多言語のテキストマイニングの統計ツールMLTP

同志社大学文化情報学部
金　明哲

MLTPについて

- MLTP: Multilingual Text Processor
- 開発用言語: JAVA(Windows)、2003年から～
- 目的：個人研究用のツール
- 基本的な考え方と現状：
  - ✓ 平テキスト、タグ付きテキストの要素を集計
  - ✓ n-gram，共起，文節の係り受け、文節のパターン
  - ✓ 統計解析は専用ソフトに任せる
  - ✓ 中国語、日本語、韓国語を中心としているがUFTコードによる左から右方向の横書き電子テキストであれば基本的には利用可能
  - ✓ 扱うファイルは*.txt形式
  - ✓ GUIのメニューは現段階では英語(言語ごとにメニュー作成に･･･)
  - ✓ 結果の出力は、Tab区切りとcsv形式

言語の種類と用いるコーパス
Type of lunguage and text

Tyep of Lenguge	Type of text		
	Plain	Tagged	Parser
Chinese	○	○	×
Japanese	○	○	○
English	○	○	×
Korean	○	○	×

文字コード(Encodings of Input files)

Japanese	Chinese	Korean
ISO-202-JP SHIFT_JIS EUC-JP	ISO-2022-CN BIG5 EUC-TW GB18030 HZ-GB-23121	ISO-2022-KR EUC-KR Johab
UTF-8, UTF-16BE/16LE, UTF-32BE/32LE/ X-ISO-10646-UCS-4-34121/4-21431		

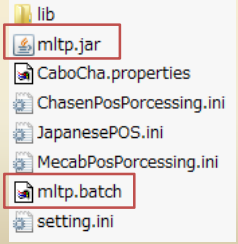
主な基本機能

	Chinese		Japanese			English		Korean	
	Plain Text	Tagged Text	Plain Text	Tagged Text	Parser	Plain Text	Tagged Text	Plain Text	Tagged Text
File List	○	○	○	○	○	○	○	○	○
Summary	○	○	○	○	×	○	○	○	○
n-gram	○	○	○	○	○	○	○	○	○
Length	○	○	○	○	○	○	○	○	○
Mark	○	○	○	○	×	○	○	○	○
KWIC	○	○	○	○	×	○	○	○	○
Word List	○	×	○	×	×	○	×	○	×
Replacement	○	○	○	○	×	○	○	○	○
Pattern	×	×	×	×	○	×	×	×	×
Co-occurrence	×	○	×	○	○	×	○	×	○

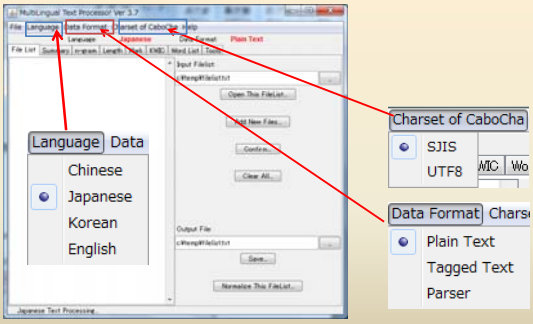
配布先など

- http://mjin.doshisha.ac.jp/MLTP/
- http://textdata.web.fc2.com/

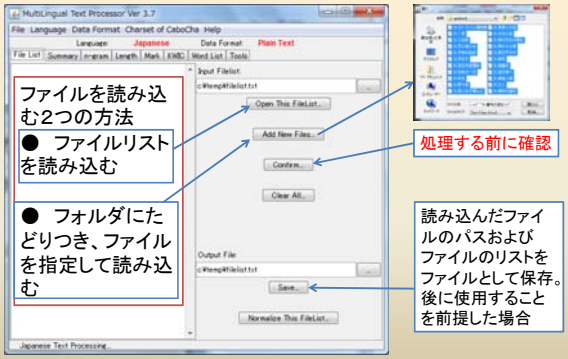
➢ダウンロードし、ローカルの適切な場所で解凍する。
➢解凍したフォルダの中には、右のようなファイルが含まれている。
➢mltp.jarをクリックするとMLTPが起動される。
➢拡張子batchをbatに直して、クリックするとより多くの量のデータに対応できる。



