コーパスツール講習会

KWIC Finder の使い方と正規表現

NICT/京都大学　金丸 敏幸

本日の予定

- KWIC Finder の紹介
- 正規表現とは
- KWIC Finder を使った事例検索

KWIC Finder とは

- KWIC Finder
  - もともとは複数テキストからの文字列検索ソフト
  - 検索結果に KWIC 形式を採用
  - 正規表現を使用した検索に対応
  - 最新バージョンでは各種検索エンジンに対応
  - テキストファイルの検索のみであれば，フリー
    - 入手先：http://www31.ocn.ne.jp/~h_ishida/KWIC.html

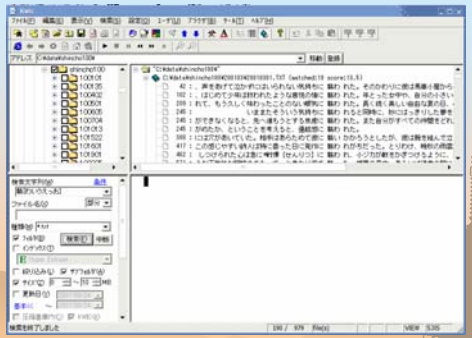
KWIC 形式とは

- Key Word In Context，つまり検索したい対象だけでなく，検索語の前後文脈まで表示する．
  - 最近の web 検索はほとんどが KWIC 形式
  - 例：

    チョムスキーの生成文法が言語能力（Competence）を調査対象とするのに対し、**コーパス**言語学は言語運用（Performance）に焦点を当てる。ある言語事象をリサーチクエスチョン（仮説）として設定し、それを**コーパス**データに基づいて記述する。...

    （google による「コーパス」の検索結果の一部より）

KWIC Finder の画面



正規表現とは

- 正規表現とは，ある要素（文字とか単語など）の前後に特定のパターンを持ったものが連接していく形の記述
- これを「正規文法」という．
  - 例：abababab → ab + ab + ab → (ab){3}
  - 正規表現では記号が意味を持つ
    - () → ()の中の表現をパターンとしてまとめる
    - {} → {} の中の数字の数だけ直前表現を繰り返す

正規表現の記号

項目	意味
[]	キャラクタクラス
()	パターンをグループ化する
^	行頭
$	行末
.	改行を除く任意の1文字
*	直前のパターンの0回以上の繰り返し
+	直前のパターンの1回以上の繰り返し
?	直前のパターンが0回または1回現われる
|	パターンの論理和
¥	エスケープ文字

記号の説明

- [] … キャラクタクラス
  - 一文字単位で指定するときに使用.
  - abc, abb, aba とあるとき, ab の後ろには, a, b, c のどれかがくるので, これらをまとめるときには, ab[abc] のように記述する.
  - 便利な方法
    - ひらがな全部:[ぁ-ん]
    - 漢字全部:[亜-黒]
    - 「-」は, その間にあるもの全てを表す.

記号の説明

- () … パターンをグループ化する
  - 同じ種類の表現をまとめたいときに使用.
  - 今日, 明日, 明後日 のどれかを検索したいときは, []は使えない. [今明明後]日では,「後日」が検索されてしまう.
  - そこで, それぞれをまとめて(今日|明日|明後日)とする.「|」は区切りの記号と覚える.
  - もう少しまとめて「(今|明|明後)日」としてもよい.

正規表現と自然言語

- 自然言語も正規表現で(一応)表現可能
  - 生成文法はそれを表したもの
    - ただし, 厳密には生成文法は文脈自由文法(2型文法)で, 正規表現で表される正規文法(3型文法)よりは, より複雑な記述が可能.
    - 正規文法は前後への連接だけだが, 生成文法は埋め込みが可能.
  - だから, 言語学者なら正規文法は使えるはず?

日本語と正規表現

- 日本語の一部は正規表現で記述しやすい
  - 動詞の活用
  - 例:五段活用=語幹+あ, い, う, え, お
    - 語幹を「要素」, 活用部分を「パターン」とすると…
    - 「書く」の場合, **書[かいきくけこ]** となる.
  - 助動詞の接続
    - 「見られている」「見つけている」「見始めている」…
    - **問題**:動詞「見る」とアスペクト「ている」の間に出現する表現を収集するには?

