English-Corpora.org を用いた言語データの採取

WORK IN PROGRESS (Ver 2020-05-29)

長谷部 陽一郎

同志社大学

yhasebe@mail.doshisha.ac.jp

1　はじめに

1.1　ユーザー・カテゴリーについて

English-Corpora.org（https://english-corpora.org）を本格的に利用するには，ユーザー登録を行う必要がある．登録ページ（各コーパス画面の右上にリンクあり）に必要事項を入力すれば，無料でユーザー・アカウントを取得できる．複数のユーザー・タイプがあり，それぞれ 1 日に可能な検索数が決められている．非登録ユーザーも基本的な検索は可能だが，回数が 1 日 20 件に制限されているほか，10〜15 回の検索ごとに登録を促すメッセージが表示される．[*1]

表 1　English-Corpora.org のユーザー・タイプ

レベル	カテゴリー	1 日の可能検索数	1 日の可能 KWIC 表示件数
4	言語学の研究者・院生（有料アカウント）	400	20,000
3	言語学の研究者・院生（無料アカウント）	200	15,000
2	言語学以外の研究者，教員など（無料アカウント）	100	10,000
1	言語学以外の院生，学部生，その他	50	5,000
0	非登録ユーザー	20	2,000

[*1] User Categories: https://english-corpora.org/userCategories_e.asp

1.2 ユーザー登録

English-Corpora.org を利用するためのユーザー登録を行うには，画面上部の my account から Register のページにアクセスし，情報を入力する（下図を参照）.

Please fill out the brief form below. Within one or two minutes, you will receive an email. Simply click on the link in that email, and you will be able to continue using the corpora. (Note: If you have already registered, please log in to see your profile)

Name	Hanako Kyotanabe (e.g. Mary | Smith)	
Email address	12345@mailx.doshisha.ac.jp	
Password	•••••	
Country	JAPAN	
Category	○ University professor: languages / linguistics ○ Graduate student: languages or linguistics	3 RESEARCHER
	○ University professor: not languages / linguistics ○ Teacher: not university; not graduate student	2 SEMI-RESEARCHER
	○ Graduate student: not languages or linguistics ● Student (undergraduate) ○ Other	1 NOT RESEARCHER
	☑ I agree to the Terms and Conditions for this website	
	SUBMIT RESET PROBLEMS ??	

2 English-Corpora.org の概要

2.1 機能と特徴

English-Corpora.org のコーパス群を使って，次のようなことができる．多くの機能と操作体系の基本的な部分は共通している．

- 語句の正確一致検索，ワイルドカード検索，レンマ検索，品詞検索を行う．これらを組み合わせることもできる．
- 最大 10 語の幅で近接語（コロケーション）の検索を行う．（例：faint に近接する名詞，woman に近接するすべての形容詞，feelings に近接するすべての動詞，など）

- 語，句，構文の検索結果に対して，頻度によるフィルターをかけたり，ジャンルごと，あるいは時代ごとの頻度比較を行う．
- 2 つの関連した語句のコロケーションを比較する．（例：little/small，democrats/republicans，men/women）
- 検索の結果として得られたワード・リストや自分で用意したワード・リストを使って，さらに別の検索を行う．
- 基本語彙については，簡便なインターフェイスを用いて，使用域，共起語，関連語などの情報を一覧表示できる．

なお，English-Corpora.org のコーパスではすべての語に品詞情報が付与されている．ただし，ICE-GB コーパス（British component of International Corpus of English）に見られるような統語解析は施されていない．[*2]

2.2 コーパスのリスト

2020 年 5 月の時点で 19 の英語オンライン・コーパスが利用可能である．[*3]ウェブ上で利用できるインターフェイス以外に，COCA などのコーパスから抽出した n-gram データをダウンロードできるサービスも提供されている．[*4] 下記ではそれらのうち，代表的な 10 のコーパスを紹介する．

