BNCwebによる汎用コーパスBNCの使用法

7113107
言語文化学部朝鮮語専攻
森　千紗

目次

1. BNCwebの概要
2. 利用者登録
3. 基本的なコンコーダンサーとしての使い方
    1. 入力方法
    2. 検索結果の表示
    3. ソート機能
4. 検索対象となるデータを絞り込む
5. 検索結果の分布を確認する
6. 任意の語句のコロケーションを調べる
7. 特殊な検索
8. 活用例：特定のサブコーパスを分析する
    1. サブコーパスを作る
    2. サブコーパスの語彙頻度リストを作る
    3. サブコーパスのキーワードを抽出する
9. おわりに

1. BNCwebの概要

- ウェブブラウザーを利用してBritish National Corpus(BNC)を検索できるサービス
- 手軽な利用方法：ランカスター大学（イギリス）で提供されているサービスを利用

- 長所：シンプルで分かりやすいインターフェースでありながら高機能
  - Ex.) 英語の使用における男女差や年齢差を調べる
    サブコーパスの語彙頻度表を作成する
- 短所：機能制限あり
  - Ex.) 検索結果の表示上限が5,000件（5,000件を超える結果があった場合、結果はランダムに抽出）

2.利用者登録

1. http://bncweb.lancs.ac.uk/bncwebSignup/user/register.php にアクセス
2. 氏名、所属、メールアドレス、国、ユーザー名、パスワードを入力

   →アカウント即時発効、確認メールが届く
3. http://bncweb.lancs.ac.uk/ にアクセス

   →ユーザー名とパスワードを入力

3. 基本的なコンコーダンサーとしての使い方

1. 入力方法

- 画面トップの検索ボックスに語句を入力して[Start Query]ボタンをクリック

※注意

- 句読点、所有格の's、-sで終わる複数形に付く「'」、縮約形の各要素はそれぞれがひとつの単語として扱われる
  →前に半角スペースを入れて検索
- ? • + , : @ / ( ) [ ] { } _ - < > |
  ↑検索時に特別な意味を持つ記号として使われる
  →そのままの意味で使いたい場合には前にバックスラッシュ(「＼」若しくは「¥」)を付ける必要あり
  Ex.) they've → [they△'ve]
  won't he? → [wo△n't△¥?] （△は半角スペース）

3. 基本的なコンコーダンサーとしての使い方

2. 検索結果の表示

- 検索式にマッチする箇所の数
  - —“〇〇hits”
- 検索式にマッチする箇所を含むテキストサンプルの数
  - —”〇〇 different texts”
- 100万語あたりの頻度
  - —“〇.〇 instances per million words”

+ マッチした件数が5,000を超える場合

  —“thinned with method *random selection* to 5000 hits”

3. 基本的なコンコーダンサーとしての使い方

2. 検索結果の表示

- [Show KWIC View]
  —検索式にマッチした部分が中央に揃う
- [Show in random order]
  —ランダムな順番の表示に切り替え可能
  (表示順は初期設定ではテキストサンプルのID順)
- [User setting]
  →[Default view:]
  [Default display order of concordances:]
  —検索結果の表示形式と表示順の初期設定を変更できる
- Filename —当該の用例を含むテキストサンプルのIDを示す
  マウスカーソルを置くと詳細な情報が表示される