答え

- まず, 共通の要素をまとめる.
  - **見**られ**ている**, **見**つけ**ている**, **見**始め**ている**
  - → 「見」と「ている」をまとめる.
- 間に(「見」の後ろに)来るパターンを記述
  - 「られ」「つけ」「始め」… 他にも「見ている」などもあるので, パターンは「0文字以上の文字」.
- 従って,「**見.*ている**」となる.
  - ただし, 多数のノイズも収集されます(後述).

最長一致の原則

- 「見.*ている」でノイズが収集される理由
  - 「．＊」が表しているのは，「0文字以上の文字」なので，間に入る文字は**何文字でも**よい．
  - 正規表現の場合，一番長い表現を正解とする．これを「最長一致」という．
  - 従って，「**見**つけられ**ている**」のような収集したいもの以外にも，「**見**ていた男が立っ**ている**」のようなものまで検索してしまう．
  - 一番短い表現「最短一致」を検索できるのもある．

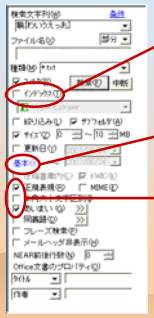
KWIC Finder による検索

- 設定の変更
  - 標準だと正規表現やあいまい検索が使用できないこともあるため．
- 比喩表現の収集
  - 通常のキーワード検索で検索してみる．
- 検索する内容を正規表現で記述
  - 特定の動詞を検索するため，活用を正規表現で記述してみる．

KWIC Finder の設定

- 検索オプションの変更
  - 正規表現とあいまい検索を有効に
    - 左下の部分の「詳細」をクリックして，詳細設定を表示．
    - 「正規表現」と「あいまい」にチェックします．
  - 検索エンジン使用の停止
    - 左下の部分にある「インデックス」のチェックを外す．
    - 検索エンジン利用は，大量のファイル検索には高速なので便利だけれども，コーパスの検索程度であれば，特に使用する必要はありません．

設定後のメニュー



- インデックス（検索エンジン）のチェックを外す
- 拡張設定を表示する
- 正規表現とあいまい検索をチェックする

検索実習

- 比喩表現の収集
  - 人生は旅である．…など，「人生」に関係ある表現を収集してみます．
- 特定の動詞の収集
  - 「襲う」などの動詞表現を収集してみます．

比喩表現の収集

- 一番簡単な検索：「人生」だけ
  - 「人生から」「人生について」など余計なものも収集してしまう．
- 人生は～だ(である)：「人生は.*[だで]」
  - 「彼等の人生は」「私の人生は」などが検索されている．
- 「人生」の前に「の」が来ないもの：
  「[^の]人生は.*[だで]」
  - [^○△]は，○や△を含まない表現

データの利用

- Excel などで利用する場合
  - メニューの「ファイル(F)」→「検索一致リスト」から「タブ区切りデータ(Y)」を選択.
  - ノートパッド(もしくはエディタ)が起動して，検索結果の一覧が表示されるので，全てを選択してコピー.
  - Excel を起動して，コピーした内容をペーストする.
    - 一行目は項目名で「ファイル名，行，位置，先行文脈，直前語，キーワード，後続文脈」

Excel にコピーした状態

データを見やすくする

- 「並び替え」をする
  - 一行目の「キーワード」のセルを選択して，「A→Z 昇順で並び替え」をクリック.
- 得られたターゲット
  - 「運命」「空虚」「形成」「賭」「夢」「地獄絵」…
    - ロクなのがないですね(笑). 小説だからでしょうか.

並び替えをした状態

特定の動詞の収集

- 「襲う」の検索
  - 活用をパターン化する.「襲-」+「わ，い，う，え，お，っ」
  - 「襲[わいうえおっ]」
- 直前格を限定して収集
  - ヲ格を収集:「を襲[いうえおっ]」「わ」は，受け身で出てくる活用なので省略.
  - 比喩表現の収集と同じ手順で Excel にコピー.

データを見やすくする

- 「並び替え」をする
  - 一行目の「直前語」のセルを選択して，「A→Z 昇順で並び替え」をクリック.
- 得られた対象
  - 「我々」「海賊船」「彼」「彼女」「首相官邸」「真珠湾」「日本」…
    - 人や場所が多く見つかることが分かります.

並び替えをした状態

まとめ

- KWIC Finder の説明
- 正規表現について
- 検索実習

- おまけ：KWIC Finder の限界
  - 品詞が使えない
  - 最短一致が使えない
  - →これを解決するのが KH Coder