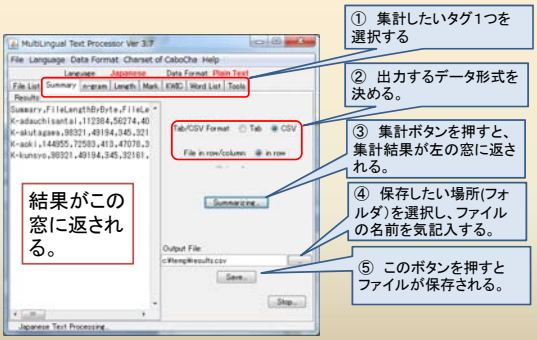
起動画面



ファイルの読み込み



集計と結果の保存



平テキストの主な処理機能とタブ

- Summary(文字数、文数、漢字・かなの数など)
- n-gram(文字、語句単位)
- Length(文の長さの分布、語句の長さの分布)
- Mark(ある記号、文字の前後のデータ)
- KWICK(ある文字列の前後の文字・記号)
- WordList(語・句のリストに基づいた集計)
- Tools(すべてファイルの一括処理。例えば、置き換え)

タブSummary

- 日本語の場合

FileLengthByByte(バイト単位の文章の長さ)
FileLengthByChar (文字単位の文章の長さ)
SentencesNum(文章における文の数)
CharNum(文章における文字の数)
KanjiNum (文章における漢字の数)
HiraganaNum (文章における平仮名の数)
KatakanaNum (文章におけるカタガナの数)
RomajiNum (文章におけるローマ字の数)
NumberNum (文章における数字の数),
ZenkakuKigoNum (文章における全角記号の数)
HankakuKigoNum (文章における半角記号の数)

タブn-gram

- 日本語、中国語の場合:文字単位
- 英語、韓国語場合:語句を単位(スペースを境)
- n-gramのnは1から6まで
- データの形式:タブとカンマ区切り、結果表を転置
- データのサイズは、Cutoff値で制御可能

  Cutoff値は、ある項目の総合計の値。

  Cutoff値より小さい項目は、すべて1つの項目“OTHERS”にまとめる

タブLength

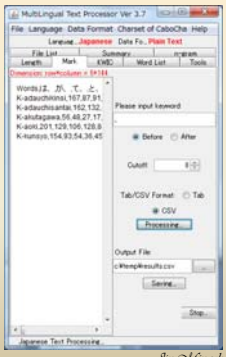
- 日本語、中国語の場合：文字単位の文の長さ
- 英語、韓国語場合：文字、語句を単位
- データの形式：タブ、カンマ区切り、結果表の転置
- データのサイズは、Category IndexとCutoff値制御

Category Index：何単位を1つの項目にまとめるか

Cutoff値： 最大の項目数。これ以上長いものは1つの項目にまとめる。

タブMark

- 指定した文字・記号列の前・後の記号・文字の集計
- データ形式：タブ区切り、カンマ区切り



タブKWIC

- 文字列、正規表現の前後一定の長さを返す。
- 前後の長さは、自由に指定できる。
- 結果はソート可能
- 結果と本文なかの対応関係の考察が可能

タブwordlist

- 指定したリスト語句項目に従って集計を行う
- 語句の指定は、論理演算（and , or）で記述可能。Andは半角記号&, orは半角記号|を用いる。
- 作成したリストは、保存して用いることが可能

タブTools

- 主な機能、一括置換、さまざまな括弧内のものを削除、文のランダムサンプリング
- サブタブReplecement：

置き換え前の文字列|置き換える後の文字列

茶筅やJUMANの形態素解析結果を＜＞タグ形式に置き換えることが可能

- サブタブParenthesis Normalizer

カギ括弧の中の文字列を削除する

- サブSentences Randomizer

テキストから文をランダムサンプリングして、複数のファイルに分割する

タグつきコーパスの集計

タグ付きデータ(Tagged text)の集計処理

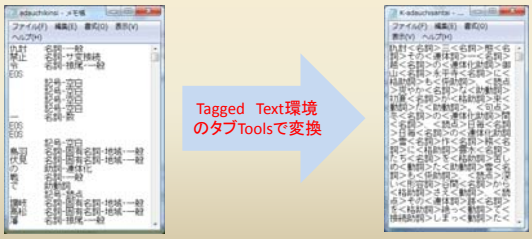
- 平文章に自由にタグを付けたコーパスを集計
- タグ記号は、日本語においては全角の＜＞
- ＜＞の中には自由に文字列を記述

どういう機会に、再び妄念に＜＞囚われ＜type5＞るかもしれない。
渇望があったように、＜＞思われ＜type2＞る。

仇討＜サ変名詞＞禁止＜サ変名詞＞令＜普通名詞＞ー＜数詞＞鳥羽＜人名＞伏見＜地名＞の＜接続助詞＞戦＜普通名詞＞で＜格助詞＞、＜読点＞讃岐＜地名＞高松＜人名＞藩＜普通名詞＞は＜副助詞＞、＜読点＞もろくも＜副詞＞朝敵＜普通名詞＞の＜接続助詞＞汚名＜普通名詞＞を＜格助詞＞取って＜動詞＞しまった＜動詞性接尾辞＞。＜句点＞