2.2.1 Corpus of Contemporary American English [COCA]

- URL: http://english-corpora.org/coca/
- 収録語数：10 億語
- 言語：アメリカ英語
- 期間：1990 年〜2019 年
- ジャンル：均衡

English-Corpora.org のコーパス群の中で最もよく利用されているコーパスの 1 つ．世界中の研究者によって実際の研究に利用されている．10 億の収録語は話し言葉，フィクション，一般雑誌，新聞，学術テキストをバランスよく含んでいる．1990 年から 2019 年の各年につき 2 千万語が収録されるように調整されており，現在の英語，そして現在英語に起こっている変化について調べるのに役立つ．

2.2.2 Corpus of Historical American English [COHA]

- URL: http://english-corpora.org/coha/
- 収録語数：4 億語
- 言語：アメリカ英語

[*2] ICE-GB: https://www.ucl.ac.uk/english-usage/projects/ice-gb/
[*3] English-Corpora.org: https://english-corpora.org/
[*4] N-grams: http://www.ngrams.info/

- 期間：1810 年〜2009 年
- ジャンル：均衡

4 億語から成る 1810 年から 2009 年にかけてのアメリカ英語テキストが検索可能．語，句，構文の出現頻度はもちろん，時系列上の意味変化や文体の変化を調べることができる．

2.2.3 TIME Magazine Corpus [TIME]

- URL: http://english-corpora.org/time/
- 収録語数：1 億語
- 言語：アメリカ英語
- 期間：1923 年〜2006 年
- ジャンル：雑誌記事

1923 年から 2006 年までの TIME 誌に掲載されたアメリカ英語 1 億語を検索可能である．語，句，構文の出現頻度や意味の変化を追うことができる．

2.2.4 Corpus of American Soap Operas [SOAP]

- URL: http://english-corpora.org/soap/
- 収録語数：1 億語
- 言語：アメリカ英語
- 期間：2001 年〜2012 年
- ジャンル：テレビ番組（ドラマ）

2001 年から 2012 にかけての 22,000 本以上のアメリカのソープ・オペラの脚本から抽出した 1 億語規模のコーパスである．通常の「話し言葉」コーパスより，さらにインフォーマルで，日常言語の姿をよく表したコーパスである．また，大多数の話し言葉コーパスより多くの収録語数を誇る．

2.2.5 British National Corpus [BNC]

- URL: http://english-corpora.org/bnc/
- 収録語数：1 億語
- 言語：イギリス英語
- 期間：1980 年代〜1993 年
- ジャンル：均衡

British National Corpus (1970 年代〜1993 年) の 1 億語からなるテキストを検索できる．BNC は 1980 年代から 1990 年代初頭にかけて Oxford University Press で開発されたコーパスで，現在ウェブ上でいくつかのバージョンを利用可能である．English-Corpora.org の BNC は最新のタグセットである CLAWS7（後述）を用いているため，他の多くのコーパスとデータ形式の互換性がある．

BNC では使用域を指定した語句検索が可能である．例えば「話し言葉」「学術」「韻文」「医療」などである．また使用域間での比較もできる．例えば，法律と医療のそれぞれの領域でどのような動詞が使われやすいか，break と共起しやすい名詞はフィクションと学術テキストとでどのように違うか，などを調べることができる．

2.2.6 Strathy Corpus [STRATHY]

- URL: http://english-corpora.org/can/
- 収録語数：5 千万語
- 言語：カナダ英語
- 期間：1970 年代〜2000 年代
- ジャンル：均衡

Queen's University の Strathy Language Unit が開発した Strathy Corpus of Canadian English を検索できる．Strathy コーパスは，1100 以上の話し言葉，フィクション，雑誌，新聞，学術テキストから得られた 5 千万語からなる．BNC と同様，English-Corpora.org の他のコーパスと共通したデータ・フォーマットを採用している．

