

ChaKi インストールマニュアル

version 0.10

August 21, 2006

浅原 正幸 松本 裕治

Copyright © 2006 奈良先端科学技術大学院大学

1 環境のインストール

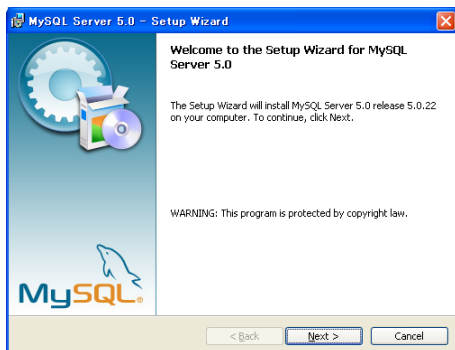
コーパスを検索する環境を構築するためには 3 つの作業が必要である :

- MySQL のインストール
- ChaKi GUI のインストール
- データ整形ツールのインストール

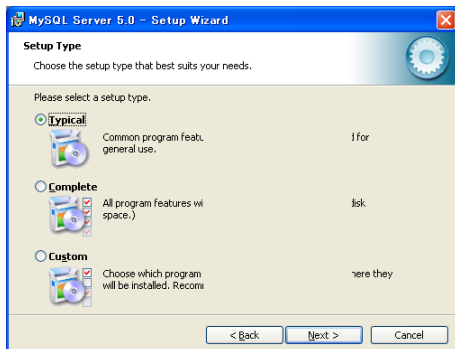
1.1 MySQL のインストール

MySQL のダウンロードページ (<http://dev.mysql.com/downloads/mysql/5.0.html>) から Windows のバイナリ (Windows Essentials) をダウンロードする。次にダウンロードした `mysql-essential-5.0.22.zip` を解凍し、`Setup.exe` を実行する。

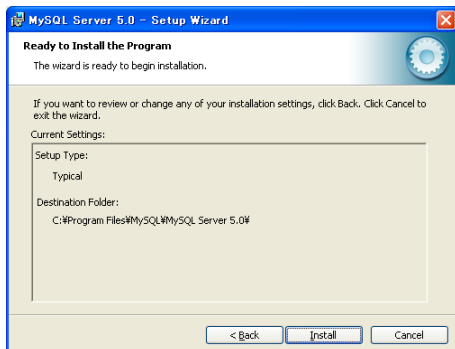
以下各画面で何をすべきかについて示す :



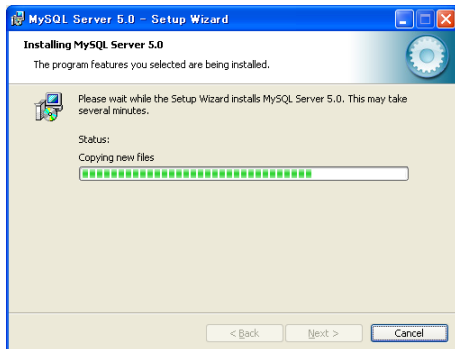
[next]



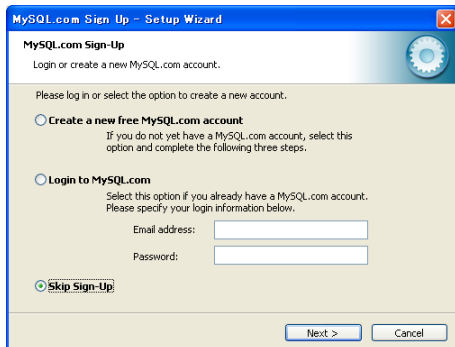
[Typical] を選択し [next]



[Install]



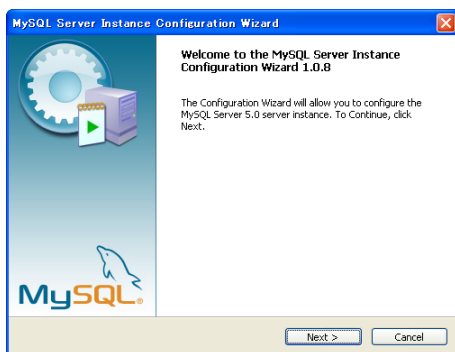
インストール作業中



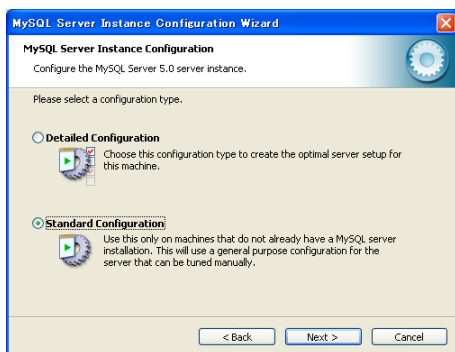
[Skip Sign-Up] [Next]



インストール作業の完了
[Configure the MySQL Server now] をチェックして [Finish]



[Next]



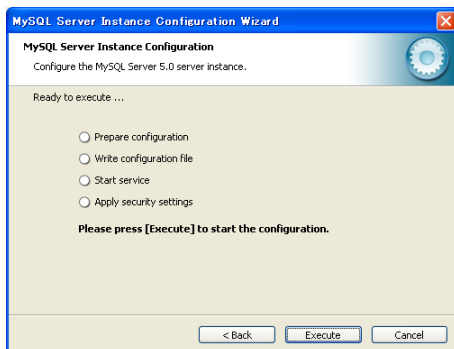
[Standard Configuration] [Next]



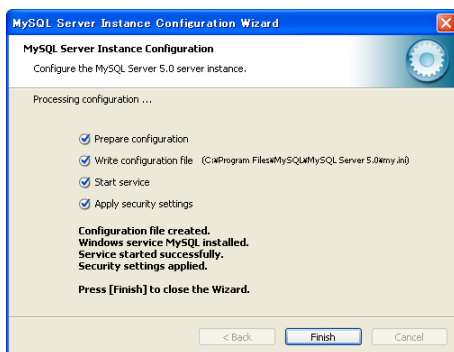
[Install As Windows Service]
[Launch the MySQL Server automatically]
Service Name:[MySQL]
[Include Bin Directory in Windows PATH]



[Modify Security Settings]
root のパスワードを入力：ここでは仮に “okage” パスワードとして話を進める。



[Execute]

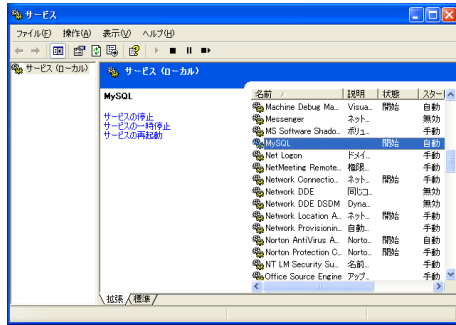


[Finish]
この手前でファイアウォールなどのセキュリティソフトを入れている場合にはファイアウォールの設定を変更する必要がある。

次に、my.ini を書き換える。通常は c:\Program Files\MySQL\MySQL Server5.0\my.ini にある。notepad.exe (メモ帳) などで開いて以下の項目を変更すること：