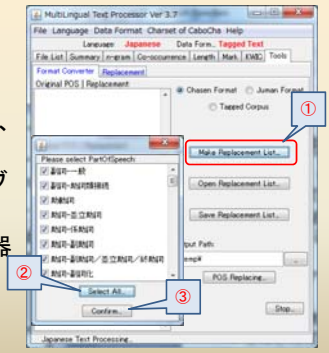
形態素解析の結果のタグ変換

- 日本語の形態素解析システム、JUMAN、ChaSen,Mecabの解析結果は、Tagged text処理環境のタブToolsで一括返還

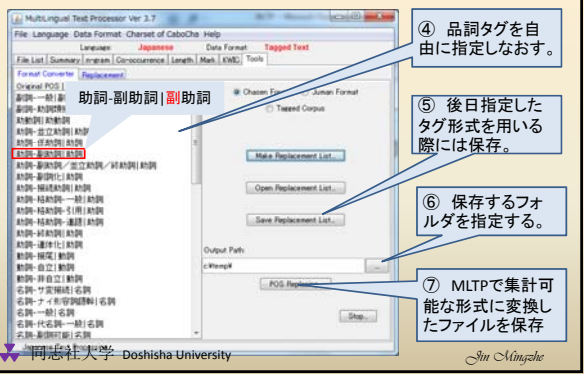


形態素解析の結果のタグ変換

- 言語の種類、データの形式を指定する
- 形態素解析結果のファイルを読み込み、確認する
- タブToolsをアクティブ化する
- 用いた形態素解析器の種類を指定する

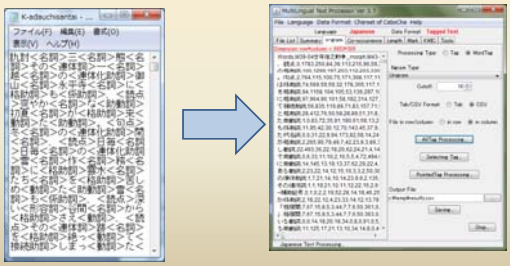


品詞情報が縦の棒|で区切られていることに注意。棒の左が、形態素解析器の品詞情報、各自が右には自由に記述。



日本語のタグ＜＞つきデータの集計

＜＞でタグ付けしたコーパスからさまざまな要素を集計することができる。

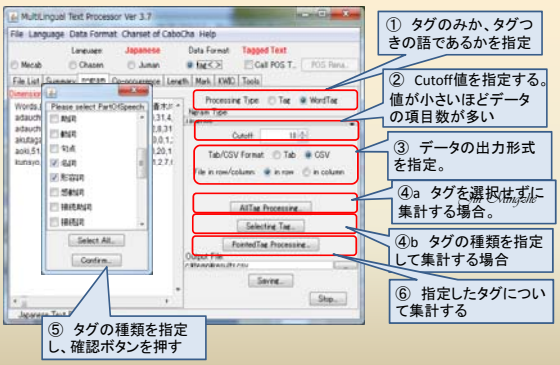


タブsummary

- FileLengthByByte,
- FileLengthByChar,
- TokenNum(述べ語数)
- TokenTypeNum(異なり語数)
- KatakanaTokenNum,
- RomajiTokenNum,
- NumberTokenNum

タブn-gram

- タグのn-gram
- タグつきの語句のn-gramを集計
- タグの種類を指定してn-gramを集計
- Cutoff値で、データのサイズを制御
- データの形式：Tab区切り、カンマ区切り、行と列の転置



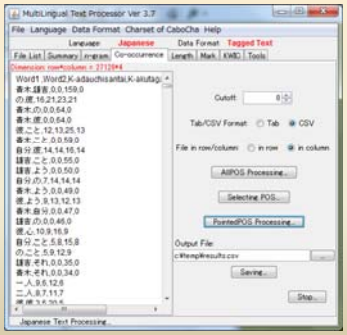
同志社大学 Doshisha University Jin Mingzhe

タブCo-occurrence

- 文の中での共起を集計する
- データのサイズ： Cutoff値で制御する
- データ形式：Tab区切り、カンマ区切り
- タグの種類をしてすることができる。
- 語句のネットワーク分析するためには,
- 出力形式をin columnにしたほうが便利

