 23 日 小木曾 智信

1. はじめに

「和文茶まめ」(茶まめ)はUniDicを使って形態素解析を行うのを補助するためのソフトウェアです。茶まめを使うことにより、UniDicで解析する際に必要な一連の作業を、わかりやすいインターフェイス(GUI)で行うことができます。



このマニュアルでは茶まめの使い方について説明します。

※ 茶まめを使わなくてもUniDicによる解析を行うことができます。このマニュアルの「5. コマンドプロンプトでの利用」をご覧ください。

2. 茶まめの使い方

以下、実際の手順に沿って茶まめの使い方を説明します。

2.1. 起動

デスクトップに作られる「茶まめ」アイコン、または、スタートメニューの「UniDic」→「茶まめ」から起動してください。



次の画面が現れます。

この画面の上から下へと順に処理方法を指定してゆき、最後に「実行」ボタンを押すことで結果を出力します。

2.2. 解析対象テキストの設定

画面上部の「解析するテキスト」で解析対象のテキストを指定します。ラジオボタンで選んでください。選んだ方法にあわせて画面が変わります。

- テキストを入力したり貼り付けたりする場合には、「テキストエリアを解析」を選びます（起動時にはこれが選ばれています）。テキストエリア（白い部分）にテキストを入力してください。ダブルクリックすると内容がクリアされます。

- 解析対象のファイルを指定する場合には、「ファイル(XML/TXT)を解析」を選びます。その後「参照」ボタンを押してファイルを指定してください。指定できるファイルはテキストファイル、XML ファイル、HTML ファイルです。HTML ファイルはタグを除去してテキストとして解析します。テキストファイルの文字コードは自動判別します (Shift_JIS, EUC-JP, JIS(ISO-2002-JP), UTF-8, UTF-16)。

※ワイルドカードを使って複数のファイルを指定し、一度に処理することができます。

例：C:¥DATA¥*.txt → C:¥DATA の中の全ての txt ファイルを解析します

- インターネット上からダウンロードして解析する場合には、「URL から取得して解析」を選んで URL を指定してください。

2.3. 解析前処理の設定

必要に応じてチェックボックスをチェック (☑) して解析前処理を指定します。

- 「☐ 踊り字を展開」をチェックすると、解析前に繰り返し記号「ゝゞゝゞ」を対応する仮名に、くの字点代用記号「／＼」「～～」を「/ \」に（「／＼」「～～」を「/ \」に）変換します。
- 「☐ カタカナひらがな反転」をチェックすると、解析前にカタカナをひらがなに、ひらがなをカタカナに変換します。近代 UniDic は原則として漢字ひらがな交じり文に対応しています。漢字カタカナ交じり文を解析する場合は、これをチェックしてください。

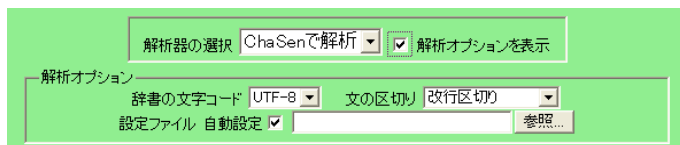
- 「☐ 半角英数字を全角に変換」をチェックすると、解析前に文字を変換します。UniDicは原則として全角文字にしか対応していません。半角文字列が入ったテキストを解析する場合にはこれをチェックしてください。
- 「☐ 数字処理 (NumTrans)」をチェックすると解析前に数字を解析しやすい形に変換します。詳細については NumTrans のマニュアルをご覧ください。

2.4. 解析器と解析オプションの設定

「解析器の選択」で「~~ChaSen~~」「MeCab」「解析しない」から選択できます。

- ~~「ChaSen」は解析器として ChaSen を使って解析します。この場合、解析後処理を選ぶことができます。~~
- 「MeCab」は解析器として MeCab を使って解析します。この場合、解析後処理を選ぶことはできません（今後対応する予定です）。
- 「解析しない」は解析せず、解析直前の XML ファイルをそのまま出力します。

