Unit A1 Corpus linguistics: the basics

学籍番号　7413104
発表者　山崎加奈
2015/4/15

Unit A1　コーパス言語学の基礎

- A1.1　導入
- A1.2　コーパス言語学の歴史
- A1.3　近代言語学における“コーパス言語学“
- A1.4　なぜ言語処理をする際にパソコンを使うのか
- A1.5　Intuition-based approach vs. Corpus-based approach
- A1.6　なぜ言語学の独立した一部門ではなく、方法論なのか
- A1.7　Corpus-based approach vs. Corpus-driven approach

A1.2 コーパス言語学の歴史

- 1950年代後半～　コーパス方法論の大部分が、主流からはずれているとして厳しく批判される

- 当時のコーパス “shoesbox corpus”
- ・データ蓄積のために、紙片を箱に詰めていた（× PC）
- ・非常に規模が小さい
- ・代表性がない
- ・少数の言語学者の音声学研究・文法研究に使われていた
- →大量の言語データを人の目と手で順番にそろえるのは、事実上ほぼ不可能

A1.2 コーパス言語学の歴史

- 1960年代前半　　世界初英語コーパス”Brown Corpus”
- →以後、コーパスの規模拡大・コーパス準拠の研究増加

- 1980年代初期～　“コーパス言語学”という単語が使われ始める
- 20世紀初頭　ベーシックコーパス方法論が、言語学の中に広がっていく

- 技術の発展
- 特に、比較的低コストで大規模なストレージと高い処理能力を持ったPC
- →大規模コーパスが利用可能に

- コーパスとコンピュータ技術の融合
- →コーパス方法論への関心に再び火をつけた

A1.2 コーパス言語学の歴史

- 現代
- ・コーパス言語学は、新しい研究分野を生み出し重要性を与えている
- ・コーパスなしでは、多くの研究はなされなかった
- ・コーパスは、言語学すべての分野に革命を起こしたといえる

- チョムスキー以前：アメリカ構造主義（言語を最小単位で考える）
- →文構造、用例、コーパスなどをデータとして研究に使用
- チョムスキー以後：innate capacity（生得能力）のモデル化が研究テーマに
- →用例やコーパスなどのデータを見なくなった

A1.3 近代言語学における“コーパス言語学”

- Leechの定義
- コンピュータ・コーパス：やみくもではなく、特定の目的に沿って集められたテキスト資料。言語や文書の代表として集められる。

- Sinclair(1996)　Leechの定義、特に代表性の重要性を強調

A1.3 近代言語学における“コーパス言語学”

- 近代言語学におけるコーパスの特徴
- １．machine-readable
- ２．authentic texts
- ３．sampled
- ４．representative

- →コーパスは、(1) 機械可読式で、(4) ある特定の言語の代表として、(2) 信頼できるテキストから(3) 標本抽出されたものの集合

A1.3 近代言語学における“コーパス言語学”

- Lancaster Corpus of Abuse (LCA)：British National Corpus(BNC)をもとに作られたコーパス。BNCのサブコーパス。
- サブコーパス：コーパスをもとに作られたコーパス

- サブコーパスがコーパスといえる理由
- １．コーパスの基準４つを満たしている
- ２．すべてのコーパスが均衡コーパスではない（特殊コーパス≠均衡コーパス）
- 特殊コーパス＝非コーパスという考え方は、言語研究に貢献しない
- ３．サブコーパスをコーパスでないとするのは非合理的

- →コーパスという用語は便利ではあるが、多少曖昧で包括的な用語としてつねに考えられるべきである

A1.4 なぜパソコンを使うのか

- 機械可読式(machine-readability)は、近代コーパスの特性
- →紙のコーパスにはないメリットがある

- １．データ処理のスピードがはやい（検索、選択、グループ分け、配列）
- しかも、低コスト
- ２．正確に矛盾なく処理できる
- ３．人間の先入観を避けることができる
- →信頼性がある
- ４．自動化された処理を可能にする
- →コーパステキストがラベル分けされることにより、言語分析の質が向上

A1.5 Intuition-based vs. Corpus-based

- Intuition-based approach（直観準拠型）
- メリット
- ・直観はすぐに使用可能なため、すぐに理論的な例を創り出すことが可能
- ・自然言語の中に存在する言語外部影響（疲労、緊張などによる言いよどみ、繰り返し、言い間違いなど）から自由

- デメリット
- ・地域方言や社会方言（特定の社会階層で使用される言語）の影響を受ける可能性がある
- ・意識的に自分の言語使用をモニターすることになる
- →個人の判断の範疇を抜け出せない
- ・内観に基づいた結果のみでは、立証が困難