2.2.7 Early English Books Online [EEBO]

- URL: http://english-corpora.org/eebo/
- 収録語数：7 億 5 千万語
- 言語：イギリス英語（近代英語）
- 期間：1470 年代〜1690 年代
- ジャンル：各種書籍

オープンソースとして公開されている古い時代の約 2 万 5 千のテキストのデータから構成される．コーパスは英国芸術・人文リサーチカウンシルによってまとめらた．品詞や意味のタグ付けはランカスター大学のチームにより行われた．

2.2.8 Global Web-Based English [GloWbE]

- URL: http://english-corpora.org/glowbe/
- 収録語数：19 億語
- 言語：20 カ国の英語
- 期間：2012 年〜2013 年
- ジャンル：ウェブ

英語使用国 20 カ国の 18 億のウェブページから採取した 19 億語からなるコーパスで，2013 年 4 月にリリースされた．地域，ジャンル，時代によって異なる様々な英語についての調査が可能になる．

GloWbE ではあらゆる語，句，構文について，20 の異なる国々のデータを得ることができる．イギリス英

語とアメリカ英語（この 2 カ国で 7 億 7500 万語を占める）を比べたり，オーストラリア（1 億 4800 万語），南アフリカ（4500 万語），シンガポール（4300 万語）といった国々の英語に関するデータを得ることができる．

2.2.9 Wikipedia Corpus [Wiki]

- URL: http://english-corpora.org/wiki/
- 収録語数：19 億語
- 言語：英語
- 期間：2014 年
- ジャンル：Wikipedia 記事

Wikipedia 英語版の 4400 万項目のテキストを採録したコーパス．2015 年に発表された．Virtual Corpus という機能が実装されており，特定のトピック（biology, video game, investment, social network, etc）に関するバーチャル・コーパスを簡単に作成できる．また，バーチャル・コーパスの中で特定語句がどれくらい出現するか，Wikipedia Corpus 全体における出現頻度との相対的な関係のもとで見ることができる．また，バーチャル・コーパスの中で特徴的な語句を見つけ出すための機能が備わっている．（**2020 年 5 月 29 日時点でシステム自体の動作が不安定？**）

2.2.10 The TV Corpus [TV]

- URL: http://english-corpora.org/tv/
- 収録語数：3 億 2500 万語
- 言語：6 か国の英語
- 期間：1950 年〜2018 年
- ジャンル：TV 番組

1950 年代から現代までの約 7 万 5 千の TV 番組からのデータで構成される．すべての番組は Internet Movie Database (IMDB) のエントリーに紐づけられており，詳細な背景的情報を得ることができる．[*5]口語的な英語表現の資料として用いるのに適している．また，時代ごとの，異なる地域（例えばアメリカとイギリス）の英語を観察することが可能である．

2.2.11 The Movie Corpus [Movie]

- URL: http://english-corpora.org/movies/
- 収録語数：2 億語
- 言語：6 か国の英語
- 期間：1930 年〜2018 年
- ジャンル：映画

[*5] Internet Movie Database (IMDB): https://www.imdb.com

1930 年代から現代までの約 2 万 5 千の映画からのデータで構成される．すべての映画は Internet Movie Database (IMDB) のエントリーに紐づけられており，詳細な背景的情報を得ることができる．口語的な英語表現の資料の資料として用いるのに適している．また，時代ごとの，異なる地域（例えばアメリカとイギリス）の英語を観察することが可能である．

2.2.12 The Intelligent Web-based Corpus [iWeb]

- URL: http://english-corpora.org/iweb/
- 収録語数：14 億語
- 言語：6 か国の英語
- 期間：2017 年
- ジャンル：ウェブ

English-Corpora.org のコーパス群の中で最大．体系的に選別された約 9 万 5 千のウェブサイトから採取された言語データで構成される．特定のトピックに限定したバーチャル・コーパスを作成することができる (e.g. chocolate, basketball, solar power, Harry Potter, etc.)．

