テキストデータの統計科学入門

- 第４章 -

テキストにおける集計モデルと集計ツール

茨城大学情報工学科
菊池　裕紀

テキストにおける集計モデル

何らかの処理や統計を行う際...

単位を決めることが必要になる！

記号列を集計するときの基本

→それぞれの記号がテキスト中に現れる度数(頻度)

- 「き」という記号

「きしゃのきしゃがきしゃできしゃする。」

→ 文章中に「き」が４つ検出される

この記号の数を増やす事で集計するモデルを拡張できる！

n-gram：n個の記号の度数を集計する方法

(nは集計のため切り取った隣接する記号列の長さ)

１つの記号：「き」　　(unigram)

２つの記号：「きし」　　(bigram)

３つの記号：「きしゃ」　　(trigram)

４つの記号：「きしゃが」(four-gram)

データ集計のツール

MLTP (multi lingual text processor)

文字単位のn-gramを抽出するソフト

複数のテキストを同時に処理できる。

RMeCab (Rパッケージ)

R言語上で日本語テキストを統計的に分析する

茶器(ChaKi)

KH-coder　　etc...

MLTP

・タグ

File List

Summary

n-gram

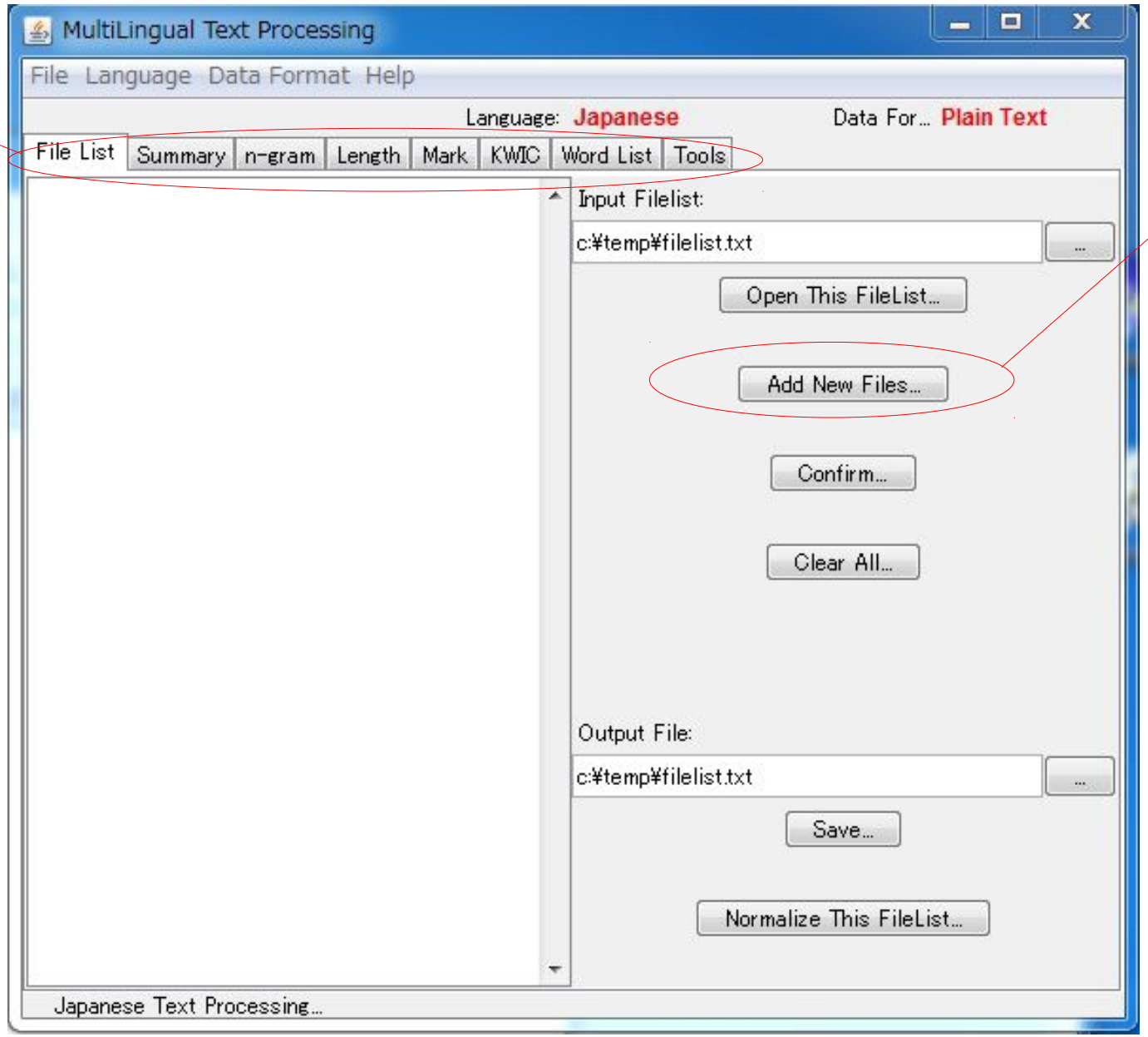
Length

Mark

KWIC

Word List

Tools



ファイルの追加