- “[mysql]” の下にある “default-character-set=latin1” を “default-character-set=sjis” に変更する。
- “[mysqld]” の数行下にある “default-character-set=latin1” を “default-character-set=sjis” に変更する。

最後に MySQL を再起動する。[スタート] [コントロールパネル] [パフォーマンスとメンテナンス] [管理ツール] [サービス] を開く。



MySQL を選択し、右クリックから再起動を選ぶ。同様のことが、Windows の再起動によっても行われる。

1.2 ChaKi GUI のインストール

ChaKi GUI のインストールは、本パッケージを適当な場所にコピーする。例えば、C ドライブの直下に `c:\chaki` というフォルダ名でコピーする。

1.3 データ整形ツールのインストール

1.3.1 英語の整形ツール

品詞タグづけ器のインストール まず品詞タグづけ器からインストールする。ここでは、品詞タグづけ器 `TreeTagger` を用いることにする。`TreeTagger` の配布ページ¹にある”Parameter files for PC (Linux and Windows, Latin1 character set)”とある項目の、English parameter file (english-par-linux-3.1.bin.gz)² と Windows version の `TreeTagger` 本体 (`tree-tagger-windows-3.1.zip`)³ とをダウンロードする。

ダウンロードしたファイルを展開する。まず、`TreeTagger` 本体である `tree-tagger-windows-3.1.zip` を展開する。展開されたフォルダを `c:\TreeTagger` とする。次にパラメータファイルを展開する。`english-par-linux-3.1.bin.gz` は、圧縮されたデータです。適切な解凍ツールを用いて解凍する。解凍ツールの例として `Lhaca` デラックス版⁴ などがある。解凍してできた `english-par-linux-3.1.bin` を `english.par` というファイル名を変更する。`c:\TreeTagger\lib` というフォルダがある。そこに `english.par` を置く。

1行1文となるテキストを作成するプログラムのインストール `answerbus`⁵ の `sentence segmenter`⁶ を用いる。Windows 用のプログラム⁷ をダウンロードして、フォルダ `c:\TreeTagger\bin` に置く。

単語単位に分割するプログラムのインストール NICT の内山さん⁸ が公開している `tokenizer.rb`⁹ を基に作成した実行ファイル `tokenizer.exe`¹⁰ を用いる。`tokenizer.exe` は、フォルダ `c:\TreeTagger\bin` に置く。

尚、先にインストールした `TreeTagger` にも単語単位に分割するプログラムがある。こちらの方は略語処理をしたり高機能だが、利用するためには `perl` をインストールする必要がある。`perl` が使いこなせる人はそちらを利用すること。

各プログラムを一括して実行するバッチファイルのインストール 今、インストールした各プログラムを一括して実行するためのバッチファイルがある。`c:\TreeTagger\bin` にある `TreeTagger` 付属の `tag-english.bat` を他の名前に変更し、本パッケージにある `tag-english.bat`¹¹ を置く。もし、`TreeTagger` を置く場所を変更した場合には、エディタで `tag-english.bat` を開いて、“`set TAGDIR=C:\TreeTagger`”の部分を `TreeTagger` を置いた場所に変更すること。

¹<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>

²<ftp://ftp.ims.uni-stuttgart.de/pub/corpora/english-par-linux-3.1.bin.gz>

³<ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger-windows-3.1.zip>

⁴<http://www.vector.co.jp/soft/win95/util/se166893.html>

⁵<http://answerbus.coli.uni-sb.de/index.shtml>

⁶<http://answerbus.coli.uni-sb.de/sentence/>

⁷<http://answerbus.coli.uni-sb.de/sentence/ss.exe>

⁸<http://www2.nict.go.jp/jt/a132/members/mutiyama/>

⁹<http://www2.nict.go.jp/jt/a132/members/mutiyama/software/ruby/tokenizer.rb>

¹⁰本パッケージ \prog\tokenizer.exe

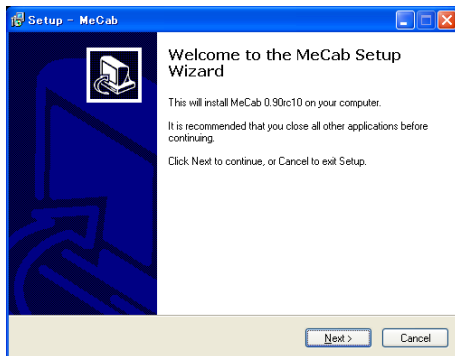
¹¹本パッケージ \prog\tag-english.bat

1.3.2 日本語の整形ツール

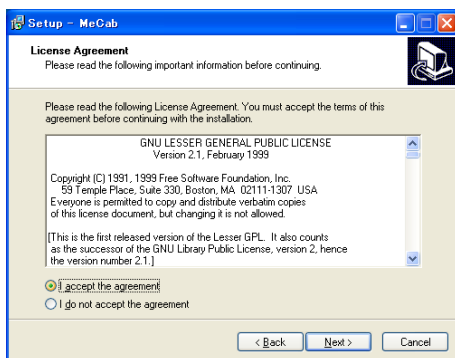
日本語の整形ツール概観 あらかじめ1行1文となるテキストを用意する。日本語の整形ツールとして、形態素解析器（わかち書きを行い品詞を付与する）と係り受け解析器を用いる。形態素解析器は MeCab, ChaSen, Juman 係り受け解析器は CaboCha, KNP などを用いることができる。

MeCab のインストール MeCab の配布ページ¹² から mecab-0.XX.exe (Binary Package for MS-Windows) をダウンロードする。2006年8月8日現在の最新版は mecab-0.93.exe。

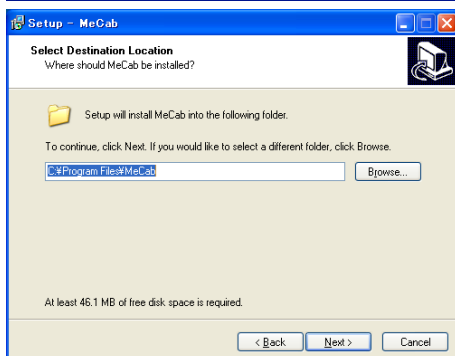
まず、mecab-0.XX.exe を実行する。その後、以下の画面のとおり作業を行う：



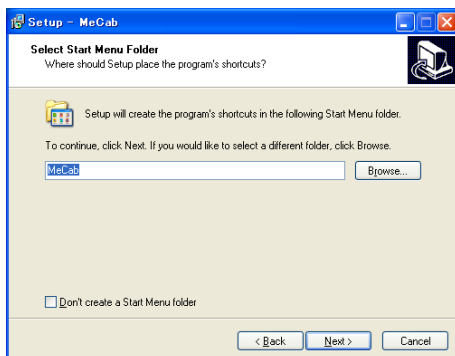
[Next]



[I accept the agreement] をチェックして [Next]