練習問題 1

上で示した中で，COCA，COHA，GloWbE など，いくつかのコーパスにアクセスし，適当な検索文字列を入れて試してみよう．

Wiki や iWeb などで使えるバーチャル・コーパスについて

作り方

SEARCH タブで List を選択し，Options から Texts/Virtual をクリック．さらに FIND WEBSITES を選んでからキーワードを入力して Find matching strings ボタンをクリック．候補となるウェブサイトがヒット数に基づいてリストアップされるので，適宜チェックマークを付けて SAVE List をクリックした後，SAVE AS にバーチャル・コーパスの名前を入力して SUBMIT する．

使い方

SEARCH タブの List に戻り, Text/Virtual で Refresh list をクリックする．その上で再び Text/Virtual をクリックすると，先ほど作成したバーチャル・コーパスが検索対象として表示されるので，これを選択して，実際の事例を調べたい語句を入力して検索を実行する．List 検索の他，Collocates や KWIC での検索も可能.

2.3 画面上の各種アイコンの意味

どのコーパスを開いても画面の上部に以下のアイコンが表示される．それぞれの意味は以下の通り．


コーパスの特徴や使い方を説明した PDF を表示


コーパスの概要を見る（目的，用法，テキスト，検索タイプなど）


コーパスの統計情報を見る（セクションごとの語数など）


他のコーパスに切り替える（同じ検索クエリを別のコーパスで実行する）


ログイン情報や以前に作成したワードリストなどを確認


検索履歴を見る

3 English-Corpora.org コーパスの機能と使い方

ここでは実際に English-Corpora.org のコーパスを使用する際に役立つ，様々な検索方法や手順をみていく．特に指定のない限り，COCA で操作を行うことを前提とするが，基本的な部分は他のコーパスでも同様である．

3.1 検索のタイプ

English-Corpora.org のコーパスでは次のような様々なタイプの検索が可能になっている．

List
: 指定されたパターンに一致する表現をリストアップする．示された表現をクリックすることで，実際の事例を確認することができる．

Chart
: 年代やジャンルにセクションが分かれているコーパスであれば，検索クエリの結果件数をセクションごとにチャート形式で示すことができる．

Word
: 指定された単語の出現頻度が全体の中で 6 万語以内である場合，その語についての様々な情報を一覧できる．関連するトピック，共起語，KWIC 表示，関連後，高頻度で出現するウェブサイトなど．

Browse
: 通常の検索よりも簡便なインターフェイスで気軽に語を検索して Word 表示ができる．

Collocates
: 中心となる語（word/phrase）と共起語（collocates）のパターンを入力し，両者の距離（中心語から 4 語以内）を指定して事例を検索することができる．

Compare

2 つの語（Word1 と Word2）と共起語（collocates）のパターンと距離を入力すると，Word1 と Word2 のそれぞれが共起語と一緒に出現する頻度を比べることができる．

KWIC

指定したパターンに一致する事例を前後の文字列（文脈）と共に表示する（KWIC = Keyword in Context）．

このうち，基本となる List モードでの検索については本節の 3.2 以降で，Collocates については 4 節で，List，Chart，Compare, KWIC については，5.1 節でより詳しく見ていく．

