7. ローカル環境での利用について

統語・意味解析コーパス (NPCMJ) チュートリアル
@弘前大学

長崎郁

2019.5.11

はじめに

- NPCMJ のデータ（ラベル付き括弧（Bracketed tree）形式），ダウンロードして利用することもできる。
- 例えば，ラベル付き括弧形式のファイルは次のようなツールを使って検索することができる。
  - TGrep2（`https://tedlab.mit.edu/~dr/Tgrep2/`）
    コマンドラインから利用，unix 環境が必要
  - Tregex
    （`https://nlp.stanford.edu/software/tregex.html`）
    java プログラム，windows や mac でも動く
- この時間は，tregex を使って NPCMJ のデータを検索する方法，および検索結果をエクスポートする方法について紹介する。

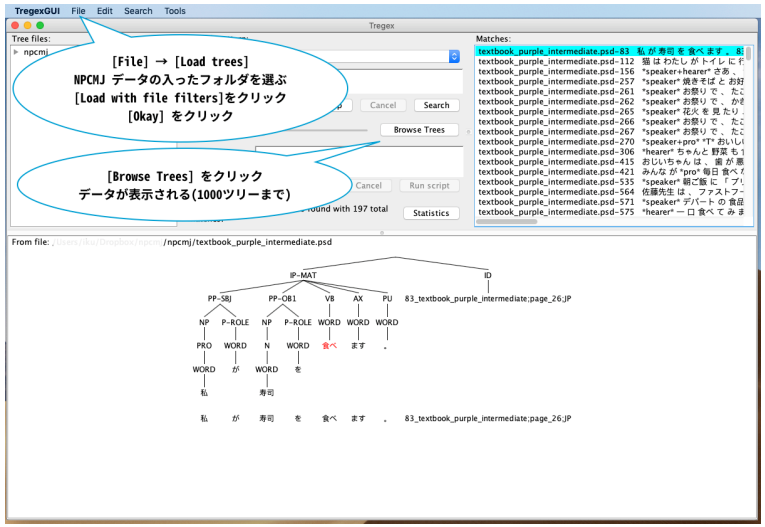
NPCMJ データのダウンロード

- データのダウンロード（以下のいずれかから）
  - トップページ
    `http://npcmj.ninjal.ac.jp/`
    Bracketed Tree ファイルをダウンロードする をクリック
  - 概要ページ
    `http://npcmj.ninjal.ac.jp/interfaces/cgi-bin/index.sh?db=npcmj&lang=jp`
    Download all bracketed trees をクリック
- ダウンロードした zip ファイルを解凍