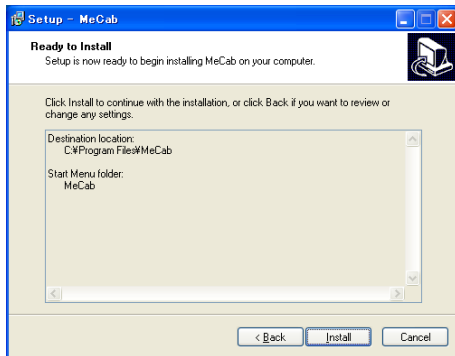


インストール先を確認して [Next]

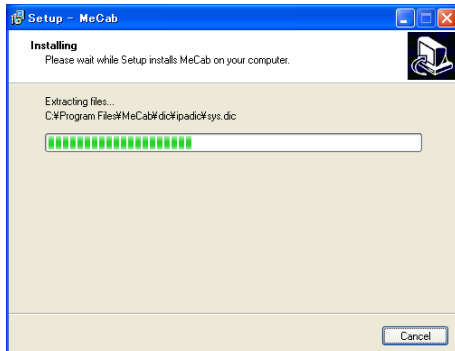


Start Menu に作成されるフォルダ名を確認して [Next]

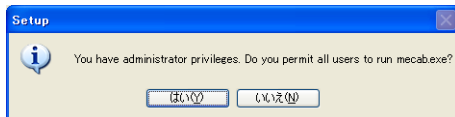
¹²<http://mecab.sourceforge.jp/>



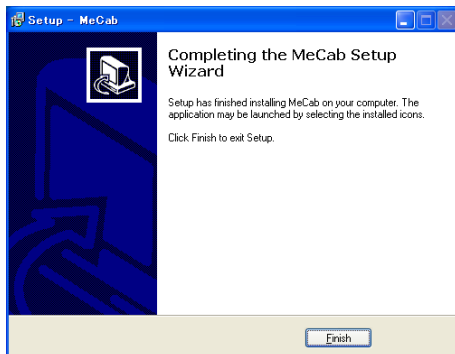
Start Menu に作成されるフォルダ名を確認して [install]



インストール作業中



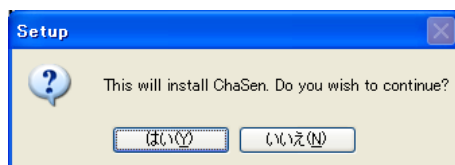
他のユーザも MeCab を利用して良いか？ 良いなら [Yes]



[Finish]

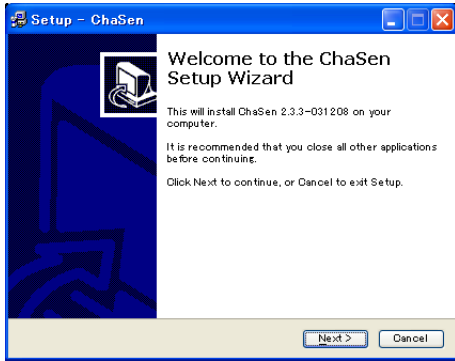
ChaSen のインストール ChaSen の配布ページ¹³ から cha233_031208.exe (Windows 版) をダウンロードする。尚、WinCha cha21244sp5.exe ではないことを確認すること。

まず、cha233_031208.exe を実行する。その後、以下の画面のとおり作業を行う：

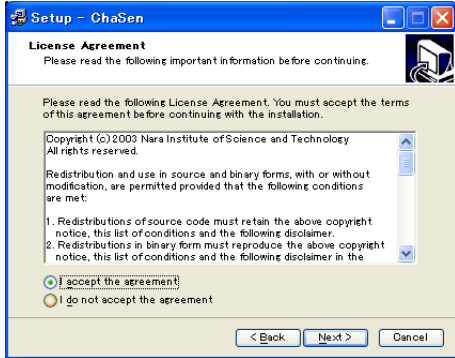


[Yes]

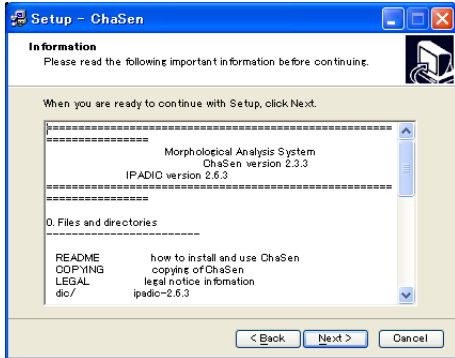
¹³<http://chasen.naist.jp/hiki/ChaSen/>



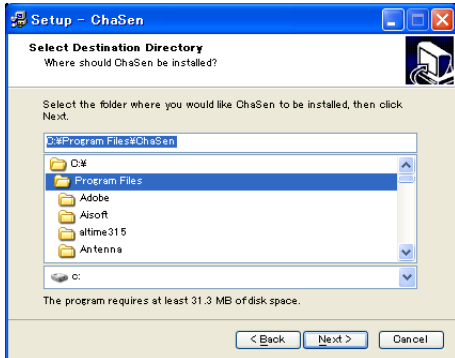
[Next]



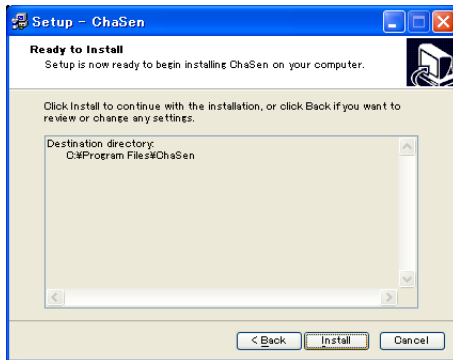
[I accept the agreement] をチェックして [Next]



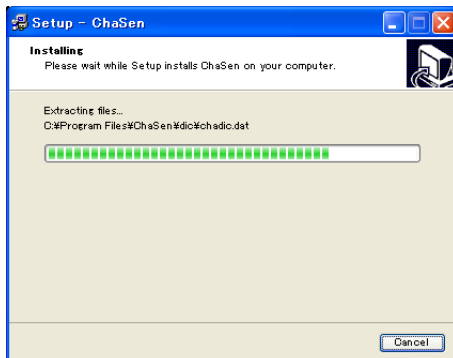
[Next]



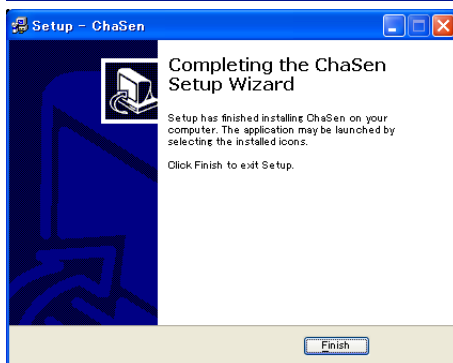
インストール先を確認して [Next]



[Install]



