UNIT A7

Using available corpora

3年　タイ語学科　江口 美優

A7.1 Introduction

〇このユニットで紹介されているもの

①一般公開されている利用可能な主なコーパス

②既成のコーパスを利用するうえでの良し悪し

・コーパス

→特定の目的のためにデザインされている

→既成のコーパスが役立つかどうかは使用者の意図する目的と合っているかどうかで判断

→よって、コーパスを研究する際に生じる疑問は今一度重要な役目を果たす

・世界には何千ものコーパスが存在している
→しかしながら、ほとんどのコーパスが特定の研究プロジェクトのために作られたもの
→したがって、我々が一般的に利用できるものではない。

・現在、英語以外の言語で作られたコーパスが数多く存在している
→このUnitでは、主に

general vs. specialized corpus
written vs. spoken corpus
synchronic vs. diachronic corpus
learner corpus vs. monitor corpus

に分類して英語コーパスについてみていく。

A7.2 General corpora(汎用コーパス)

〇様々な研究目的に広く対応できるように、多種多様なソースからバランスよく集められたコーパス
・書き言葉に関するデータを含む場合、話し言葉に関するデータを含む場合、または書き、話し言葉両方のデータを含む場合がある。

・よく知られている汎用コーパス = the British National Corus(BNC)

〇 BNC
・1億10万6008語
・4124の書かれた文書と現代のイギリス英語のトランスクリプトから構成
・できる限り広範囲で現代のイギリス英語を代表するように作成
・90% の書き言葉に関する部門(the spoken section)と10%の話し言葉に関する部門(the written section)から構成
・品詞タグの情報に加えて、豊富なデータでアノテーションされている
・辞書編集や人工知能、スピーチの評価、文学研究、そして言語学といった幅広い研究に役立つ
・現代のイギリス英語を研究するのに便利であるように、BNCは対照的な言語研究に信頼できる論拠を与えてくれる
・BNCに匹敵するコーパスもデザインされている
ex) the American National Corpus, the Korean National Corpus, the Polish National Corpus

A7.3 Specialized corpora（特殊目的コーパス）

〇特殊目的コーパス

→汎用コーパスと比較して、ある専門分野に特化したコーパスのこと

ex) the Guang zhou Petroleum English Corpus

・ 41万1612語

・ 全て石油化学の分野に関する英語の書き言葉の単語で構成されている

ex) the HKUST Computer Science Corpus

・ 100万語

・ 大学生のコンピュータに関する教科書からサンプリングした英語の書き言葉をもとに作成

・近年、特殊目的コーパスの作成や探究に関してかなりの興味、関心が抱かれている。

ex) the Corpus of Professonal Spoken American English
・専門的な場でのやり取りのトランスクリプトから選別して作成されたコーパス
・2つのサブコーパスから成り立っており、それぞれ100万語含んでいる
i)職員会議やテストに関する委員会のような教育に関する話し合いから成り立っているもの
ii)White House press の会議のトランスクリプトを含んだもの

他にも…
〇 the Michigan Corpus of Academic Spoken English
・ミシガン大学における現代の大学のスピーチに焦点をあてたコーパス
・170万語(200時間近くのレコーディング)

・the Professional English Research Consortium によるさらなる大掛かりな計画が進められている。
→100-million-word Corpus of Professional English の作成

・言語は、分野やジャンルをこえてかなり異なっているように、特殊目的コーパスは分野やジャンルに関連した価値のある研究材料を提供してくれる。
・また、特殊目的コーパスは、汎用コーパスから引用、抜粋されたコーパスにもなり得る。

A7.4 Written corpora

・最初の英語の現代コーパスは、アメリカ英語の書き言葉をもとに作られたコーパスであった。

① the Brown university Standard Corpus of Present-Day American English
・約2000語の500チャンクを使って編纂されている
・テキストは15の分野からサンプリングされている
・全て作られたのは1961年

・Brownモデルに倣ったコーパスがいくつかある

② the Lancaster-Oslo-Bergen Corpus of British English(LOB)
・英語版のBrown corpus
・Brown corpusと同じサンプリング技術を用いており、1961年に使用されたイギリス英語の書き言葉も加えられている。