3.2 基本検索シンタックス

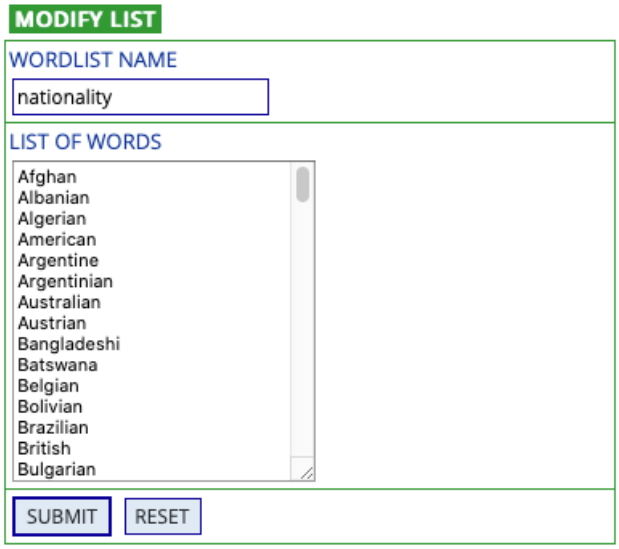
English-Corpora.org のコーパスの検索シンタックスでは，スペースで区切られた 1 つ 1 つのまとまりを「スロット（slot）」と呼ぶ．各スロットは「語」に対応しており，スロットの中にスペースを含めることはできない．

表 2　基本的な検索

フォーマット	検索種別	実際の例	結果の例
[pos]	品詞検索	[vvg]	going, using
[lemma]	レンマ検索	[sing]	sing, singing, sang
		[tall]	tall, taller, tallest
[=word]	同義語検索	[=strong]	formidable, muscular, fervent
@listname	ワード・リスト	@clothes	tie, shirt, blouse, coat, jacket, etc.

ワード・リストについて

ワード・リストを作成するにはユーザー登録が必要．登録後，画面上部のユーザーアイコンをクリックして ACCOUNT 画面を表示し，Saved lists という項目下にある Customized word lists をクリックすると，下図のようにリスト名と単語リストを記入することができる．なお，リストを構成するのはスペースを含まない「単語」でなければならないという制約がある．ワードリストを使用する際には，リスト名の前に@を付けたものを検索クエリの中に含める（例：`@nationality`）．



練習問題 2

基本問題

次のような語句・構文を検索してみよう．

(a) 形容詞 + record
(b) foreseeable + 名詞
(c) sing（レンマ） + a + 形容詞 + song
(d) hold の同義語 + a party
(e) surprising の同義語 + news の同義語

ヒント 形容詞は `[j*]`，名詞は `[n*]`

発展問題

(a) 検索モードを Chart や KWIC に変更して上記の検索を試してみよう．
(b) 検索モードを Compare に変更し，Word(S) に `[idea]` と `[concept]` を入力してみよう．

3.3 ワイルドカード検索

ワイルドカードを用いた検索は，異なる語尾形式の語をまとめて検索したり，品詞検索の粒度を調整するのに役立つ.

表 3　ワイルドカード検索

フォーマット	検索種別	実際の例	結果の例
`*xx`	*は 0 以上の数の文字	`un*ly`	unlikely, unusually
`x?xx`	?は 1 文字	`s?ng`	sing, sang, song
`x?xx*`	上記の組み合わせ	`s?ng*`	song, singer, songbirds
`[pos*]`	品詞検索で	`[v*]`	find, does, keeping, started

練習問題 3

次のような語句を検索してみよう.（検索モードは List に戻しておくこと）

(a) holic で終わる語
(b) 接頭辞 un と接尾辞 able を共に含む語
(c) 接頭辞 under と接尾辞 ed を共に含む語
(d) it で終わる 4 文字の単語

3.4 OR/NOT 検索

OR と NOT といった意味を表す論理演算子を利用した検索も可能である.

表 4　論理演算子を用いた検索

フォーマット	検索種別	実際の例	結果の例
`word|word`	OR 検索	`stunning|gorgeous|charming`	stunning, charming, gorgeous
`-word`	NOT 検索	`-[nn*]`	the, in, is

練習問題 4

次のような語句を検索してみよう．

(a) e-mail，email もしくは electronic-mail
(b) thank you so much もしくは thank you very much
(c) look（レンマ） + forward 以外の語 + to

3.5 複合検索

ピリオドを使って，1 つのスロットの中で要素を組み合わせることができる．この機能は，語の特定の品詞としての用例を抽出するような場合に役立つ．例えば表 5 の最後の例であれば，動詞を指定しているので，rhythm や drumming のような名詞は結果から除外される．

