Unit A2
Representativeness, balance and sampling

タイ語科3年 行木彩花

発表の流れ

1. 導入
2. 代表性
3. 外部基準/内部基準
4. サンプルコーパス/モニターコーパス
5. 汎用コーパスと特殊コーパスの代表性
6. 均衡
7. サンプリング
8. まとめ

導入

代表性

コーパスに不可欠な特徴
コーパス≠無作為に文章を収集したもの（archive)
全ての発話や文章を収集することは不可能
→抽出（サンプリング）が必要

どうしたら言語や言語の多様性を代表する抽出ができるのか？
→均衡を考慮し、代表性を保証する抽出を行う

代表性

Leechによるコーパスの定義
→代表性は不可欠

Biberによる代表性の定義
・代表性はサンプルが母集団の多様性の範囲に関係
・サンプリングは生きている言語集合の寄せ集め

▷代表性を決定づけるもの
コーパスが含む範囲（均衡）
テキスト集合がどう収集されたか（サンプリング）

外部基準/内部基準

外部基準(External Criteria)
言語特徴の分布に関わらない状況的な基準

内部基準
言語学的特徴を考慮した基準

▷Biber
テキストカテゴリー（言語学的にテキストタイプ）
→分野(genre)言語使用域(register)と定義

外部基準/内部基準

内部基準

言語学的特徴の分布を考慮に入れる

→コーパス作成前に言語特徴の分布を想定するのはおかしい

　自然発生的特徴を捉えられない

外部基準

言語的特徴は選択過程から独立、考慮しない

コーパス分析がコーパスの代表性を改善するものとして使用される

外部基準/内部基準

- Hunsonの主張

代表性は時間の経過によって変化（新たな代表性の側面）
コーパスが更新されない→代表性を失う

永続的な妥当性はコーパスをどう見るかによる

サンプルコーパス/モニターコーパス

サンプルコーパス(sample corpus)

静的な視点が適用される

長期的な変化:同じsampling frameを使用すれば可能

モニターコーパス(monitor corpus)

動的な視点が適用される

急速な言語変化、新表現に対応可能

長期的な変化:diachronic corpus通時的コーパスを使用すれば可能

汎用コーパスと特殊コーパスの代表性

汎用コーパス

言語、言語の多様性をすべて含む（例）BNC

特殊コーパス

領域や分野を特定

→どちらも様々な言語の多様性を代表させるため幅広いデータを収集

汎用コーパスと特殊コーパスの代表性

どちらも言語を代表するものであるべき
代表性の測定方法が異なる

汎用コーパス
幅広い分野から抽出されているか
特殊コーパス
語彙レベルにとどまる
→言語特徴の閉鎖度と飽和度合いで測定できる(closure/saturation)

均衡

言語の代表性

どれだけコーパスの均衡がとれているか
＝どれだけ幅広いカテゴリーを含んでいるか
　コーパスバランスの許容範囲は使用目的による

→よって汎用コーパスは話し言葉も書き言葉も含む

均衡

- 不可欠なもの
- 信頼できる科学的測定方法は存在しない

- 言語類型、分類、カテゴリーごとの分類

→コーパスの均衡を保つことに役立つ

均衡

British National Corpus(BNC)

- 均衡のとれたコーパス
- のちの多くのコーパスの構成に影響を与えた
- 100万語（90%書き言葉、10%話し言葉）

選択肢標（書き言葉）

‘domain’,’time’,’medium’の指標が用いられた

選択肢標（話し言葉）

demographic,context-governedという2つの指標が用いられた

→これらの設計指標はコーパスバランスの概念を表す

均衡

サンプルコーパス
- 均衡が特に重視される

モニターコーパス
- 常に更新される

→均衡の優先度は低い
- コーパスのサイズが大切
- 十分な規模に達したコーパスは

それ自体が均衡に影響を与えるという仮説

均衡

- 代表性と同じく作成者側にも使用者側にも重要

<u>代表性</u>

- 調査質問につながっている
- 流動的な概念

- 作成者はできる限り均衡に配慮しなければならない
- 読者はコーパスデータが研究に適切に使用されているか
  注意を払わなければならない

サンプリング

- 代表性と均衡のために適切なサンプリングが必要
- サンプリング方法が重要
- より大きな母集団から抽出を行う
- 統計上、抽出は母集団を縮小化したものである

サンプリング

- 母集団から代表性のあるサンプルを抽出するために
  Sampling Unitと母集団の境界を定義する

- Sumpling Unit（例）本や新聞

- Sumpling frame（Sumpling Unitの集合）
（例）Brown Corpusでは
Brown University LibraryとProvidence Athenaeumの定期刊行物

サンプリング

➢母集団の定義

- Language production
- language reception　　→人口動態（例）性別、階級
- language as a product

→話し言葉で母集団や構成を定義することは難しい

サンプリング

- Simple random sampling

  無作為にサンプリングする

  →比較的珍しいデータが集まらない

- Stratified random sampling

  母集団を比較的均質な層に分けて無作為に収集する

サンプリング

➢サンプルの規模

<u>完全なテキストから抽出</u>

著作権の問題

対象が完全な文章を含んでいること≠均衡

<u>部分から抽出</u>

むしろバランスが取れており適している

サンプリング

- 層ごとの割合

・使用頻度とターゲット層の大きさを考慮する
　→割合を決めるのは難しい

・直観的洞察に頼ることが多々ある

Summary

- 代表性、均衡、サンプリングが重要
- 言語や言語の多様性を表している
  →コーパスに代表性がある
- 代表制は均衡によってもたらされる
- 現在コーパスの代表性や均衡を測る方法はない
- コーパスを使用する際は調査質問に合っているか考える