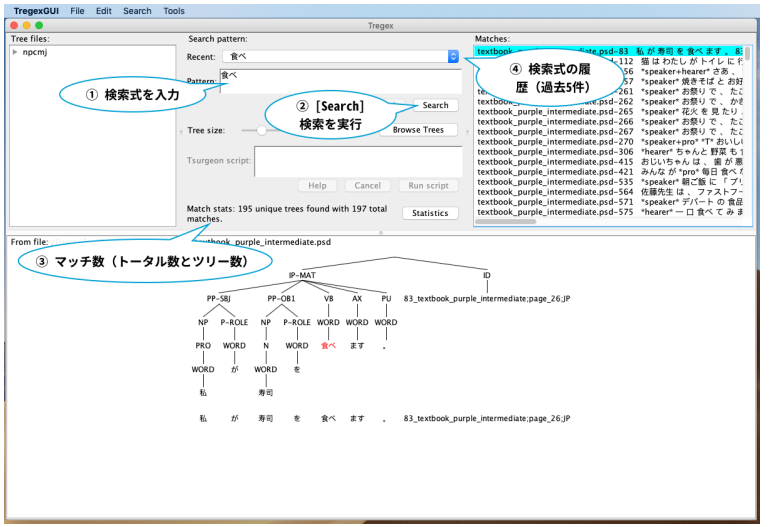
tregex のインストールと起動

- `https://nlp.stanford.edu/software/tregex.html`
  - タイトル（Tregex, Tsurgeon and Semgrex）の下の
    | Download | をクリック
  - 最新版（3.9.2）をダウンロード，適当な場所（例えば Desktop）に保存
  - zip ファイルを解答し，フォルダの中の “stanford-tregex.jar” をクリック
  - tregex ブラウザーが立ち上がらなかったら，java をインストールする。
    - `https://www.java.com/en/download/`
    - Free Java Download
    - Agree and Start Free Download
    - ダウンロードしたファイルをクリックし，指示に従う。

tregex：ファイルの読み込み



tregex：検索



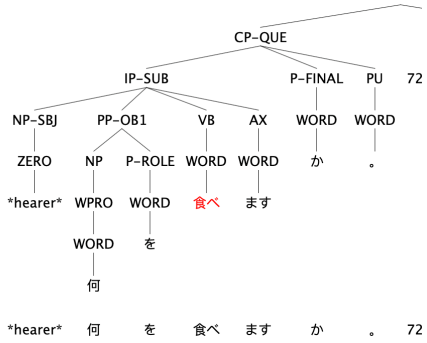
適当な語（例：食べ）を検索してみましょう

tregex：検索式（TGrep-lite との違い）

- TGrep-lite とは書き方が少し異なる。
  1. 終端ノードも非終端ノードも “/.../” や “[...]” で囲まずに書くことができる。その場合「完全一致」の意味になる。
  2. 終端ノードも非終端ノードも正規表現環境にするときは “/.../” で囲む。単に文字列を囲めば「部分一致」の意味になる。
  3. 利用できる正規表現の種類が TGrep-lite よりも多い。
     - 例えば，選言の “|” は /.../ で囲まなくても使える。
     - “食べる|食べろ|食べよ|食べ” と “/^(食べる|食べろ|食べよ|食べ)$/” は同じ (TGrep-lite は常に後者を使わなくてはならない)
     - 詳しくは [Help] を参照

tregex：検索式（TGrep-lite との違い）

4. 終端ノード（空要素をのぞく）は “WORD” によって直接支配される (空要素は “ZERO” によって直接支配)。よって，終端ノードと品詞タグの支配関係を記述する場合は，“A␣<<:␣B”，“B␣>>:␣A” を使うと良い（”␣” は半角スペース)。
   - VB␣<<:␣/^食べ/ (VB（動詞）で「食べ」で始まるもの)
   - か␣>>:␣P-FINAL (「か」で P-FINAL（終助詞))
   - か␣!>>:␣P-FINAL (「か」で P-FINAL（終助詞）でないもの)



tregex：検索結果のエクスポート

- 例：助詞「しか」の用例を検索し，エクスポートしてみる
- 「しか」を検索 → 62 total matches
  - (「しか」はすべて P-OPTR（とりたて助詞）として扱われているか，確認したい場合は以下のようにして確認すると良い。
    “しか␣!>>:␣P-OPTR”
- エクスポート
- メニュー [File] → [Save matched trees]（タグあり）
- メニュー [File] → [Save matched sentences]（タグなし）

tregex: ツリーの一部を表示・エクスポートする

- データには非常に長いツリーも含まれる。そのため検索結果を確認するのが困難な場合は，必要な部分だけ表示させることもできる。
- 例：助詞「しか」を含む節（IP）
- メニューから [TreeGUI] → [Preferences]
  [Show only matched portion of tree] をチェック
  - 検索式中のマスターノード（とそれに支配されるノード）を表示する。
- 助詞「しか」を支配する助詞句（PP）をもつ節（IP）を検索
  “/^IP/␣<␣(/^PP/␣<␣(P-OPTR␣<<:␣しか))”（55 total matches）
- 上述と同じ方法でタグあり，タグなしのデータをエクスポートすることができる。