・Brown corpus とLOB は1960年代初期に使用された英語の主な2つの多様性を比較する理想的な基本原理を生み出した
・2つとも品詞のタグ付けがされている

・2つのコーパスからのサブサンプルも構文分析されていた。

①' the Susanne Corpus
・Brown corpus から借用した構文分析化されたサブサンプルをもとに作成されたコーパス
・12万8000語

②' the Lancaster Parsed Corpus
・LOBから借用した構文分析家された約13万3000語のサブサンプルをもとに作成されたコーパス

・2つのFreiburg corpus は1990年代初期のBrown corpus とLOB の関係を反映している。

①'' the Freiburg-Brown Corpus of American English(Frown)
②'' the Freiburg-LOB Corpus of British English(FLOB)

→1991年に使われたアメリカとイギリスの書き英語を代表している
→さらに、Brown/LOBとFrown/FLOBの間に介在する30年間にわたるアメリカとイギリス英語の言語的変化をたどることが可能となる

・Brown のサンプリングモデルを利用して多様化した英語のコーパスも作られている。

ex) the Australian Corpus of English
・1986年～のオーストラリア英語の書き言葉をもとに作成されたコーパス

ex) the Wellington Corpus of Written NZ English
・1986~1990年のニュージーランド英語の書き言葉をもとに作成されたコーパス

ex) the Kolhapur Corpus
・1978年～のインド英語をもとに作成されたコーパス

・英語の多様性に焦点をあてていないコーパスもある。

ex) the Lancaster Corpus of Mandarin Chinese
・中国語と2つの主な英語の多様性を対比することが可能になった

A7.5 Spoken corpora

・汎用コーパスではあるが、一部が話し言葉のデータから作成されたコーパス。
・BNCや特定の分野のSpoken Corpus の他にもいくつかの英語のSpoken Corpus が利用可能。

① the LLC
・1960年代～1970年代中頃のイギリス英語をもとにしたSpoken Corpus
・100のテキストから構成、5000 words 、合計で50万 wordsが含まれている
・ダイアログ(どのような状況で人と対話するのか)とモノローグ(無意識に長々と話すのか準備したうえで話すのか)という点で区別している。
・韻律学的にアノテーションされている

② the SEC corpus
・1980年代半ば以降のイギリス英語をもとにしたSpoken Corpus
・主にラジオ放送を参考にしている
・5万3千 words が含まれている
・実況解説やニュース、講義や対談のような幅広い部門に対応している

③ CANCODE
・the Cambridge International Corpus の一部
・1995~2000年にイギリスで集められた自然と生じる会話を文字に起こして作成された(日常的な会話、一緒に働いている人たちの会話、買い物している人たちの会話、ディスカッションなど様々なシチュエーションを参考にしている)
・面白い特徴→話者同士の関係に関する情報も一緒にコード化されている
⇒話者間の親密度がどのように話し方に影響を与えているのかも知ることが可能

④ SBGSAE
・ the International Corpus of English のアメリカの部分の一部
・異なる出生地、年齢、職業、民族、社会的背景をもつアメリカ中の人々の自然に生じる何百もの会話をもとに作成
・人々が日常生活の中で使用する言語を反映
・それぞれのスピーチファイルはトランスクリプトがついているので、スピーチの評価を研究するのに特に便利

⑤ the WSC
・1988~1994年に集められたニュージーランド英語の話し言葉のデータをもとに作成(99%が1990年~1994年に集められた)
・データの 12%= formal speech/monologue
13%= semi-formal speech/elicited monologue
75%= informal speech/dialogue

私的なデータが多くを占めるため、くだけた話し言葉の使用域を調べるのに有効

A7.6 Synchronic corpora(共時コーパス)

・Unit7.4で便利だと紹介された英語のwritten corpusによって、英語の多様化を比較することができる。
→より信頼のおける世界の英語の多様化を比較するうえでの基礎となるものは、 “the International Corpus of English(ICE)”である。

○ the International Corpus of English(ICE)
・100万wordsがそれぞれ含まれた20個のコーパスを集めて編纂
・世界中で使用されている英語研究の比較を容易にすることが目的
・テキストのマークアップ、品詞のタグ付け、そして統語論の構文分析を含む様々なレベルでコード化

