Representativeness, balance and sampling

英語専攻4年　廣池桜子

本発表の流れ

- A2.1　導入
- A2.2　 Representativeness(代表性)とは
- A2.3　汎用コーパスと特殊目的コーパスの代表性
- A2.4　均衡
- A2.5　サンプリング
- まとめ

A2.1　導入

例）アメリカ英語とイギリス英語の書き言葉を比較

それぞれの英語で書かれた出版物すべてをコーパスに収めるのは不可能

サンプルを抽出する事が不可欠！！

↑

この場合、サンプルがアメリカ英語とイギリス英語の書き言葉を代表するようなもの（reprsentative）でなければ研究は成立しない。

A2.1　導入

つまり、

コーパスには、収集されたテキストが研究対象となる言語変種を代表しているという大前提が必要!!!

代表性を確保するには均衡とサンプリングについてよく考える必要がある。

A2.2　Representativeness(代表性)とは

- コーパスの代表性を決める要因

その①　コーパスに含まれるジャンルの範囲

その②　テキストの固まりの選定方法

A2.2　Representativeness(代表性)とは

- コーパスのためのテキストの選定に使われる基準

→External criteria (外部の基準)
(言語学的特徴には関わらない)

- Internal criteria（内部の基準）

(言語学的特徴に関わる)

→問題アリ
(コーパスはそもそも言語学的な語の配列を研究するものなので、分析する意味がなくなってしまう!)

- 従って、External criteria (外部の基準)を使うべき。

A2.2　Representativeness(代表性)とは

- 時間とともにコーパスの内容に変化が生じる場合:　変化→代表性に関わるファクター

内容の変わらないコーパス:　サンプルコーパス

内容が更新されるコーパス:　モニターコーパス

A2.3 汎用コーパスと特殊目的コーパスの代表性

- **汎用コーパス**:　言語全体を総合的に記述したコーパス
- 例)　BNC→現代イギリス英語全般の代表

- **特殊目的コーパス**:　ある分野や形式に特化したコーパス
- 例）分野:医療、法律　形式:新聞や学術的文章など

A2.3 汎用コーパスと特殊目的コーパスの代表性

- コーパスの代表性の測り方→コーパスの種類によって異なる。

- **汎用コーパス**→

そのコーパスが幅広い形式からサンプルを収集しているかどうか

- **特殊目的コーパス**→

そのコーパスの閉鎖性と充足性

A2.4 均衡

- コーパスの代表性→そのコーパスがいかに均衡のとれたものであるかに関わる

- 均衡のとれたコーパスは、その言語を代表するような幅広いテキストの種類を網羅している。

A2.4 均衡

- この「均衡性」という概念は、コーパスの必要条件と考えられているが、それを測る信頼のおける科学的な方法はない。

- 「均衡性」を保つために…

コーパスの作成者たちは、既存のコーパスモデルを採用している。

例）ANC,KNCはBNCをモデルに作られた!

A2.4 均衡

- BNCについて

9割が書き言葉、1割が話し言葉で構成

書き言葉→

‘domain’ ‘time’ ‘media’という三つの基準で選定

話し言葉→

‘demographic governed criteria’(個人の日々の会話から)

‘contextual governed criteria’(公共のイベントから)

＊このようにさまざまに基準を設けているのは、

均衡のとれたコーパスを作るため!!

A2.4 均衡

- Atkins

「科学的に均衡がとれたコーパスが登場するまでコーパスを使わないでいるのは、近視眼的であり、

均衡性が証明されていないからといって、そのコーパスの分析結果を退けてしまうのは短絡的である。」

A2.5　サンプリング

- 均衡のとれたコーパスを作るには、サンプリングの仕方が大事!
- サンプリングユニットと母集団の境界を決める必要がある
- 母集団も収集対象も定義が難しいものである

A2.5　サンプリング

サンプリングの方法

- simple random sampling
- stratified sampling

A2.5　サンプリング

- 集めるサンプルのサイズについて

書き言葉のデータが欲しい時、テキスト全体と一部、どちらを対象にすればよいのか？

→テキストの一部からサンプリングするべき

A2.5　サンプリング

Sampling Unit

→ひとつのサンプルのサイズ

例）ある作家の小説から5000語など

Stratified Sampling

→大きな枠の中のproportionを決めてからサンプルを採集する。

例）話し言葉50%，書き言葉50%

まとめ

- 代表性と均衡、サンプリングは深く関連

- コーパスから収集したデータが、その言語の特徴を表すようなものになっていれば、そのコーパスは代表性があるものだと言える。

参考文献

斉藤俊雄, 中村純作, 赤野一郎編(2000)
英語コーパス言語学: 基礎と実践. 第2版, 研究社.

赤野一郎, 堀正広, 投野由紀夫編(2014)
英語教師のためのコーパス活用ガイド. 大修館書店.