「☐ 解析オプションを表示」をチェックすると指定画面が表示され解析オプションを変更することができます。通常はそのままにしておいてください。



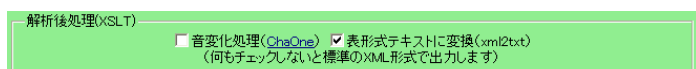
- ~~ChaSen の設定ファイル (chasenre) を自分で設定したい場合には、ここで「☒ 自動設定」のチェックを外し、ファイルを指定します。~~
~~「自動設定」がチェックされていると、設定ファイルとして同梱の chasenre が (ChaOne による処理を行う場合には chasenre-chaone が) 使われます。~~
- ~~「文の区切り」は ChaSen の -j オプションに相当します。~~
- 「☐ 解析と解析後処理を行わない」をチェックすると、解析対象となる XML ファイルを解析しないでそのまま出力します。茶まめによる処理過程を確認したい場合にお使いください。

2.5. 解析後処理の設定

~~必要に応じてチェックボックスをチェック (☒) して解析後処理を指定します。~~

~~解析器に ChaSen を利用した場合のみ、各種のオプションが利用できます。~~

~~解析器に MeCab を利用した場合は、常に表形式テキストで出力されます。~~



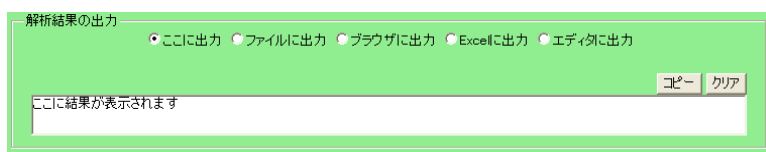
- ~~「☐ 音変化処理 (ChaOne)」をチェックすると、ChaOne による音変化処理を行います。処理内容については ChaOne のマニュアルをご覧ください。~~
- ~~「☐ 表形式テキストに変換 (xml2txt)」をチェックすると、解析結果の XML を表形~~

~~式のタブ区切りテキストに変換します（起動時にチェックされています）。~~

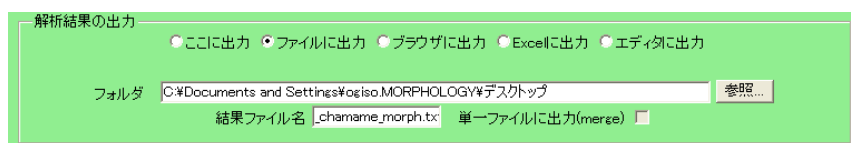
- ~~• 大きなファイルは ChaOne 処理に時間がかかる場合があります。また、前処理・後処理でたくさんの処理を行う（チェック☑をたくさん付ける）ほど時間がかかります。~~
- ~~• 何もチェックしないと UniDic 標準の XML 形式（3.3 参照）で出力します。~~

2.6. 解析結果の出力

画面下部の「解析結果の出力」で解析結果の出力方法を指定します。ラジオボタンで選んでください。選んだ方法にあわせて画面が変わります。



- 「ここに出力」を選択すると、茶まめのテキストエリアに結果を表示します。短い文章を解析してみるときに使ってください。
- 「ファイルに出力」を選択すると、出力先の指定画面が現れ、指定したファイルに解析結果が出力されます。比較的大きなファイルを処理する場合にはファイルに出力するようにしてください。



ワイルドカードを使って入力ファイルを指定した場合にはこのボタンしか選択できません。また、この場合に限って「単一ファイルに出力」チェックボックスをチェックすることで結果を一つのファイルにまとめることができます。

- 「ブラウザに出力」を選択すると、解析結果を Web ブラウザ（Internet Explorer）に出力します。解析結果の XML ファイルを閲覧する場合にお使いください。
- 「Excel に出力」を選択すると、解析結果を Microsoft Excel（表計算ソフト）に出力します。解析結果を表形式テキストに変換して表として使いたい場合にお使いください。
- 「エディタに出力」を選択すると、解析結果をテキストエディタに出力します。選択されるエディタは、Internet Explorer の「ソースの表示」用に指定されているエディタです。