・地方の方言に関するコーパスで利用可能なものは少ない。
→方言の比較は、言語の多様性、または2つの異なった言語を言語を比較することよりも意味がないと想定
ex) the Longman Spoken American Corpus
→アメリカの地方の方言を比較する際に利用

・英語の方言を研究するために作られたコーパスは、the Survey of English dialects のspoken corpusである。

〇 the Survey of English dialects
・1948年にリード大学でハロルド・オートンによって始められる
・イギリス農村部の318か所でのインタビューをもとに方言の調査を行う
・調査の際、細かいインタビューとレコーディングが作成
・レコーディングにはトランスクリプトに関連した音声ファイルが付いていた

・現在、英語以外の方言や言語の多様化の研究に適した共時コーパスは少ない。

ex) the LIVAC (Linguisitic Variation in Chinese Speech Communication) Corpus
・存在する数少ないコーパスのひとつ
・香港や台湾、北京など中国語が話される6つの共同体における電子メディアや新聞を代表する文書が含まれている
・香港の大学で作成途中だが、いくつかのサンプルはすでに利用可能

A7.7 Diachronic corpora(通時コーパス)

〇同じ言語で様々な時代区分から集めたテキストを含むコーパス
・サンプルを集める時代は、Brown や Frown , LOBや FLOB, またはモニターコーパスのものよりも広範囲
・言語の進化の過程で生じた変化を知るために利用される
・書き言葉しか含まれていないが、コーパス作成者は、昔の時代から話されていた話し言葉のコーパスも作成しようとしている

① the Helsinki Diachronin Corpus of English(= the Helsinki Corpus)
・最もよく知られている英語の通時コーパス
・約150万語
・8世紀~18世紀が対象
・広いジャンルに対応しており、3つの区分に分けることができる

i) Old English (750~1150)
ii) Middle English (1150~1500) + さらに11の時代に細かく分類可能
iii) Early Modern English (1500~1710)

② the ARCHER Corpus
・1650年~1990年のイギリス、アメリカ英語をもとに作成
・話し言葉のデータも含まれている

③ the Lampeter Corpus of Early Modern English Tracts
・the ARCHER Corpus よりも利用しやすいコーパス
・1640年,1740年~のパンフレットをもとに作成
・110万語
・小さなサンプルからというよりはむしろテキスト全体が含まれており、初期の現代英語の文書がそのように構成されていたのかを研究するうえでも役だつ

A7.8 Learner corpora(学習者コーパス)

〇言語を学ぶ教室に直接関係のあるコーパスのタイプ
・第二言語(L2)を習得しようとしている学習者が書いたものやスピーチを集めて作成

① the International Corpus of Learner English(ICLE)
・14の異なる母語の背景をもつ上級の英語学習者によって書かれた小論文をもとに作成
・300万語
・エラータグ、品詞タグ付けされたものも近い将来利用可能に
・異なる母語背景をもつ学習者の書いたものを比較できる + the Louvain Corpus of Native English Essaysと組み合わせる = 英語の母語話者と学習者のものを比較できる
・言語研究のためには有効、営利的な目的で利用することはできない

② the Longman Learner's Corpus
・20もの異なるL1に背景を持ち、英語の熟練度の異なる学生が書いたテキストをもとに作成
・1000万語
・生徒がどのような知識を持っているか、どんな間違いをするのかが分かる
・辞書編集者や教科書の題材の作成者にとっては便利な資料
・営利的な目的で一般に公開

③ the Cambridge Learner Corpus
・世界中の英語学習者が書いた例を集めて作成
・2000万語以上、拡大し続ける
・世界中でCambridge ESOL English試験を受けた学生が書いた匿名の答案を参考
・それぞれの答案は、生徒のL1、国籍、英語のレベル、年齢などの情報と一緒にコード化
・今のところ、ケンブリッジ大学で働いている作成者、ケンブリッジESOLのスタッフのみ利用可能

