Unit C1 Collocation and pedagogical lexicography Case Study 1

尹美(インメイ)

C1.1 Introduction

学習者用辞書でコロケーション情報を充実
させるためのBNC Web の使い方

１９８７　Cobuild- OALD,LDOCE,CIDE など

コーパス活用例：
１．頻度情報
2.定義の際の語彙の簡略化（defining vocabulary）
３．コーパスからの例文の選択

コーパス例文のみ：Cobuild
書き直しと組み合わせる：OALD,LDOCE
上記と同様、それ以上に期待されているのがコロケーションへの応用

次のセクション以降ではコロケーション情報の抽出と、分析に使える様々な
統計法を扱う

C1.2 Collocation Information

- BNC Web からコロケーションの情報を引き出す（Sweet と共起する名詞）

C.1.2.1Collocation analysis using BNC Web

手順（ p 210 - 211 ）

8 →範囲を + 1 ~ + 3にする

8 →リストとして出できたのがsweetと本当に共起しているか確認する必要がある

C.1.2.2 Collocation Statistics

- Raw frequency
  最も基本的なもの
  BNCでそもそも高頻度のものがリストの上位に来てしまう
- Observed/expected score
  中心後と共起語の全体の頻度を考慮する尺度
  偶然によるものとどのぐらい結果が違っているかを示す
- Z-score
  全体的な頻度を調整して、それから予測されるものよりどのぐらい頻度が高いのかを示す
  前提としてデータが正規分布してなければならない
  低頻度の語が上位にきてしまう

- Log-likelihood（LL）
  データが正規分布していなくてもよい（小さいデータでも使える）
  低頻度語と高頻度語両方がリストに入ってくる（比較可能）

- Mutual Information (MI)
  LLほど厳格ではないが広く使われている
  低頻度語が上位に来てしまう

  低頻度語は、一般的な目的の辞書手は扱われるべきだが、教育目的の辞書では基本的なコロケーションのほうが重要

- MI3 score
  高頻度語のより重きを置く（Cubing）
- Log-Log formula
  MIの弱点を調整したもの
  高頻度語がリストの上位に入ってくる

C1.3 Using Corpus Data for Improving a Dictionary Entry

Longman Dictionary of Contemporary Englishの第一版と第４版を比較

C1.3.1 Focusing on high-frequency word

LDOCE1と LDOCE4を見比べる（P221)

行数が多くなっている

重要な語の扱い方が二つの版の大きな違い

高頻度語により焦点を当てるという編纂方針

例：（S2）（W3）=top 2000 in the spoken corpus/top 3000 in the written corpus

C1.3.2 Providing Examples

- 用例を供給することもコーパスデータが果たす大きな役割
- 用例はきわめて重要（統語、意味，語用的な情報を示す）
- LDOCE1とLDOCE4を見比べる（P222)

  LDOCE４ではComplete sentenceの用例を載せている（見出し語が使われる文脈を示すので、学習者により有益）

C1.3.3 Providing collocation information

- コーパスデータにより供給されるコロケーションの情報がより重要
- LDOCE1とLDOCE4を見比べる（P 223)

  LDOCE1の用例は短く、adjective+nounのパターンだけ

  LDOCE4の用例は長く、完全な文で与えられている。

・手順（P223の4～）

５→範囲を＋／－３にする（Sweetの前後両方の情報が必要だから）

- コロケーションの情報と辞書の用例を照らし合わせる

  LDOCE４は上位３０の78.57%、上位１０の42.86%のコロケーションを用例に含んでいる

  LDOCE１では上位10の33.33%、上位50の 62.5%

  用例にコロケーション情報がいかされている

- Further Study

  P.２２５にある三つのコロケーション辞書の妥当性の調査

　ただし、特定の辞書が他の辞書よりよいということを強調しすぎるべきではない

辞書は、target userに応じて異なるタイプの情報を掲載しているので、それぞれ長所、短所がある