3. 茶まめの出力形式

茶まめが出力するファイルの形式について説明します。

3.1. ファイルの文字コード

解析結果は辞書と同じ UTF-8 で出力されます。

3.2. 表形式テキストのフィールド

標準状態で出力される表形式テキスト (xml2txt によって変換したタブ区切りテキスト) のフィールドは左から順に次の通りです。

出典 文境界 書字形 発音形 語彙素読み 語彙素 品詞 活用型 活用形 語形 語種

※文境界は B が文頭、I がそれ以外。

表形式テキスト出力の例

出典	文境界	書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
chamame	B	此处	ココ	ココ	此处	代名詞			ココ	和
chamame	I	に	ニ	ニ	に	助詞・格助詞			ニ	和
chamame	I	解析	カイセキ	カイセキ	解析	名詞・普通名詞 -サ変可能			カ イ セ キ	漢
chamame	I	せ	セ	スル	為る	動詞・一般	文語サ行変格	未然形・一般	ス	和
chamame	I	ん	ン	ム	む	助動詞	文語助動詞・ム	終止形・撥音便	ム	和
chamame	I	と	ト	ト	と	助詞・格助詞			ト	和
chamame	I	する	スル	スル	為る	動詞・一般	文語サ行変格	連体形・一般	ス	和
chamame	I	文章	ブンショウ	ブンショウ	文章	名詞・普通名詞 -一般			ブ ン シ ョウ	漢
chamame	I	を	オ	ヲ	を	助詞・格助詞			ヲ	和
chamame	I	入力	ニューリョク	ニュウリョク	入力	名詞・普通名詞 -サ変可能			ニ ュ ウ リョク	漢
chamame	I	す	ス	スル	為る	動詞・一般	文語サ行変格	終止形・一般	ス	和
chamame	I	べし	ベシ	ベシ	べし	助動詞	文語助動詞・ベシ	終止形・一般	ベシ	和
chamame	I	。			。	補助記号・句点				記号

3.3. 出力される XML のタグ

標準の XML 形式による出力では、茶まめの処理により次のようなタグ・属性が付与されます。

すべて名前空間 URI として「<http://www.unidic.org/chasen/ns/structure/1.0>」を使用します。名前空間接頭辞として「cha:」を使用しています。

要素・属性	説明
cha:D	テキスト・HTML ファイルを処理した場合に、XML 文書のルートタグとして使われます。
cha:S	W1 を含む込む文レベルの要素です。テキスト・HTML ファイルの場合には ChaSen の EOS / BOS の範囲に相当します。XML ファイルの場合には解析後に cha:W1 の親として挿入します。
cha:W1	ChaSen によって解析された短単位のタグです。元のテキストを内容として、次の属性が付けられます (chasenrc による解析の場合)。 <ul style="list-style-type: none">• orth : 書字形• kana : 仮名形• pron : 発音形• pos : 品詞• cType : 活用型• cForm : 活用形• orthBase : 書字形基本形• kanaBase : 仮名形基本形• pronBase : 発音形基本形• lForm : 語彙素読み• lemma : 語彙素表記• form : 語形• aType : アクセント型• aConType : アクセント結合型• goshu : 語種
@cha:src	XML 文書のルートタグに付けられる属性で、解析対象の出典を表します。入力がテキストエリアの場合は「chamame」、ファイルの場合はファイル名、URL 指定の場合は URL が値となります。 (グローバル属性で名前空間接頭辞がつきます)

NumTrans, ChaOne によって処理を行った場合にはこれ以外のタグも使われます。それらのタグについては NumTrans, ChaOne のマニュアルをご覧ください。

4. 茶まめが行う処理

※ この項目は茶まめが内部的に行っている処理について説明したものです。一般的な利用を行うだけであれば読み飛ばしていただいてもかまいません。

4.1. 茶まめの処理の流れ

茶まめは次のような流れで処理を行います。

- 1 入力
- 2 解析対象の XML 文書化
- 3 解析前処理 XSLT
 - 3.1 踊り字の変換 (オプション)
 - 3.2 ひらがなカタカナの変換 (オプション)
 - 3.3 半角文字の変換 (オプション)
 - 3.4 NumTrans (オプション)
- 4 解析器による解析
- 5 解析後処理 XSLT (ChaSen 利用時のみ)
 - 5.1 S タグ挿入
 - 5.2 ChaOne (オプション)
 - 5.3 表形式テキストへの変換 (オプション)
- 6 出力