表 5 要素の組み合わせ

フォーマット	検索種別	実際の例	結果の例
`word.[pos]`	語＋品詞	`strike.[v*]`	strike
`word*.[pos]`	語＋品詞	`dis*.[vvd]`	discovered, disappeared, discussed
`[lemma].[pos]`	レンマ＋品詞	`[strike].[v*]`	strike, struck, striking
`[=word].[pos]`	同義語＋品詞	`[=beat].[v*]`	hit, strike, defeat

練習問題 5

次のような語句を検索してみよう．

(a) book（レンマ・動詞） + a + 名詞
(b) you + 動詞 + beautiful の同義語（形容詞に限る）

角型括弧（`[ ]`）を余分に加えることで，「同義語のレンマ検索」を実現できる．もちろん，これに品詞指定を加えることも可能である．

表 6　同義語のレンマ検索

フォーマット	検索種別	実際の例	結果の例
`[[=word]]`	同義語＋レンマ	`[[=publish]]`	announced, circulating publishes, issue socked, shirt
`[[=word]].[pos]`	同義語＋レンマ＋品詞	`[[=clean]].[v*]`	mop, scrubs, polishing

練習問題 6

次のような語句を検索してみよう．

(a) advice の同義語
(b) advice の同義語（レンマ）
(c) help の同義語（動詞）
(d) help の同義語（動詞・レンマ）

TIPS

検索結果として示された語の後の `[s]` をクリックすると，さらにその語の同義語をみることができる．

3.6　句や構文の検索

すでに述べた通り，要素をスペースで区切ることで複数の語（＝複数のスロット）から成る句を検索できる．下にいくつかの例を示す．

表 7　句の検索

実際の例	結果の例
`nooks and crannies`	nooks and crannies
`fast|quick|rapid [nn*]`	fast food, rapid transit
`pretty -[nn*]`	pretty smart, pretty as
`[get] her to [v*]`	get her to stay, got her to sleep
`.|,|; nevertheless [p*] [v*]`	. Nevertheless it is
	, nevertheless he said
`[break] the [nn*]`	break the law, broke the story
`[beat].[v*] * [nn*]`	beat the Yankees, beaten to death
`[=gorgeous] [nn*]`	beautiful woman, attractive wife
`[put] on [ap*] @clothes.[n*]`	put on her hat, putting on my pants

練習問題 7

English-Corpora.org の COCA 以外のコーパスでの検索を試してみよう．

(a) TIME コーパスで検索モードを Chart に設定し，greenhouse effect と global warming をそれぞれ検索してみよう．

(b) COHA コーパスで検索モードを Chart に設定し，「help（動詞・レンマ）＋代名詞＋to＋動詞」の構文パターンと「help（動詞・レンマ）＋代名詞＋動詞」の構文パターンをそれぞれ検索してみよう．

(c) GloWbE コーパスで検索モードを Chart に設定し，`[wait] in a queue` と `[wait] in a line` をそれぞれ検索してみよう．

3.7　CLAWS7 タグセット

ここでは，English-Corpora.org のコーパス検索で利用できる品詞タグ（CLAWS7 タグ）のうち主なものを示す．English-Corpora.org のコーパスで CLAWS7 を使う際には次の 2 点に注意する必要がある．

- 名詞句に前置される所有格代名詞 (例：my, your, our) のタグは本来 `[APPGE]` であり，代名詞を意味する `[p*]` ではなく限定詞を意味する `[a*]` にマッチする．
- システム上では `noun.ALL` すなわち名詞すべてにマッチするタグとして `[nn*]` が示されているが，これは固有名詞（曜日名や月名を含む）にマッチしない．

