Chapter7 Using Available Corpora Latter half (7.7-7.9)

2013/6/13 (Thu) 3rd Period

Presented by Yoshimi Sugihara

A7.7
通時的コーパス Diachronic Corpora

- 通時的コーパスとは？
  - 1つの言語の異なる時期のデータ集めたもの
  - 言語の変化を辿る際に使用
  - Brown, Frown, LOB, FLOBやmonitor corpusより広範

- 既存の通時的コーパス
  - Helsinki Dialect Corpus
  - Corpus of English Dialogues
  - A Representative Corpus of Historical English Registers (the ARCHER)
  - Lampeter Corpus of Early Modern English Tracts

A7.7
通時的コーパス Diachronic Corpora

- Helsinki Dialect Corpus（最も有名）
  - 約150万語 / 400テキスト / 8-18世紀
  - 様々なジャンル/社会言語学的要素
  - 古、中、初期近代の3区分 / 11の下位区分
- ARCHER corpus (A Representative Corpus of Historical English Registers)
  - イギリス英語＆アメリカ英語
  - 1650-1990を50年ごとに区分
- Lampeter Corpus of Early Modern English Tracts
  - 110万語 / 広告印刷物のテキスト　/　1640-1740
  - 初期近代英語のテキスト構造の研究に有効

A7.8
学習者コーパス Learner Corpora

- 学習者コーパスとは？
  - 第二言語学習者の話し言葉、書き言葉の集積
  - 典型例の分析、個人の長期的分析　両方に使用
  - ⇔Developmental Corpus
    （母語として学ぶ子供の発話、作文データ）
    - Child Language Data Exchange System (CHILDES), Polytechnic of Wales Corpus (POW)

A7.8
学習者コーパス Learner Corpora

- International Corpus of Learner English(ICLE)
  - 300万語 / 14の異なる母語の上級学習者の作文
  - エラータグ、品詞タグ付与予定
  - 異なる母語を持つ学習者の比較、ネィティブスピーカーと学習者の比較（Louvain Corpus of Native English Essays使用）
  - 商用不可
- Longman Learners' Corpus
  - 1000万語 / 20の異なる母語の様々なレベルの学習者の様々なデータ)
  - L1 / レベルにより区分
  - 一部エラータグ付き
  - 辞書編纂者、教材開発者の材料
  - 商用可能

A7.8
学習者コーパス Learner Corpora

- The Cambridge Lerner Corpus (CLC)
  - 2000万語（拡大中） / 世界中（150ヵ国）の学習者のケンブリッジ英語検定（Cambridge ESOL）の5万のデータ
  - 母語、英語のレベル、年齢等で区分
  - 8万以上のエラータグ
  - Cambridge University Press, Cambridge ESOLの関係者のみ使用可

A7.8
学習者コーパス Learner Corpora

- 背景母語が一種のみのコーパス
- HKUST Corpus of Learner English
  - 1000万語 / 香港の中国人学習者の作文、試験
- The Chinese Learner English Corpus 100万語 / 中国人学習者（中等学校～高等学校）
- Standard Speaking Test (SST) Corpus 100万語 / 日本人学習者
- JEFLL(Japanese EFL Learner) corpus 100万語 / 日本人学習者
- JPU (Janus Pannonius University) learner corpus 40万語 / ハンガリー人学習者（ハンガリーの大学の上級学習者）
- USE (Uppsala Student English) corpus 100万語 / スウェーデン人学習者（スウェーデンの大学の上級学習者）
- The Polish Learner English Corpus 50万語 / ポーランド人学習者（レベル様々）

A7.9

モニターコーパス　Monitor Corpora

- モニターコーパスとは？
  - 新しいテキストを加えて拡大/テキストタイプは一定
- 例
  - Bank of English : 1980年代に創始 / 現時点で5億2400万語
  - The Grobal English Monitor Corpus 主要新聞のデータ / 数年で数十億語
  - AVIATOR (Analysis of Verbal Interaction and Automated Text Retrieval)

    バーミンガム大学で開発 / フィルターを用い自動で言語変化を観察

A7.9
モニターコーパス　Monitor Corpora

- モニターコーパスの目的：
  - 言語使用、意味論的変化の観察
  - 様々な地域の英語の同化あるいは分化の研究

- 否定的な意見
  - 'ongoing archive'(Leech)　サイズが非限定的＝完成品でない

A7.9
モニターコーパス　Monitor Corpora

- 1992　ALLCとACHの合同会議
  - Balanced sample corpus (Quirk, Leech) VS Monitor corpus (Sinclair Meijs)
- Monitor corpusの考え
  - Sinclairが発展
  - 20年前の人々のコーパスへの見解を反映
    - Sample corpusへの反論　コーパス全体のサイズVSサンプルのサイズ
  - コーパスの拡大=サンプルのサイズ及び数の拡大
  - あらゆる言語学的特徴が1つのサンプルにも含まれる様に

A7.9

モニターコーパス　Monitor Corpora

- モニターコーパスの難点
  1. 均衡をサンプルの大きさのみに頼っている
     →量的分析における信頼性低い
  2. テキスト全体を重要とするため、入手が困難
  3. 含有語彙数を示しても内容が分からない
  4. 拡大を続けるため比較が困難
     - ‘standard reference’としての価値無
     - サイズが一定のコーパスを内包するべき（Grönqvist）
  5. 長期的に基準が保たれる保証がなく、比較が困難

A7.9
モニターコーパス　Monitor Corpora

- **潜在的有用性**
  - **通時的コーパスよりも短い期間だが敏感**

    →急速な言語変化の研究　ex. Neologism

    →（長期的に存在すれば）速度の緩やかな言語変化

    （文法など）

    現在は存在しない

Chapter7のまとめ

- **既存の利用可能なコーパスの導入**
  - general VS specialized, spoken VS written, synchronic VS diachronic, learner corpus and monitor corpus
- **実際のコーパスは様々なタイプの混合**
  - ex. BNC: written＆spoken, general＆specialized, synchronic ＆diachronic
- **サンプルコーパスとモニターコーパスはそれぞれ異なる目的に有用**

- **既存のコーパスの有用性は、使用者の研究目的に拠る**
  - **目的にあったものが無い場合は自身で創る必要**