4.2. 解析対象の XML 文書化処理

unidic 関連ツールは入出力を XML で行います。そのため、テキストエリアに入力された文字列なども XML 文書に変換してから解析前処理へ進みます。この XML 文書化の処理方法は、入力の種類別に表のようになっています。

入力の種類		処理方法
テキストエリア		テキストファイルと同じ処理を行う。
ファイル	テキスト	ルートを cha:D タグで囲み、「。」を区切りとして cha:S タグを挿入して XML 文書にする。特殊文字 (<>&) は全角文字に置き換える。文字コードは次の各種エンコーディングを自動判別する。 (Shift_JIS, EUC-JP, JIS(ISO-2002-JP), UTF-8, UTF-16)
	HMTL	タグを除去 (br, tr, li, p などは改行を挿入) してテキスト化したのち、テキストファイルと同様に XML 文書にする。実体参照や特殊文字は全角文字や＝に置き換える。
	XML	ルートにネームスペース宣言、cha:src 属性を付加する。文書型宣言付きの場合は削除する。 (解析後、Inserts.xsl により cha:S タグを挿入する)
URL 指定		ファイルを取得して保存後、種類別に上と同じ処理を行う。 (ただし XML ファイルとして処理するのは拡張子 XML, RDF のファイルのみ)

5. コマンドプロンプトでの利用 (ChaSen)

きわめて大きなファイルを解析する場合や、独自の処理を行いたい場合などには、コマンドプロンプト上で解析を行ってください。

ChaSen の解析用辞書として UniDic (UniDic-chasen) を使うには、付属の chasenrc を利用してください (インストール先の dic フォルダの中に入っています)。ChaSen の -r オプションなどでこのファイルを指定することにより、UniDic を使った解析が行えます。

本パッケージ付属の辞書は文字コードが UTF-8 です。 -i オプションで w を指定してください。その他、ChaSen のコマンドラインオプションの詳細については、ChaSen の説明書をご覧ください。

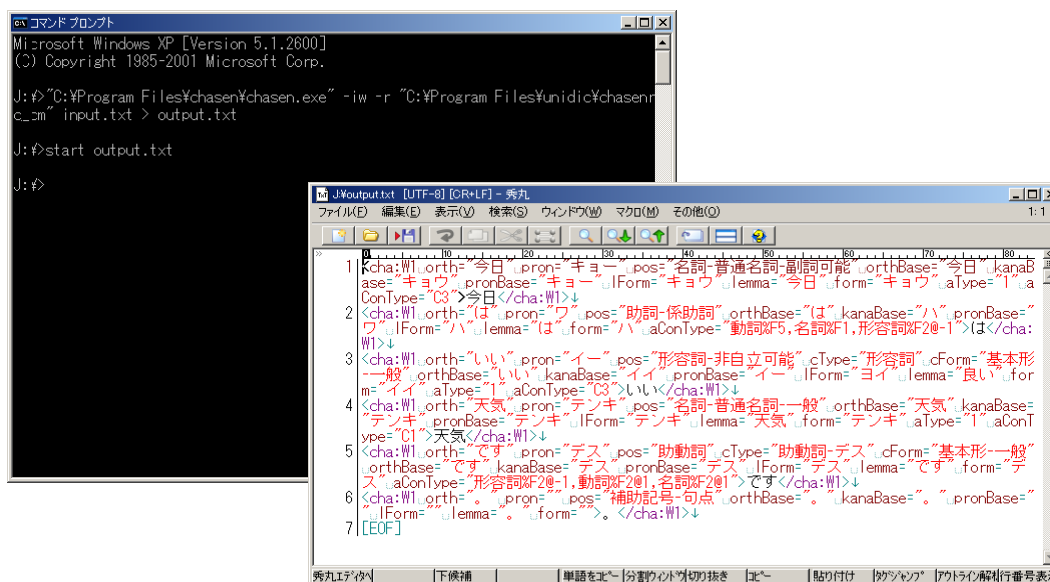
インストール直後の状態では、コマンドプロンプトで次のように打ち込むことで UniDic による解析が可能です。