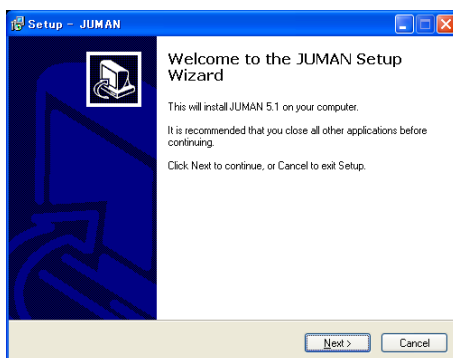
インストール中



[Finish]

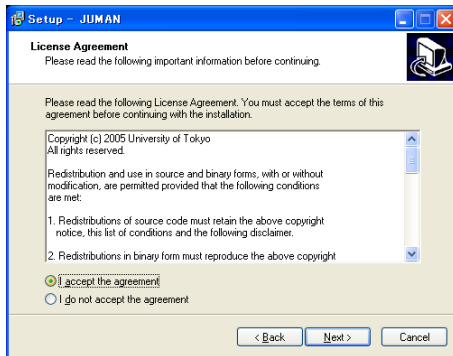
JUMAN のインストール JUMAN の配布ページ¹⁴ から juman-X.X.exe (Windows 版) をダウンロードしてください。2006 年 8 月 8 日現在の最新版は juman-5.1.exe.

まず、juman-5.1.exe を実行する。その後、以下の画面のとおり作業を行う：

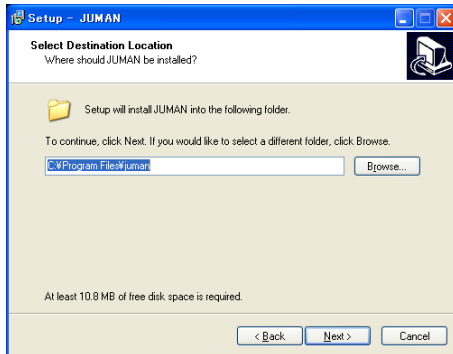


[Next]

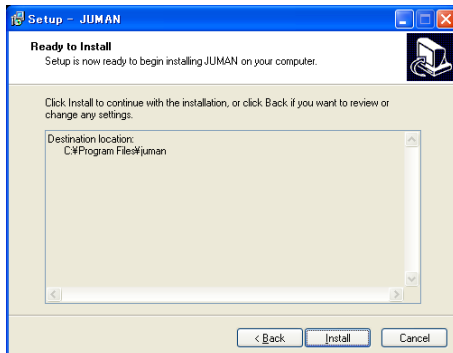
¹⁴<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>



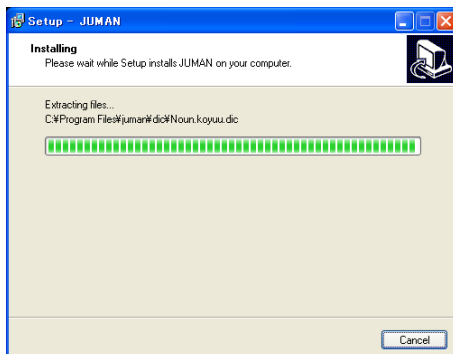
[I accept the agreement] をチェックして [Next]



インストール先を確認して [Next]



[Install]



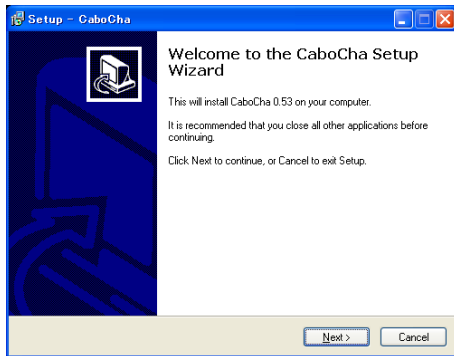
インストール中



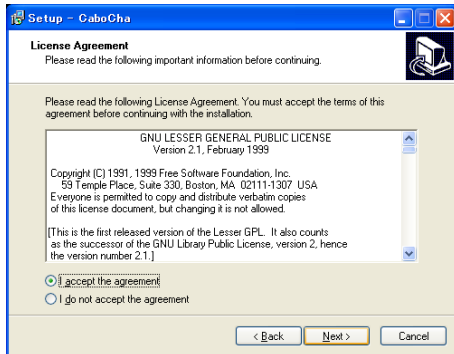
[Finish]

CaboCha のインストール CaboCha の配布ページ¹⁵ から cabocha-X.XX.exe (Windows 版) をダウンロードする。2006 年 8 月 8 日現在の最新版は cabocha-0.53.exe。

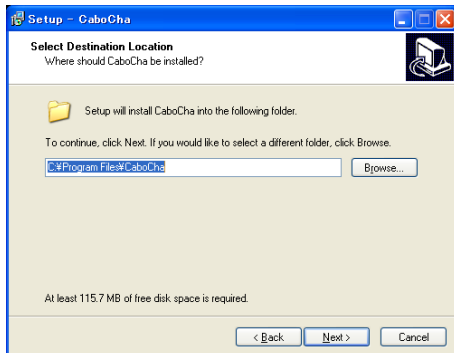
まず、cabocha-0.53.exe を実行する。その後、以下の画面のとおり作業を行う：



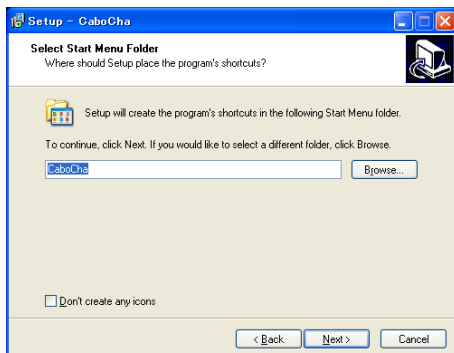
[Next]



[I accept the agreement] をチェックして [Next]

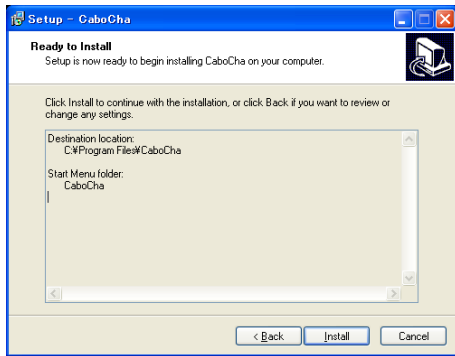


インストール先を確認して [Next]

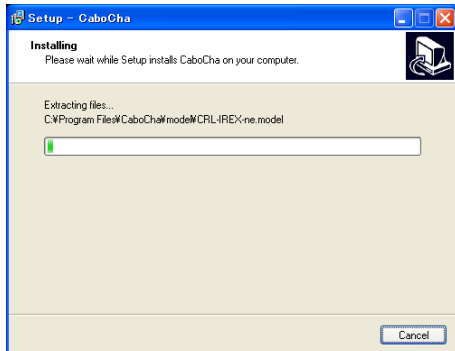


Start Menu に作成されるフォルダ名を確認して [Next]

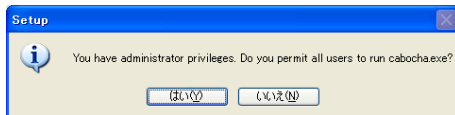
¹⁵<http://chasen.org/~taku/software/cabocha/>