3. 基本的なコンコーダンサーとしての使い方

3. ソート機能

- 検索結果画面のドロップダウンリスト（初期値では[New Query]）で[Sort]を選択し、[Go!]をクリック

→1語右がアルファベットの昇順(ABC順)でソートされる

- 詳細設定
  - [Position:] ―指定する位置
  - [Tag restriction:] ―品詞の制限

  →[Submit]をクリック

※新しい検索式で検索を行いたい場合

―ドロップダウンリストを[New Query]に設定

→[Go!]をクリック

4. 検索対象となるデータを絞り込む

- 書き言葉または話し言葉のみに制限したい場合

  [Restriction:]（検索ボックスの下）

  →[Written Texts]または[Spoken Texts]に変更

- 更に細かい条件を指定する場合

  [Written / Spoken restriction]（初期画面の左側にあるメニュー）

  →検索対象に含めたい属性をチェックボックス形式で選択

5. 検索結果の分布を確認する

- 分布
  - 話し言葉・書き言葉どちらに多いのか
  - どのジャンルで多いのか
  - どのような発話者の属性（性別、年齢、社会階層など）に多いのか

- 検索結果画面から[Distribution]を選択し、[Go!]をクリック

6. 任意の語句のコロケーションを調べる

1. 検索結果画面のドロップダウンリストで[Collocations...]を選択→[Go!]をクリック
2. 3つのパラメーターをそれぞれ指定し、[Submit]をクリック
   - [Include lemma information]
     —"yes"にすると前後の単語の見出し語の情報を含めて集計される
3. 品詞、位置、ソート方法を指定
   - 初期値では対数尤度比(log-likelihood)の値の高いものから並べられている

※この機能で得られるデータは、あくまで一定の範囲にある単語を集計したものに過ぎない
（＝信頼できるとは限らない）

7. 特殊な検索

- 特殊な記号を使うことで、通常では出来ない語句・単語の検索ができるようになる
- 検索方法の概略
  - —Simple Query Syntax helpを参照
    （PDFファイル、初期画面の検索ボックス下）

7. 特殊な検索

特殊な記号を用いることで可能な検索方法

- 「任意のx語」「単語中の任意のx文字」
  →ワイルドカード（+, *, ?）の使用
- 見出し語形による検索→{ }の使用
- 品詞の指定→「_」（アンダースコア）+品詞タグの使用

—BNCで使われているCLAWS5のタグセット、BNC独自の簡易品詞タグも使用可能

※上記はそれぞれ併用が可能

8. 活用例：特定のサブコーパスを分析する

1. サブコーパスを作る

1. [Make/edit subcorpora]（初期画面左のメニュー）を選択
2. [Define new subcorpus via:]で[Written metatextual categories]を選択→[Go!]をクリック
3. [Genre:]内から作成したい分野のリストを選択
   →[Get text IDs]をクリック
4. [include all]（右上）を選択→[Add]をクリック
5. サブコーパスに付ける任意の名前を入力して
   [Submit name]をクリック

8. 活用例：特定のサブコーパスを分析する

2. サブコーパスの語彙頻度リストを作る

1. 再度[Make/edit subcorpora]を選択
2. 作成したサブコーパスの[Frequency list]が[Compile]と表示されている場合→[Compile]をクリック
   →語彙頻度データが作成され、[Compile]が[Available]という表示に変わる
3. [Frequency lists]（初期画面左のメニュー）を選択
4. [Choose one or more lemma classes:]（[Headword or lemma frequencies]の下）で[no restrictions]を選択
5. [Range of texts:]で[Subcorpus: 名前]を選択
6. [Show list]をクリック

8. 活用例：特定のサブコーパスを分析する

3. サブコーパスのキーワードを抽出する

分野別のコーパス：上位100単語以下の頻度では分野毎に特徴的な単語が多く見られるようになる

1. 初期画面左にあるメニューで[Keywords]を選択
2. [Select frequency list 1]を
   [作成したサブコーパスの名前]に変更
3. [Compare:]を[Lemmas]に変更
   →[Calculate keywords!]をクリック

9. おわりに

- BNCwebでは、バランスを考慮したサブコーパスから構成されているBNCの特徴を最も活かした分析が出来る
- 英語の使用域ごとの分布を調べるのに適する
- コロケーションのきめ細やかな分析が可能
- 話し言葉のデータのうち約550万語の音声を聴くことが出来る

教材研究や自己研鑽に計り知れない恩恵