"C:¥Program Files¥chasen¥chasen.exe"

-i w -r "C:¥Program Files¥unidic¥dic¥chasenrc" 《入力ファイル》 > 《出力ファイル》

※ 入力ファイルの文字コードは UTF-8 にしておく必要があります。またインストール先を変更した場合にはそれに合わせてください。

※ 実際に利用する場合には ChaSen のインストールパスを環境変数 Path に追加したり、付属の chasenrc を標準の設定ファイルにしたりして環境を整備してください。 chasenrc の指定方法など詳しくは ChaSen の説明書をご覧ください。



※ XSLT による変換をコマンドプロンプト上で行いたい場合には、msxsl.exe*などのコマンドラインツールをダウンロードして使用してください。

* <http://www.microsoft.com/downloads/details.aspx?familyid=2FB55371-C94E-4373-B0E9-DB4816552E41>

6. FAQ（よくある質問）

Q：大きなテキストを解析したいのですが、どのくらいのサイズまで解析できますか。

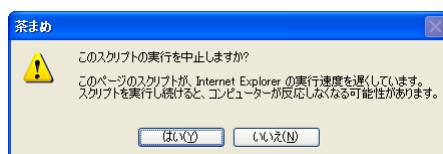
A：解析器（ChaSen・MeCab）自体は数百 MB 程度のファイルでも解析することができます。しかし XSLT による前処理・後処理を行う場合には多くのメモリを必要としますので、数 MB 程度のテキストでも変換できない場合があります。

茶まめでは XSLT を多用するため、標準で 5MB 以上のファイルは処理しないようにしています。また、一つの入力ファイルの解析結果が 20MB を越えると XSLT の後処理をパスするようになっています。

茶まめを使って大きなデータを解析する場合には、小さめのファイルに分割して処理してください。茶まめでは、ワイルドカードを使って複数のファイルをまとめて解析することができるようになっています。また、「単一ファイルに出力（merge）」をチェックすることで、出力結果を一つのファイルにまとめることもできます。

どうしても大きなファイルを処理する必要がある場合は、コマンドプロンプトで解析のみを行ってください。

Q：大量のファイルを一度に処理したら「このページのスクリプトが、Internet Explorer の実行速度を遅くしています。スクリプトの実行を続けると、コンピュータが反応しなくなる可能性があります。スクリプトを中断しますか？」という警告が出た。



A：「いいえ」を押すことでそのまま処理を続行できます。概ね 256 ファイル以上を一度に処理する場合にこの警告が出ます。

大量のファイルの処理を頻繁に行う場合は、UniDic をインストールしたフォルダ（C:\Program Files\unidic）にある timeout_off.reg ファイルを使ってレジストリを書き換えることで警告を抑制できます。このファイルをダブルクリックして実行し、「はい」をクリックしてください。レジストリを元に戻す場合は同じ場所にある timeout_default.reg を実行し、「はい」をクリックしてください。

Q：短いファイルを解析すると文字化けします。

A：あまり短いファイルだと文字コードの判別には失敗することがあります。余分に文を入力するなどして、長めにして解析してみてください。

Q：茶まめを強制終了したい。

A：万一応答がなくなってしまった場合は、タスクマネージャを起動して「chasen.exe」
「mshta.exe」のプロセスを終了させてください。

Q：解析結果を表示するテキストエディタを指定したい。

A：Internet Explorer でソースの表示をするエディタを変更してください。

(参考) @IT：Internet Explorer の [ソースの表示] メニューで起動するエディタを指定する

<http://www.atmarkit.co.jp/fwin2k/win2ktips/286iesourceview/iesourceview.html>

7. 著作権・問い合わせ先

「和文茶まめ」及び本マニュアルの著作権は小木曾智信が保持します。

問い合わせ先：net@ogiso.net