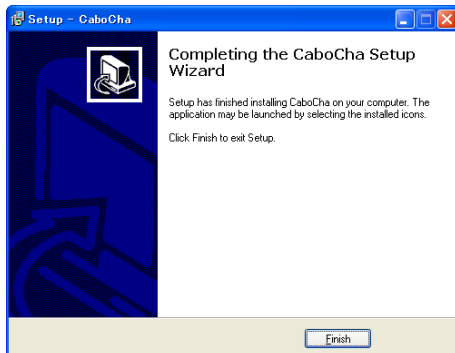
Start Menu に作成されるフォルダ名を確認して [Install]



インストール中



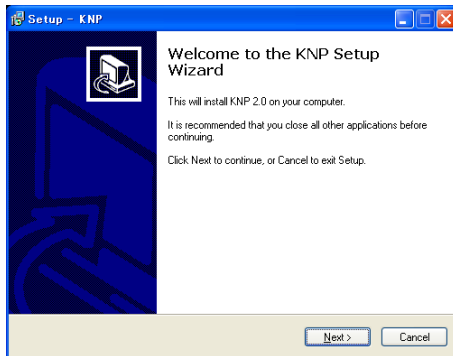
他のユーザも CaboCha を利用して良いか？ 良いなら [Yes]



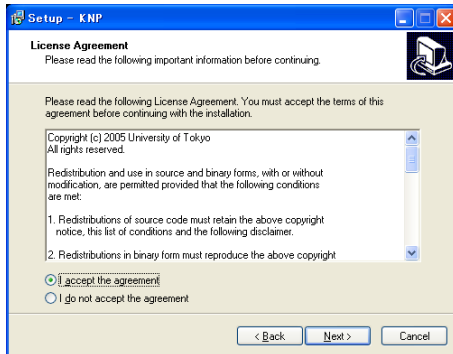
[Finish]

KNP のインストール KNP の配布ページ¹⁶ から knp-X.X.exe (Windows 版) をダウンロードする。2006 年 8 月 8 日現在の最新版は knp-2.0.exe.

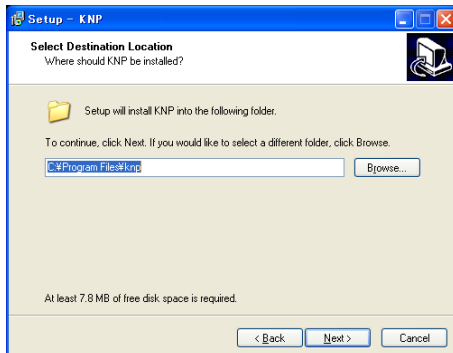
¹⁶<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html>



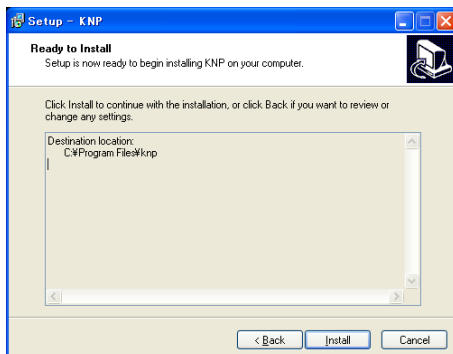
[Next]



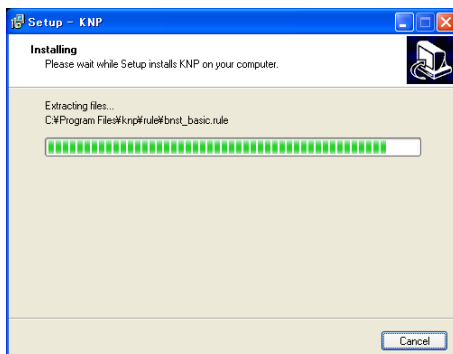
[I accept the agreement] をチェックして [Next]



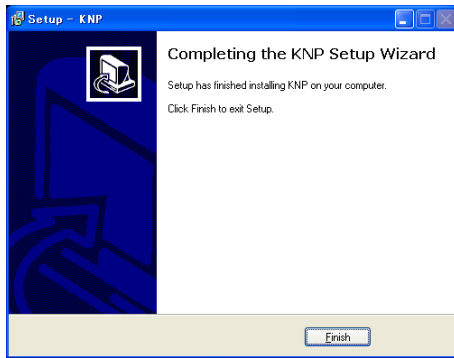
インストール先を確認して [Next]



[Install]



インストール中



[Finish]

2 データの格納

2.1 サンプルデータのインポート

2.1.1 サンプルデータについて

フォルダ db 以下に ChaKi のサンプルデータとして英語のデータと日本語のデータの 2 種類ある。英語データは Project Gutenberg¹⁷ から取った 6 ファイル、日本語データは青空文庫¹⁸ から取った 4 ファイル。ファイルの内容は以下の通り：

- English
 - Dickens Charles: A Christmas Carol
 - Dickens Charles: A Tale of Two Cities
 - Dickens Charles: Oliver Twist
 - Jane Austen: Emma
 - Jane Austen: Persuasion
 - Jane Austen: Pride and Prejudice
- Japanese
 - 芥川龍之介:鼻
 - 芥川龍之介:羅生門
 - 夏目漱石:こころ
 - 夏目漱石:三四郎

2.1.2 サンプルデータのインポート

英語のデータは English という名前のデータベースに、日本語のデータは Japanese という名前のデータベースに格納される。

まず、一時ファイルを置くフォルダを作成する。ここでは c:\temp というフォルダを作成することにする。次にフォルダ db 以下にあるバッチファイル、english.bat と japanese.bat をエディタで開く。ファイルは次のような内容である（改行されているが実際には 1 行）：

```
..\prog\cabocha2dat.exe -f english -t c:\temp -h localhost -u root  
-p okage -d english --corpusformat=English --spacing=English
```

そして次の項目をチェックすること：

- 一時ファイルを置くフォルダを変更した人は “-t c:\temp” の “c:\temp” の部分を書き換える。

¹⁷<http://www.gutenberg.org/>

¹⁸<http://www.aozora.gr.jp/>

- MySQL のパスワードを変更した人は “-p okage” の “okage” の部分を書き換える。
- 他のデータベース名で格納したい人は “-d english” の “english” の部分を書き換える。japanese.bat も同様。

各バッチファイルを（ダブルクリックして）実行する。

最後に、フォルダ db 以下にある english.def と japanese.def を、ChaKi のフォルダに（もしなければ）格納する。これらのファイルは以下のような形式になっている：

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

そして次の項目をチェックする：

- MySQL のパスワードを変更した人は “password=okage” の “okage” の部分を書き換える。
- 他のデータベース名で格納したい人は “corpusname=english” の “english” の部分を書き換える。japanese.def も同様。

2.1.3 サンプルデータの削除

サンプルデータを MySQL のデータベースから削除する場合、まずコマンドプロンプトを立ち上げる。その後、次のコマンドで mysql.exe を実行すること。

```
mysql.exe -uroot -pokage
```