・さらに、L1の背景のみをカバーした英語学習者コーパスも利用可能

ex) the HKUST Corpus of Learner English
ex) the Chinese Learner English Corpus
ex) the SST(Standard Speaking Test) Corpus
ex) the JEFLL(Japanese EFL Learner) Corpus
ex) the JPU(Janus Pannonius University) Learner Corpus
ex) the USE(Uppsala Student English) Corpus
ex) the Polish Learner English Corpus

A7.9 Monitor corpora

・これまで紹介してきたサンプルコーパスのサイズは一定
・モニターコーパスは、絶えず新しい情報が加えられ、サイズは大きくなり続ける

① the Bank of English
・1980年代に作られ始めてから次第にサイズが大きくなり続け、現在5億語を超えている

② the Global English Monitor Corpus
・2001年から作られ始めたが、数年以内に10億語を超えると予測されている

③ AVIATOR
・バーミンガム大学で発展したコーパス
・新しい語の形やペア、または語句や意味上の変化を一連のフィルターを用いて特定、分類することで自動的に言語の変化を観察している

・しかしながら、モニターコーパスのサイズは無限であるため、コーパス言語学者の中には「モニターコーパスは本物のコーパスというよりはむしろ発達中のアーカイブである」と主張する人もいる。

・1992年にクリストチャーチ大学において開かれたthe Association for Literary&Linguistic Computingとthe Association for Computing in the Humanitiesの協働会議で重要な議論が行われた。

Quirk&Leech→バランス型サンプルコーパスに賛成
Sinclair&Meijs→モニターコーパスモデルに味方

→1992年の論争では、モニターコーパスが勝利
→サンプルコーパスは現代のコーパス作成において主要なもの+今日まで作成されたコーパスのほとんどがバランス型サンプルコーパスである
→さらなる論争でサンプルコーパスモデルが勝利することは明らか

・モニターコーパスの取り組みはSinclairによって行われる
・SinclairはBrownのような静的なサンプルコーパスに反論、動的なモニターコーパスには賛同するような主張
・Sinclairの主張は、未だに普及している
・Sinclairがサンプルコーパスに反対して示したコーパスサイズとサンプルサイズについての主張は、再検討するに値するとされている。

・コーパスサイズとサンプルサイズについての懸念→コンピュータの製版によって大量の電子テキストが入手可能になったので中和されつつある
・Aston やBurnard→「絶え間ないコーパスサイズの成長は、サンプルの数と同様に、サンプリングのサイズも増えることをほのめかしている」

・飽和状態に達したモニターコーパス→古い情報は捨てて、新しい情報は加え、最新の状態を維持
⇒しかしながら、困った点が存在する
　ⅰ)コーパスのバランスをとるための理念に基づいた試みを拒み、その代わりに問題を取り扱う上でごく薄いサイズに頼っている点。量的にサンプルコーパスより信頼できない。

　ⅱ)全文採用のデータを経時的に次々と収集→著作権の問題で文書の使用が困難になる点。

　ⅲ)特定のコーパスのバージョンを指し示すのがとても困惑な点。使用者ではなく、作成者のみが特定のバージョンに何が含まれているか把握可能。

　ⅳ)サイズが大きくなり続けているため、容易に検索結果を比較できない点。

・モニターコーパスは潜在的にはとても便利
・異なる時代区分のにおける言語的変化を追うようにデザイン→通時的コーパスと似ている面も
・新しい語や表現の登場から消失までのように比較的速い言語変化を研究するのに利用可能
・長い間存在するのであれば、文法の変化などよりゆっくりとした速さで生じる変化を研究することも可能になるかもしれない

Summary

・このUnitでは、一般公開されていて利用可能である主な英語コーパスを紹介

・よく知られている、大きな影響を及ぼし得る英語または他の言語のコーパスのさらなる幅広い調査が知りたい→コーパス作成者のウエブサイトへ

・ここで紹介されたコーパスごとの違いは便利であるものの、これから巡り合うコーパス全てに適応できるかどうかは分からない

・繰り返されるテーマ→与えられたコーパスの利便さは使用者が意図する研究目的に依拠している
・既成のコーパスは多数存在するものの、自分の研究に役立たないコーパスも存在する
→自分自身のコーパスを作らなければならない