同志社大学 Doshisha University Jin Mingzhe

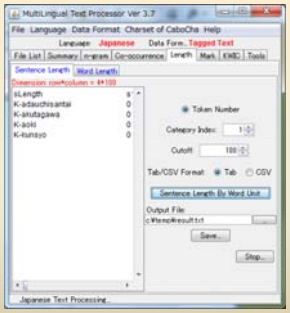
条件の指定は、n-gram と同様



同志社大学 Doshisha University Jin Mingzhe

タブLength、Mark、Tools

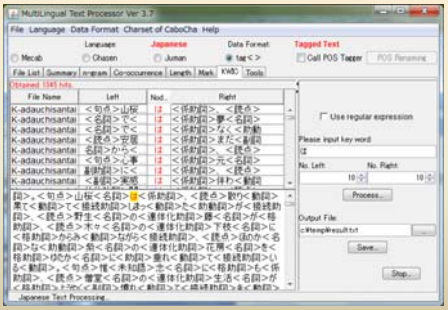
- タグ単位の文の長さの分布を集計できる。
- タブMark、KWICは、Plain textの場合と同じ
- タブToolsでは、一括置き換えなど



同志社大学 Doshisha University Jin Mingzhe

タブKWIC

- タグつきのKWIC検索。正規表現が利用可能



同志社大学 Doshisha University Jin Mingzhe

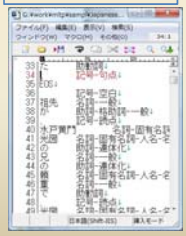
形態素解析器の結果から集計

形態素解析器ChaSen、 MeCab、JUMANの出力結果から直接集計する機能。ChaSen、MeCab、については、品詞情報が多層になっているので、品詞について名前を再定義（Renaming）して集計を行うことができる。品詞の再定義行わない場合は、第1層の品詞情報にもとついて集計する。JUMANの場合は第1層情報のみ用いるので、品詞の再定義（Renaming）を行う必要がない。

形態素解析器の結果の形式

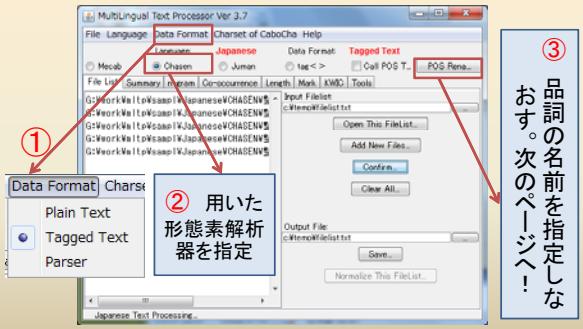
- 形態素解析器の出力結果は次の形式を基本とする。MeCabはデフォルトオプションの形式。

ChaSenの場合



JUMANの場合

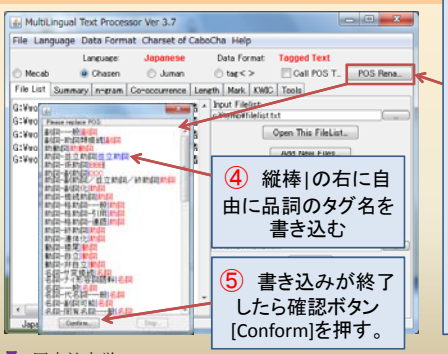
形態素解析器の結果を用いる手順



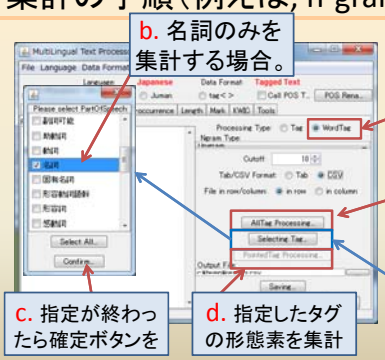
品詞名前の定義し直し（POS Renaming）

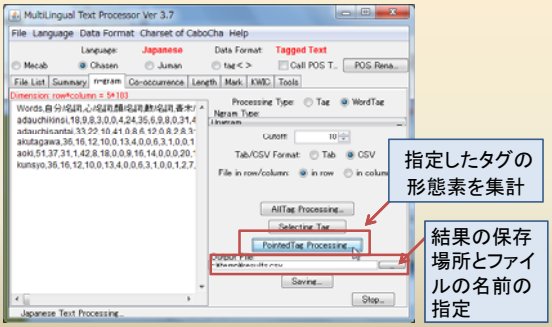
- 形態素解析器によっては品詞を第1層、第2層のように分けて出力する。例えば、名詞の場合は、一般名詞、代名詞、数詞、固有名詞などに細分類される。テキストを内容的に分析する際には、代名詞や数詞が必要ではない場合もある。このような語を集計対象から除外する一つの方法は、品詞で区別して名前をつけておくことである。