- MySQL のパスワードを変更した人は “-pokage” の “okage” の部分を書き換える。

次に mysql のプロンプトに対し、以下を実行する。

```
drop database english;
drop database japanese;
```

2.2 自作データのインポート (英語)

現在、手元に ChaKi に入れたい英語のテキストファイル、もしくは HTML ファイルがあると仮定する。あらかじめ「データ整形ツールのインストール」を行うこと。

まず、整形するためのデータを準備する。例えば、Project Gutenberg ¹⁹ などのページから著作権が切れたテキストを集めてくる。

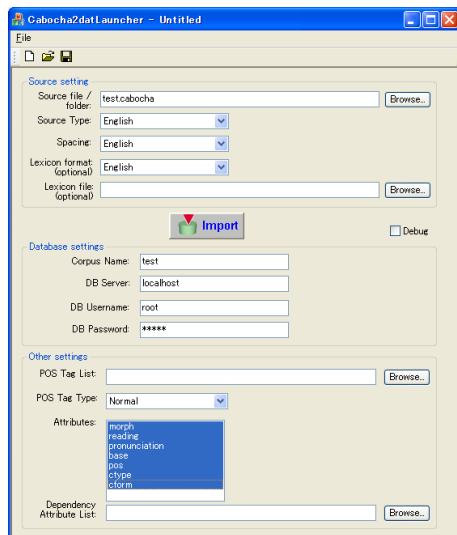
データ整形は \prog にある tag-english.bat により行う。準備したデータを c:\test.txt とする。コマンドプロンプトを開き、次のように入力することによって行う。

```
c:\TreeTagger\bin\tag-english.bat c:\test.txt c:\test.tnt
```

c:\test.tnt が整形されたデータである。エディタなどで開いて中身を確認すること。

データ格納は、ChaKi フォルダにある Cabocha2datLauncher.exe を用いる。例えば、以下のように設定する。

¹⁹<http://www.gutenberg.org/>



各オプションの意味は以下のとおり：

- Source file/folder: データ整形済コーパスのファイル名もしくはデータ整形済コーパスが入っているフォルダを指定する。
- Source Type: コーパスの形式（上記英語の整形ツールを使用時には “English”）
- Spacing: 単語間のスペース（英語データの場合 “English”）
- Corpus Name データベース名（自分で適当な名前をつける）
- DB Server ホスト名（通常 localhost）
- DB Username MySQL におけるユーザ名
- DB Password MySQL の上記ユーザ名に対応するパスワード

同様のことが ChaKi フォルダにある cabocha2dat.exe によっても行うことができる。コマンドラインから次のように入力する（改行が入っているが実際には 1 行）：

```
c:\chaki\chaki\cabocha2dat.exe -c test.tnt -t c:\temp -h localhost -u root
-p okage -d test --corpusformat=English --spacing=English
```

各オプションの意味は以下のとおり：

- -c コーパス名
- -t 一時ファイルを出力するフォルダ
- -h ホスト名（通常 localhost）
- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名（自分で適当な名前をつけてください）
- --corpusformat コーパスの形式（上記英語の整形ツールを使用時には “English”）
- --spacing 単語間のスペース（英語データの場合 “English”）

2.3 自作データのインポート (日本語)

現在、手元に ChaKi に入れたい日本語の 1 行 1 文となっているテキストファイルを仮定する。あらかじめ「データ整形ツールのインストール」を行っておくこと。

まず、整形するためのデータを準備する。例えば、青空文庫²⁰などのページから著作権が切れたテキストを集めてくる。

²⁰<http://www.aozora.gr.jp/>

2.3.1 MeCab を用いる方法

MeCab は、形態素解析器であり、デフォルトの設定では IPA 品詞体系をタグづけする。

まず、格納したい日本語のテキストファイルを用意する。テキストファイルのフォーマットは、1 行 1 文となっている文字コードが SJIS のファイルとする。テキストファイルを入力.txt とすると

```
\Program Files\MeCab\bin\mecab.exe input.txt > input.mecab
```

により、解析済ファイル input.mecab を得る。

次に、コマンドラインから次のように入力する（改行が入っているが実際には 1 行）：

```
c:\chaki\chaki\cabocha2dat.exe -c input.mecab -t c:\temp -h localhost -u root  
-p okage -d inputmecab --corpusformat=MeCab --spacing=Japanese
```

各オプションの意味は以下のとおり：

- -c コーパス名
- -t 一時ファイルを出力するフォルダ
- -h ホスト名（通常 localhost）
- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名（自分で適当な名前をつけてください）
- --corpusformat コーパスの形式（MeCab 使用時には“MeCab”）
- --spacing 単語間のスペース（日本語データの場合“Japanese”）

2.3.2 ChaSen を用いる方法

ChaSen は、形態素解析器であり、デフォルトの設定では、IPA 品詞体系をタグづけする。

MeCab の際と同様に文字コードが SJIS である 1 行 1 文となっている日本語のテキストファイル input.txt を用意する。

ChaSen のデフォルトの出力では発音情報が得られない。次のようにして -F オプションを用い、

```
\Program Files\ChaSen\chasen.exe -F "%m\t%y\t%a\t%M\t%U(%P-)\t%T \t%F \n" input.txt > input.chasen
```

により、解析済ファイル input.chasen を得る。

同様のことが \Program Files\ChaSen\dic\chasenrc に、以下のように記述することで、-F オプションなしで行うことができる：

（出力フォーマット "%m\t%y\t%a\t%M\t%U(%P-)\t%T \t%F \n"）

次に、コマンドラインから次のように入力する（改行が入っているが実際には 1 行）：

```
c:\chaki\chaki\cabocha2dat.exe -c input.chasen -t c:\temp -h localhost -u root  
-p okage -d inputchasen --corpusformat=ChaSen --spacing=Japanese
```

各オプションの意味は以下のとおり：

- -c コーパス名
- -t 一時ファイルを出力するフォルダ
- -h ホスト名（通常 localhost）
- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名（自分で適当な名前をつけてください）
- --corpusformat コーパスの形式（chasen 使用時には“ChaSen”）
- --spacing 単語間のスペース（日本語データの場合“Japanese”）

2.3.3 JUMAN を用いる方法

JUMAN は、形態素解析である。デフォルトの設定では、益岡・田窪品詞体系をタグづけする。

MeCab の際と同様に文字コードが SJIS である 1 行 1 文となっている日本語のテキストファイル `input.txt` を準備し、

```
\Program Files\juman\juman.exe < input.txt > input.juman
```

により、解析済ファイル `input.juman` を得る。

次に、コマンドラインから次のように入力する（改行が入っているが実際には 1 行）:

```
c:\chaki\chaki\cabocha2dat.exe -c input.juman -t c:\temp -h localhost -u root  
-p okage -d inputjuman --corpusformat=JUMAN --spacing=Japanese
```

