コンコーダンサ、
頻度、
コロケーション

おしながき

- コンコーダンサの紹介
- 頻度
- コロケーション

コンコーダンサ

コンコーダンサ

- コーパスを検索するためのソフトウェア
  - Keyword in Context (KWIC) 形式で表示
  - 頻度表の作成、コロケーションの計算機能など
- 個人的にはあまり使っていません。

英語の代表的なコンコーダンサ

- オンライン
  - COCA など
  - Sketch Engine

英語の代表的なコンコーダンサ

- ローカル
  - WordSmith (Win) (有料)
  - AntConc (Win/Mac) (フリー)
  - CasualConc (Mac) (フリー)

英語の代表的なコンコーダンサ

- 英語用のコンコーダンサは単語単位で検索するようになっているので、日本語を扱いたい場合は、ChaSen などを用いて単語間をスペースで区切る必要がある（文字コードの問題が生じる場合もある）

日本語に特化したコーパス検索ソフトウェア

- ひまわり（国語研。ひまわり向けに作られたXMLデータを用意する必要がある）(Win/Mac)
- Chaki.NET
- KH Coder（形態素解析と一体化、統計処理機能もついた高機能ソフト） (基本的にWindows)

頻度

タイプ頻度（異なり語数）とトークン頻度（のべ語数）

- タイプ頻度（異なり語数）
  - 語彙の豊富さ、パターンの生産性などに対応
- トークン頻度（のべ語数）
  - コーパス全体の規模、特定の語の用例の数の豊富さに対応

語彙項目 (lemma) と語形 (wordform)

- 英語であれば *give, gave, given, gives, giving* のような違いを別々に数えるか、同じ *give* という語の現れとみるか
- 日本語であれば「読ま」「読め」「読ん-（だ）」のような違いを別々に数えるか、同じ語として数えるか

語彙項目 (lemma) と語形 (wordform)

- 英語の場合、語形の違いを吸収するためのデータが必要
  - http://www.laurenceanthony.net/software/antconc/
- 日本語の場合、ChaSen などで形態素解析を行ったうえで、語彙素を数えることになる

百万語あたり (PMW)

- コーパスの規模が異なるため、そのままでは比較できない場合に有用
  - BCCWJ (NINJAL-LWP) の例
  - 全体語数 - http://nlb.ninjal.ac.jp/site_media/pdf/NLB.manual.v.1.30.pdf
- 各コーパスが100万語に満たない場合はミスリーディングになり得るので注意

コロケーション

コロケーション

- 語と語の結びつき。とくに、**統計的に、語Aと語Bが偶然よりも高い頻度で一緒に現れること** NINJAL-LWP で例
- コロケーションは、語の意味を詳しく調べたいとき（とくに類似した語の微妙な違いを調べたいとき）、教育目的などに特に有用
- コロケーションが理論的にどういう意味を持つかは難しい
  - 意味的な関連性を示す？
  - イディオム性を示す？ cf. 心理的実在性

ダイス係数

$$\frac{2*(\text{語A,Bの共起頻度})}{(\text{語Aの頻度})+(\text{語Bの頻度})}$$

cf.「抱く」「抱える」論文

Tスコア

$$\frac{1}{\sqrt{\text{共起頻度}}}\left(\text{共起頻度} - \frac{\text{語Aの頻度*語Bの頻度}}{\text{コーパス総語数}}\right)$$

相互情報量 (MI)

$$\log_2 \frac{\text{共起頻度} * \text{コーパス総語数}}{\text{語Aの頻度} * \text{語Bの頻度}}$$

コロケーションの指標には性質の違いがある

- Tスコア
- ダイス係数
- MIスコア

実際に頻度が高いものに高いスコアが出る傾向

↓

頻度は低くても、確実に一緒に出てくるものに高いスコアが出る傾向

4種類の共起 (Stubbs 2002)

- **コロケーション**：特定の語と語が共起すること
- **コリゲーション**：語が特定の品詞と共起すること
- **優先的意味選択** (semantic preference)：語が特定の意味カテゴリに属する語と共起すること
- **意味的韻律** (semantic prosody)：語が特定の話者の態度などと共起すること

コリゲーションの例

- deeply, greatly, highly の違い (p.181)
- 「状態」と「状況」の違い (新屋2010) (p.177)

宿題

- 英語または日本語の作品を選んで
  - その作品ののべ語数、異なり語数を調べてください
  - 2つの語を取り上げ、コロケーションの指標を用いて、それぞれ結びつきの強い語のリストを作り、2つの語の用法の違いについて論じてください。