Unit A7
Using available corpora

英語専攻４年　岸本枝里香

A7.1 Introduction

この章では・・・

* 公に利用可能となっている主要なコーパス（英語）の紹介をします。
* 既成のコーパスを使うことに対する良い点と悪い点を挙げます。

* General vs. specialized corpus
* Written vs. spoken corpus
* Synchronic vs. diachronic corpus
* Learner corpus
* Monitor corpus

A7.1 Introduction

この章で紹介されているコーパスの分類

* 対象としている言語の多様なサンプル（ジャンル・領域）を含んでいるという点で、balanced である。
* 書き言葉データ、話し言葉データ、もしくは両方を含む。
* 有名なgeneral corpusとしては、the British National Corpus (BNC)が挙げられる。

* A7.2 General corpora とは・・・

A7.2 General corpora
BNCについて

- BNCは1億10万6千8語を含み、現代イギリス英語の4,124の書き言葉・話し言葉のテキストで構成されている。
- 90%が書き言葉・・・新聞、雑誌、学術書、フィクション、手紙、学校のエッセイなど。
- 10%が話し言葉・・・ビジネス、政府の会議、ラジオショーなど様々な文脈の中の話し言葉。

A7.2 General corpora
BNCの利点

* 品詞の情報に加え、豊富なメタデータが付与されている。
* 様々な研究目的にかなったコーパスである。
* 多くのアクセス方法がある。
* 同じsampling frameを用いた他の言語のコーパスがあるので、contrastive language studyの基盤となりうる。例：ANC, the Korean National Corpus, PNC

* ある特定の領域（医学、法律など）やジャンル（新聞記事、フィクションなど）を対象としたコーパス。
* 特定の領域やジャンルを研究する際に有用。
* General corporaの中から取り出すことも可能。

* A7.3 Specialized corpora とは・・・

* The Corpus of Professional Spoken American English (CPSA)
* The Michigan Corpus or Academic Spoken English (MICASE)
* Professional English Research Consortium (PERC)によるCorpus of Professional English (CPE)

A7.3 Specialized corpora
学術的・専門的なものの例

* The Brown University Standard Corpus of Present-Day American English (the Brown corpus)・・・初の現代英語コーパス（1961年）
* 約2000語、500チャンクの書き言葉テキストから作られた。

A7.4 Written corpora
the Brown corpusについて

A7.4 Written corpora
the Brown corpusとLOB

* The Lancaster-Oslo-Bergen Corpus of British English (LOB)・・・the Brown corpusのイギリス英語版で、そのモデルに沿って作られた。
* The Brown corpusとLOBを比較→1960年代初期のアメリカ・イギリス英語を比較できる。
* The Lancaster Parsed Corpus (LPC)・・・parsingされたLOBから取ってきた下位サンプル。
* The Susanne corpus・・・parsingされたthe Brown corpusから取ってきた下位サンプル。

* The Freiburg-LOB Corpus of British English (FLOB)・・・1991年のイギリス英語の書き言葉
* The Freiburg-Brown Corpus of American English (Frown)・・・1991年のアメリカ英語の書き言葉
* FLOBとFrownを比較→1990年代初期のイギリス・アメリカ英語を比較できる。
* LOBとFLOBを比較→イギリス英語の変遷が分かる
* BrownとFrownを比較→アメリカ英語の変遷が分かる

* A7.4 Written corpora
FLOBとFrown

A7.4 Written corpora
イギリス・アメリカ英語以外

* The Australian Corpus of English (ACE)・・・1986年以降の書き言葉のオーストラリア英語。
* The Wellington Corpus of Written NZ English (WWC)・・・1986～1990年の書き言葉のニュージーランド英語。
* The Kolhapur corpus・・・1978年以降のインド英語。
* The Lancaster Corpus of Mandarin Chinese (LCMC)・・・Brownのsampling frameを用いた中国語（マンダリン）のコーパス。

A7.5 Spoken corpora
the London-Lund Corpus (LLC)

* 1960年代～1970年代半ばまでも話し言葉のイギリス英語コーパス。
* 5000語から成る100のテキスト（計50万語）で構成されている。
* 対話と独白に分けられている。
* プロソディーに注釈がつけられている。

A7.5 Spoken corpora
the Lancaster/IBM Spoken English Corpus (SEC)

* 約53,000語の話し言葉のイギリス英語コーパス。
* 1980年代半ば以降。
* 実況放送、ニュース、講義や対話を含むラジオ放送から主に取ってこられた。
* 書き起こし版、品詞タグ付け・parsingされた版、プロソディーに注釈をつけた版も使用可能。

* The Cambridge International Corpus (CIC)の一部で、500万語から成る。
* 1995～2000年の間に集められた話し言葉のイギリス英語。
* 打ち解けた会話、買い物、話し合いなど様々な状況を含む。
* 話者間の関係性に関する情報が付与されている。
* 話者間の関係性が発話にどのような影響を与えるのかを見ることができる。

A7.5 Spoken corpora

the Cambridge and Nottingham Corpus of Discourse in English (CANCODE)

* The International Corpus of English (ICE)のアメリカ英語編の一部。
* アメリカ中から集めた数百の発話の録音から成る。
* 様々な故郷、年齢、職業、民族・社会的地位の話者。
* 会話、噂、議論、講義、スピーチなど色々な言語使用を反映している。
* 発話ファイルと書き起こしがつながっている。

A7.5 Spoken corpora

the Santa Barbara Corpus of Spoken American English (SBCSAE)

* 1988～1994年の話し言葉のニュージーランド英語コーパス。
* 100万語、551の抜粋から成る。
* インフォーマルな話し言葉の割合が高い(75%)。

* A7.5 Spoken corpora

the Wellington Corpus of Spoken New Zealand English (WSC)

A7.6 Synchronic corpora
National varieties

* 国ごとの英語の違いを見るには、共時コーパスであるthe International Corpus of English (ICE)が信頼できる。
* 100万語ずつから成る20のコーパスから成り、それぞれ1989年以降の話し言葉と書き言葉を含む。どれも同じコーパスデザインに沿っている。
* 英語が第一or主要言語である国や地域が対象。（アメリカ、イギリス、オーストラリア、カナダ、東アフリカ、香港など）

A7.6 Synchronic corpora
Regional dialects

* ある一国内の地域方言を比べるコーパスは、世界の英語を比べるコーパスよりもずっと少ない。
→あまり意味がないと思われているから。
* BNCの話し言葉の部分や、それに合うように作られたThe Longman Spoken American Corpusも使用可能。
* The Survey of English Dialects (SED)・・・1948～1973年のイギリスの田舎の方言が分かる。録音は高齢者の約60時間に渡る対話から成る。

A7.6 Synchronic corpora
英語以外の言語

* 英語以外で言語のバラエティーや方言の研究に適したコーパスは現在ほとんど存在しない。
* The Linguistic Variation in Chinese Speech Communities (LIVAC)はその一つ。
* 中国語を話す六地域（香港、台湾、北京、上海、マカオ、シンガポール）の新聞と電子メディアのテキストから成る。
* 共時コーパスなので全て比較可能。
* 現在作成途中だが、一部は既に利用可能。