各オプションの意味は以下のとおり：

- -c コーパス名
- -t 一時ファイルを出力するフォルダ
- -h ホスト名（通常 localhost）
- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名（自分で適当な名前をつけてください）
- --corpusformat コーパスの形式（JUMAN 使用時には“JUMAN”）
- --spacing 単語間のスペース（日本語データの場合“Japanese”）

2.3.4 CaboCha を用いる方法

CaboCha は、係り受け解析器であり、内部で ChaSen もしくは MeCab を呼び出すことにより、デフォルトの設定では IPA 品詞体系と係り受け情報をタグづけする。

MeCab の際と同様に文字コードが SJIS である 1 行 1 文となっている日本語のテキストファイル `input.txt` を準備し、

```
\Program Files\CaboCha\bin\cabocha.exe -f1 input.txt > input.cabocha
```

により、解析済ファイル `input.cabocha` を得る。

次に、コマンドラインから次のように入力する（改行が入っているが実際には 1 行）:

```
c:\chaki\chaki\cabocha2dat.exe -c input.cabocha -t c:\temp -h localhost -u root  
-p okage -d inputcabocha --corpusformat=cabocha --spacing=Japanese
```

各オプションの意味は以下のとおり：

- -c コーパス名
- -t 一時ファイルを出力するフォルダ
- -h ホスト名（通常 localhost）
- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名（自分で適当な名前をつけてください）
- --corpusformat コーパスの形式（CaboCha 使用時には“CaboCha”）
- --spacing 単語間のスペース（日本語データの場合“Japanese”）

2.3.5 KNP を用いる方法

KNP は、係り受け解析器であり、形態素解析器 juman の出力に対し係り受け情報をタグづけする。

MeCab の際と同様に文字コードが SJIS である 1 行 1 文となっている日本語のテキストファイル input.txt を準備し、

```
\Program Files\juman\juman.exe < input.txt | \Program Files\knp\knp.exe -tab > input.knp
```

により、解析済ファイル input.knp を得る。

次に、コマンドラインから次のように入力する（改行が入っているが実際には 1 行）：

```
c:\chaki\chaki\cabocha2dat.exe -c input.knp -t c:\temp -h localhost -u root  
-p okage -d inputknp --corpusformat=KNP --spacing=Japanese
```

各オプションの意味は以下のとおり：

- -c コーパス名
- -t 一時ファイルを出力するフォルダ
- -h ホスト名（通常 localhost）
- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名（自分で適当な名前をつけてください）
- --corpusformat コーパスの形式（KNP 使用時には“KNP”）
- --spacing 単語間のスペース（日本語データの場合“Japanese”）

2.4 Penn Treebank コーパスの格納

Penn Treebank を ChaKi に格納するための手引。

2.4.1 とりあえず入れる

Penn Treebank のパッケージの中の combined フォルダを用いる。combined フォルダをこのパッケージの中にある db\ptb フォルダに置く。

次に、db\ptb フォルダにある ptbimport.bat を編集する。確認すべき項目は ptbimport.bat の 2 行目のオプションである。

- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名（自分で適当な名前をつけてください）

ptbimport.bat を実行する。実行には数分（マシンによっては数十分）かかるので注意すること。

最後に、db\ptb フォルダにある ptb.def, ptb.pos を ChaKi フォルダに移動する。ptb.def の中のサーバ名、ユーザ名、パスワードは適宜変更すること。

2.4.2 各処理の説明

上のバッチファイルは次のような作業を行う：

1. CaboCha 出力形式への整形
2. データの格納

CaboCha 出力形式への整形

```
combined2s.exe combined\wsj | ptbconv.exe -D |dep2cabocha.exe > ptb.cabocha
```

ここで利用している 3 つのプログラムは以下の作業をする。

- combined2s.exe
Penn Treebank のディレクトリを入力とし、ディレクトリ以下の全ての .mrg ファイルを S 式 (Parsed Tree を括弧で表現したもの) に変換する。
- ptbconv.exe 北陸先端科学技術大学院大学²¹ の山田寛康氏作成のプログラム。Penn Treebank の Parsed Tree を様々な形式へと変換することができる。オリジナルのプログラムは彼のページ²² からダウンロードが可能。
- dep2cabocha.exe
ptbconv の出力を CaboCha 出力形式へと変換するプログラム。

データの格納

```
..\ChaKi\cabocha2dat.exe -c ptb.cabocha -t c:\temp -h localhost -u root \  
-p okage -d ptb --corpusformat=CaboCha --spacing=English
```

各オプションの意味は以下のとおり：

- -c コーパス名
- -t 一時ファイルを出力するフォルダ
- -h ホスト名 (通常 localhost)
- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名 (自分で適当な名前をつける)
- --corpusformat コーパスの形式 ("CaboCha" にすること！)
- --spacing 単語間のスペース (英語データの場合 "English")

2.5 BNC コーパスの格納

BNC コーパス を ChaKi に格納するための手引。

2.5.1 とりあえず入れる

用いるデータは BNC コーパスの CD-ROM の disk 1 にある texts.tar.gz。適当な解凍ツールを用いて、まずこのファイルを解凍する。但し、BNC コーパスは非常にサイズが大きいため、ディスクの空き容量が十分であることを確認してから解凍する。解凍してできた Texts フォルダをこのパッケージにある db\bnc フォルダに置く。

次に、db\bnc フォルダにある bncimport.bat を編集する。bncimport.bat の 偶数行目のオプションを確認する。

- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名 (自分で適当な名前をつける)

尚、bnc コーパスはデータ量が大きいため、各フォルダ毎に格納する。デフォルトの bncimport.bat では、Texts\A フォルダのみを入れる。他のデータも格納する場合には、bncimport.bat の「@rem」の部分を削除すること。

bncimport.bat を実行する。実行には数分 (マシンによっては数十分) かかるので注意すること。全データを入れる場合には、ディスク容量が十分であることを確認すること。

²¹<http://www.jaist.ac.jp/>

²²<http://www.jaist.ac.jp/~h-yamada/>

2.5.2 各処理の説明

上のバッチファイルは次のような作業を行う：

1. TreeTagger 出力形式への整形
2. データの格納

TreeTagger 出力形式への整形

```
bnc2tnt.exe Texts\A > bncA.tnt
```

Texts\A フォルダ以下のコーパスファイルを読み込み、TreeTagger の出力形式ファイル bncA.tnt を出力する。

データの格納

```
..\..\ChaKi\cabocha2dat.exe -c bncA.cabocha -t c:\temp -h localhost -u root\\  
-p okage -d bncA --corpusformat=English --spacing=English
```

各オプションの意味は以下のとおり：

- -c コーパス名
- -t 一時ファイルを出力するフォルダ
- -h ホスト名 (通常 “localhost”)
- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名 (自分で適当な名前をつける)
- --corpusformat コーパスの形式 (“English” にすること！)
- --spacing 単語間のスペース (英語データの場合 “English”)