A1.5 Intuition-based vs. Corpus-based

- Corpus-based approach（コーパス準拠型）
- ・信頼できるテキスト、もしくは、本物のテキストに基づく
- （信頼性に関しては、議論の対象になりうる）
- ・直観だけでは気付けない違いを見つけることができる
- ・信頼性のある量的データを生み出す

- →昨今の言語研究は研究の幅が広いため、コーパスデータを使う際には、Intuition-basedとCorpus-basedのバランスをとることが大切

A1.6 方法論か理論か

- コーパス言語学は、理論ではなく方法論である
- （すべての言語学者が同意しているわけではないが、この考え方が主流）

- 音声学、統語論、意味論、語用論：言語使用をある特定の側面から研究
- コーパス言語学：側面が制限されていない
- むしろ、言語学研究のほぼすべての分野で用いられる

- →社会科学など他の学問も両方の側面を持つように、コーパス言語学は、方法論的側面と理論的側面を持っている
- →コーパス言語学は、言語学の多くの分野や理論にわたって、広範囲に適用できる方法論としてみなされるべきである

A1.7 corpus-based vs. corpus-driven

- Corpus-based approach（コーパス準拠型）
- コーパスが言語研究の前提となる前に形作られた理論や記述を、解釈、検証、例示するために主に用いられる
- 不利な証拠を捨てており、全体としてのコーパスデータにコミットしていないとして非難されている

- Corpus-driven approach（コーパス駆動型）
- 全体としてのデータの統合性に完全にコミットしている
- 理論上の主張が、コーパスに基づく根拠と完全に一致し、直接反映している

- Corpus-based vs. corpus-driven → 強調されすぎてる

A1.7 corpus-based vs. corpus-driven

- Corpus-driven approach に対する反論 ４つのポイント

- １．使用するコーパスの種類
- ２．既存の理論や知識に対する態度
- ３．研究の焦点
- ４．基本となる主張

- ［Corpus-driven approachの主張 → 反論］という流れでまとめました

A1.7 corpus-based vs. corpus-driven

- １．使用するコーパスの種類 (the type of corpus data used)

- ３つの論点
- （１）代表性　（２）コーパスの規模　（３）アノテーション

- （１）代表性
- コーパスは十分に規模が大きくなった時に自身で均衡をとるため、コーパスの均衡性と代表性を達成する必要はない
- ＝累積的代表性の獲得
- →この推定は正当性を欠く
- e.g.　the corpus of Zimbabwean English

A1.7 corpus-based vs. corpus-driven

- （２）コーパスの規模
- ・規模がとても大きい
- ・頻度をフィルターとして使用（頻度n以上の単語だけ分析するなど）
- →分析者が自身の分析からいくつかのデータを除外することを可能にする

- （３）アノテーション：テキストに情報を加えること
- １．属性情報付与　２．言語情報付与（e.g.タグ付け）
- 解釈の付与により、他の解釈で分析することが難しくなるため、アノテーションに強く異議を唱えている
- →implicit annotation（頭の中で行う無意識なアノテーション）は避けられない

A1.7 corpus-based vs. corpus-driven

- ２．既存の理論に対する態度
- コーパス以前の理論は軽視すべきではないとするが、既存の理論なしのコーパスを扱う
- （⇔Corpus-based：既存の理論に対して批判的ではない）
- →implicit annotationは、PC上のアノテーション（タグ付など）よりも信頼できない

A1.7 corpus-based vs. corpus-driven

- ３．研究の焦点 (research focuses)
- コーパス以前の概念である、語彙論、統語論、語用論、意味論、談話を区別しない
- →コロケーションは、KWICコンコーダンスのアノテーションされていないデータを容易に認識できるが、文法的に正しくタグ付けされないと、そのコリゲーションは不明瞭
- →文法的タグ付けは、既存の理論に基づくため、情報不足（ラベルがあることによる使いにくさ）

A1.7 corpus-based vs. corpus-driven

- コンコーダンス：コーパスの検索結果を検索語とその前後のコンテクストと共に表示する形式
- KWICコンコーダンス：特に、検索対象語(key word)を中央に配置し、その前後に一定の長さのコンテクストを表示する形式

- コロケーション：単語と単語の連結パターンの情報
- コリゲーション：単語と結び付きやすい文法構造のパターン情報

A1.7 corpus-based vs. corpus-driven

- ４．基本となる主張(paradigmatic claims)
- ある言語すべてを説明できる、新しい理論的枠組みであると主張
- →言い過ぎではないか

- まとめ
- ・現在では、corpus-drivenを主張する人は少ない
- ・corpus-basedを用いる研究者が多い
- ・この本では、corpus-based という用語を広義で用い、corpus-based と corpus-driven 両方を含む