第3章
汎用コーパス概観

2015/4/14

発表者 : 及川　ほのか

本発表の流れ

1. 汎用コーパスとは
2. Brown系コーパス
3. British National Corpus
4. WordbanksOnline
5. Corpus of Contemporary American English (COCA)
6. Sketch Engine
7. 新聞・雑誌・その他のアーカイブ
8. コーパスとしてのウェブサイトとWebCorp Live
9. まとめ・疑問点

1．汎用コーパスとは

コーパス

汎用コーパス

様々な研究目的に広く対応できるよう，多種多様なソースからバランスよく集められ，電子化されたもの。「均衡コーパス」(balanced corpus)と呼ばれることもある。

特殊目的コーパス

特定の言語研究の目的に限定したもの。

2. Brown系コーパス

Brownコーパス[1961]

- 当時の米語の書き言葉の実態を反映
- 世界初の電子テクスト集
- 「第1世代コーパス」の原型
- Sketch Engineにて無料で使用可

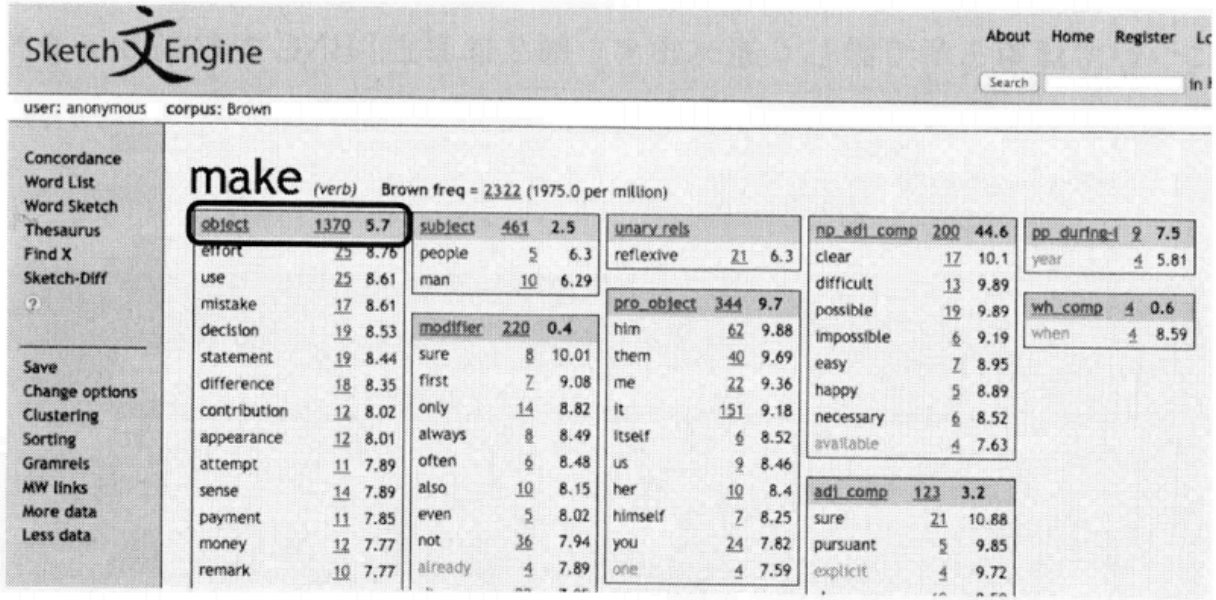

図1　Brownコーパスにおけるmakeのコロケーション

2. Brown系コーパス

コーパス[年]	言語の種類	サイズ	利用形態
Brown[1961]	アメリカ英語	各100万語	• The ICAME Corpus Collection (CD-ROM) • Sketch Engine
Lancaster-Oslo/ Bergen Corpus(LOB) [1961]	イギリス英語		
Freiburg Brown(Frown) [1990年代]	アメリカ英語		
Freiburg LOB(FLOB) [1990年代]	イギリス英語		
AmE06 [2005-2007]	アメリカ英語		Sketch Engine
BE06 [2005-2007]	イギリス英語		

3. British National Corpus (BNC)

British National Corpus（BNC）

[1990年代]