2.6 京都テキストコーパスの格納

京都テキストコーパス (形態素解析済み) を ChaKi に格納するための手引。

2.6.1 とりあえず入れる

用いるデータは 京都テキストコーパスの dat/ フォルダ以下にある syn/ フォルダに生成される拡張子が .KNP のデータ。

次に、db\kc フォルダにある kcimport.bat を編集する。kcimport.bat の 2 行目のオプションを確認する。

- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名 (自分で適当な名前をつける)

2.6.2 各処理の説明

上のバッチファイルは次のような作業を行う：

1. CaboCha 出力形式への整形
2. データの格納

TreeTagger 出力形式への整形

```
kc2cabocha.exe syn
```

syn フォルダ以下のコーパスファイルを読み込み、CaboCha の出力形式ファイルを kc4 フォルダ以下に出力する。これは、京都テキストコーパス .KNP ファイルのフォーマットが、係り受け解析器 KNP の出力と若干異なるために行う。

データの格納

```
..\..\ChaKi\cabocha2dat.exe -f syn -t c:\temp -h localhost -u root\  
-p okage -d kc --corpusformat=CaboCha --spacing=Japanese
```

各オプションの意味は以下のとおり：

- -f コーパスが格納されているフォルダ名
- -t 一時ファイルを出力するフォルダ
- -h ホスト名（通常 “localhost”）
- -u MySQL におけるユーザ名
- -p MySQL の上記ユーザ名に対応するパスワード
- -d データベース名（自分で適当な名前をつける）
- --corpusformat コーパスの形式（“CaboCha” にすること！）
- --spacing 単語間のスペース（英語データの場合 “Japanese”）

3 サーバとクライアントを分離する

これまでの節では、1 台のマシンにサーバ (MySQL サーバ) とクライアント (ChaKi GUI) を設定する方法について説明した。本節ではサーバとクライアントを別々のマシンに持たせる方法について説明する。

3.1 サーバ (Windows) の設定

サーバとして Windows マシンを用いる場合、1.1 節にある方法で MySQL サーバをインストールする。サーバマシンの IP address が 192.168.0.2 (もしくは hostname が server.your.domain), クライアントマシンの IP address が 192.168.0.3 (もしくは hostname が client01.your.domain) とする。クライアントマシンが用いるユーザ名を “user”、パスワードを “shika” とする。

検索対象のコーパス名を “japanese” とする。クライアントマシンのユーザにこのコーパスの検索、修正を許す場合、サーバマシンのコマンドプロンプトから以下のように設定する。

```
mysql -u root -p  
Enter password:<<mysql の root のパスワードを入力>>
```

```
mysql> GRANT SELECT, INSERT, UPDATE, DELETE ON japanese.*  
TO user@192.168.0.3 IDENTIFIED BY 'shika';  
mysql> exit;
```

```
mysqladmin -uroot -p flush-privileges  
Enter password:<<mysql の root のパスワードを入力>>
```

hostname で指定する場合、‘user@192.168.0.3’ の箇所を ‘user@client01.your.domain’ とすれば良い。

クライアントマシンのユーザにこのコーパスの検索のみを許す場合、サーバマシンのコマンドプロンプトから以下のように設定する。

```
mysql -u root -p
Enter password:<<mysql の root のパスワードを入力>>
```

```
mysql> GRANT SELECT ON english.* TO user@192.168.0.3 IDENTIFIED BY 'shika';
mysql> exit;
```

```
mysqladmin -uroot -p flush-privileges
Enter password:<<mysql の root のパスワードを入力>>
```

3.2 クライアント (Windows) の設定

クライアントの設定は、コーパス定義ファイル .def ファイルの書き換えのみによる。サーバマシンの IP address が 192.168.0.2 (もしくは hostname が server.your.domain), クライアントマシンの IP address が 192.168.0.3 (もしくは hostname が client01.your.domain) とする。クライアントマシンが用いるユーザ名を “user”、パスワードを “shika” とする。

検索対象のコーパスを “japanese” とすると japanese.def ファイルの内容は以下のようになる：

```
corpusname=japanese
server=192.168.0.2
user=user
password=shika
```

hostname で指定する場合、 ‘ ‘server=192.168.0.2’ ’ の箇所を ‘ ‘server=server.your.domain’ ’ とすれば良い。

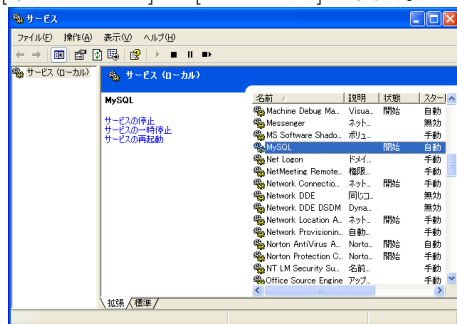
4 UTF8 環境の構築

4.1 MySQL の設定変更

my.ini を書き換える。通常は c:\Program Files\MySQL\MySQL Server5.0\my.ini にある。notepad.exe (メモ帳) などを開いて以下の項目を変更すること：

- “[mysql]” の下にある “default-character-set=latin1” を “default-character-set=utf8” に変更する。
- “[mysqld]” の数行下にある “default-character-set=latin1” を “default-character-set=utf8” に変更する。

最後に MySQL を再起動する。[スタート] [コントロールパネル] [パフォーマンスとメンテナンス] [管理ツール] [サービス] を開く。



MySQL を選択し、右クリックから再起動を選ぶ。同様のことが、Windows の再起動によっても行われる。

4.2 コーパスの格納

文字コードが utf8 である解析済みのコーパスデータの場合、cabocha2dat.exe に --encode=UTF8 オプションをつけることにより、データを格納することが可能である。

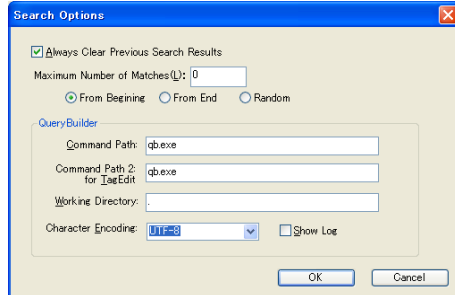
例えば、コマンドラインから次のように入力する（改行が入っているが実際には 1 行）：

```
c:\chaki\chaki\cabocha2dat.exe -c input.cabocha -t c:\temp -h localhost -u root
-p okage -d inputcabocha --corpusformat=cabocha --spacing=Japanese
--encode=UTF8
```

プログラム cabocha2dat.exe は、文字コードの変換を行わないため、あらかじめ input.cabocha を他の手段で utf8 に変換しておく必要がある。

4.3 GUI の設定

[Options] [Search Options] の画面を開き、[Character Encoding] を以下のように、UTF-8 に変更する：



日本語以外の言語の場合、用いるフォントの制約により文字化けすることがある。その場合には、[Options] [Font Setting] の画面を開き、[KwicColumnPrimary] などの項目を、当該言語の文字集合を含むフォントに変更すること。

5 コーパス定義ファイル(.def ファイル) について

lexicontype

- 0: none
- 1: english
- 2: cabocha
- 3: chasen
- 4: comp(複合語)