品詞名前の定義手順



集計の手順（例えば, n-gram）



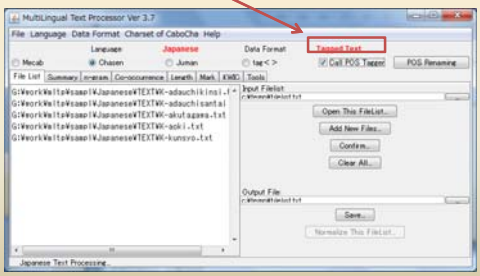


平テキストからタグ付けして集計

平テキストを読み込み、形態素解析器ChaSen、MeCab、JUMANを指定し、形態素解析を行い、タグ情報に基づいて集計を行うことができる。ただし、形態素解析器をインストールし、パスを切ることが必要である。

平テキストからタグ付き要素の集計

Call POS taagerにチックを入れる。それ以外はすべて、形態素済みのデータを用いた場合と同じである。



構文解析器CaboChaの結果を集計

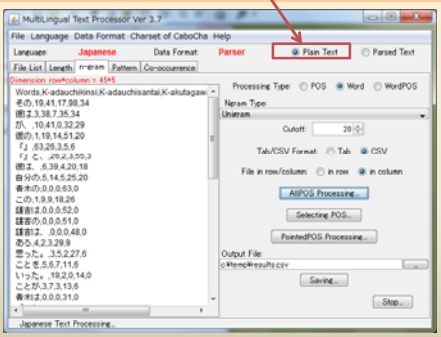
- 文節に関する集計は次の2つの方法がる。
- CaboChaの出力結果を読み込んで集計する方法。
- 平テキストを読み込み、CaboChaを呼び出し、解析を行い、集計をする方法。

文節と文節の係り受け

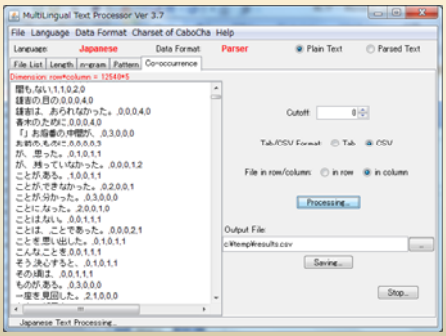
- メニューのData FormatでParserを指定
- 処理データの形式：①平テキスト、②係り受け解析済みのデータ
- Length: 文節の長さの分布、文節を単位とした文の長さの分布
- n-gram: 文節単位のn-gram
- Pattern：文節の品詞によるパターン
- Co-occurrence：文節の係り受けペア

文節のn-gram

平テキストを用いる場合



文節の係り受けペア



同志社大学 Doshisha University

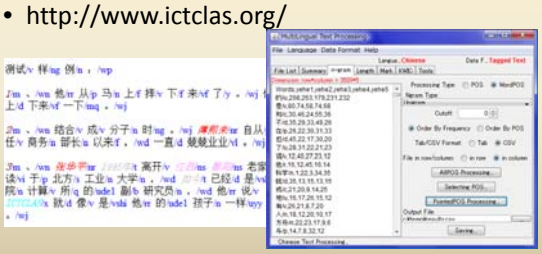
Jin Mingzhe

その他の言語

- 中国語
- 韓国語
- 英語

中国語

- 形態素解析は、中国科学院のICTCLASと北京大学基準の出力形式
- http://www.ictclas.org/

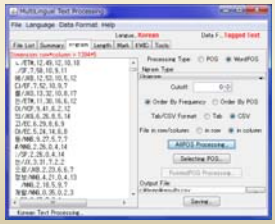


同志社大学 Doshisha University

Jin Mingzhe

韓国語

- 形態素解析はPOSTAG_SEJONG/K
- http://isoft.postech.ac.kr/Research/POSTAG/sejong/postag_sejong_k.php

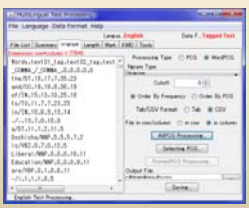


同志社大学 Doshisha University

Jin Mingzhe

英語

- 英語の形態素解析はBrill Tagger＋GoTaggerによる次の形式の結果を用いる。
- ttp://uluru.lang.osaka-u.ac.jp/~k-goto/use_gotagger.html



同志社大学 Doshisha University

Jin Mingzhe