- 「第2世代コーパス」の1つ
- 現在、多くの言語学者に使用されている

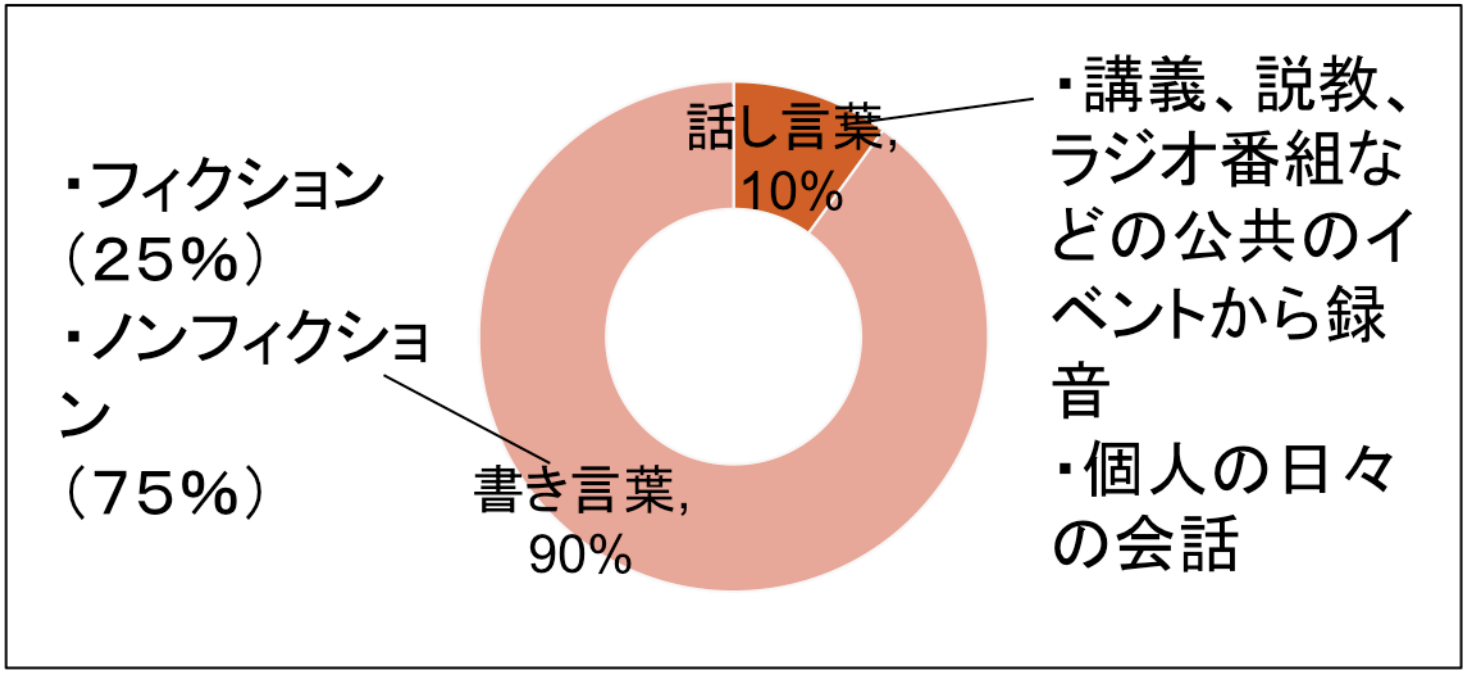

図2　BNCの話し言葉、書き言葉の割合とその内容

3. British National Corpus (BNC)

コーパス[年]	言語の種類	サイズ	利用形態
British National Corpus（BNC） [1990年代]	イギリス英語	1億語	• BYU-BNC • BNCweb • Intellitext • Phrases in English

それぞれの利用形態により、使い勝手が異なる。
⇒用途に合わせて使い分けるとよい。

4. WorldbanksOnline

WorldbanksOnline
[1985-1998]

- Bank of English(BoE)のうちから一部を抽出し、再構成したもの。

コーパス[年]	言語の種類	サイズ	利用形態
Bank of English(BoE) [1980~]	• イギリス英語 • アメリカ英語 • オーストラリア英語	6億5千万語	-
WorldbanksOnline [1985-1998]		5,400万語	• WorldbanksOnline • 小学館コーパスネットワーク

書籍、新聞、話し言葉、放送原稿など。
話し言葉(全体の約25%)は放送原稿＋自由会話

5. Corpus of Contemporary American English（COCA）

- 1990年以降現在まで1年ごとに2000万語のセクションで構成されている。
- 無料で提供されるコーパスの中で最大、かつアメリカ英語コーパスとしては唯一の大規模均衡コーパス。
- データはサイズの等しい5つのジャンル（話し言葉、フィクション、大衆雑誌、新聞、学術誌）に分けられている。

コーパス[年]	言語の種類	サイズ	利用形態
Corpus of Contemporary American English(COCA) [1990~]	アメリカ英語	4億6千万語	• COCA（CORPUS.BYU.EDU）

6. Sketch Engine

- 数多くのコーパスへのアクセスや多機能な分析ツール、コーパス作成支援などのサービスを提供する有料サイト。
- Word Sketchなど、非常に多機能。
- 膨大なデータから検索して用例を抽出できる。

⇒低頻度の語句の用例を見つけることも可能。

コーパス	言語の種類	サイズ	利用形態
ukWaCや enTenTen12など 100以上	変種を含め50以上の言語	ukWacは15億語、enTenTen12は100億語など	Sketch Engine

7. 新聞・雑誌・その他のアーカイブ

- TIME, CORPUS.BYU.EDU,
  New York Times
  The Oxford Text Archive,
  Project Gutenberg
- 語法の確認や適当な用例を探すのに役立つ。
- テクストからコーパスを自作可能。

8. コーパスとしてのウェブサイトと WebCorp Live

検索エンジンを使用し、インターネット上にある情報を丸ごと言語資料として利用。⇒情報の宝庫

[問題点]⇒[活用するためには]

1. ウェブサイトは常に変化し続けている

⇒必要な情報・データは記録・保存

2. データの信頼性の問題

⇒「検索オプション」等の利用

3. コーパスツール機能がない

⇒WebCorp Liveの利用

9. まとめ

調査の目的に応じて、正しいサイズのコーパスを使用することが検索の成功に繋がる。

⇒それぞれのコーパスの特徴を知っておくべきである。