なお，CLAWS7 タグの詳細については http://ucrel.lancs.ac.uk/claws7tags.html を参照のこと．

表 8　基本品詞タグ

タグ	意味	実際の例
[n*]	名詞	sheep, book, books, inch, IBM
[v*]	動詞	be, was, can, do, have, give
[j*]	形容詞	old, better, strongest, able
[r*]	副詞	kindly, else, namely, very
[xx*]	否定辞	not, n't
[d*]	限定詞	such, little, this, which
[p*]	代名詞	none, who, it, anyone, he, them
[app*]	所有格代名詞	my, your, our
[i*]	前置詞	for, of, in, with
[c*]	接続詞	and, or, but, if, as, than

表 9　名詞類のタグ

タグ	意味	実際の例
[nn1*]	普通名詞単数形	book, girl
[nn2*]	普通名詞複数形	books, girls
[nn0*]	不可算名詞	. aircraft, data, committee
[np*]	固有名詞	IBM, Andes, Smith, Sunday, October
[nn*]	普通名詞	sheep, cod, headquarters, book, girls

表 10　動詞類のタグ

タグ	意味	実際の例
[VV0*]	語彙動詞・原形	give, work
[v?i*]	動詞・不定詞	be, do, have, give, work
[vvi*]	語彙動詞・不定詞	give, work
[vm*]	動詞・モーダル	can, will, would, ought, used
[v?z*]	動詞・3 人称単数	is, does, has, gives, works
[v?d*]	動詞・過去	was, did, had, gave, worked
[v?n*]	動詞・過去分詞	been, done, had, given, worked
[v?g*]	動詞・ING	being, doing, having, giving, working
[vv*]	語彙動詞	give, work, gives, giving, worked
[vb*]	BE 動詞	be, is, was, were, been, being
[vd*]	DO 動詞	do, does, did, done, doing
[vh*]	HAVE 動詞	have, has, had, having

表 11　形容詞・副詞類のタグ

タグ	意味	実際の例
[jjr*]	形容詞・比較級	older, better, stronger
[jjt*]	形容詞・最上級	oldest, best, strongest
[rp*]	不変化詞	about, in
[rrq*]	WH 副詞一般	where, when, why, how, wherever

表 12　代名詞類のタグ

タグ	意味	実際の例
[pn1*]	不定代名詞・単数	anyone, everything, nobody, one
[pp*]	代名詞	it, I, you, him, her, they, mine, yourself
[pnq*]	WH 代名詞	whom, who, whoever
[ppx*]	再帰代名詞	myself, yourself, herself, themselves

表 13　その他のタグ

タグ	意味	実際の例
[mc*]	数詞	one, two, three, sixes, 40-50
[md*]	助数詞	first, second, last, next
[cc*]	等位接続詞	and, or, but
[cs*]	従属接続詞	if, because, unless, so, for
[uh*]	間投詞	oh, yes, um
[y*]	句読点など	, . ? ! : ;

練習問題 8

数多くの言語学研究の対象となってきた英語の二重目的語構文（ditransitive construction）と to-与格構文（*to*-dative）を COCA で検索するためのパターンを考えてみよう．

二重目的語構文：動詞 + 代名詞 + 冠詞 + 名詞
to-与格構文：動詞 + 冠詞 + 名詞 + to + 代名詞

なお，Goldberg（2011）では，COCA から採取した上記 2 種の構文データを用いて，両構文の違いについての議論を行っている．

4　コロケーション検索の基礎

画面左側のパネルで，Collocates の検索モードを選ぶと，コロケーション検索を行うことができる．
ここで注意する必要があるのは次のことである．

1. 検索の中心語となるのはあくまで Word/phrase の方であり，コロケーションの幅の指定は，中心語から「左右に何語以内」という形式で行う．
2. 検索結果として画面右側にリストアップされるのはコロケーションの方である．
3. 複数のスロットから成る中心語句の左側のコロケーションを調べるときは，中心語句の最も左の語を起点として（中心語句に含まれる，その語より右の語も数えた上で）コロケーションの幅指定を行う．

以下にコロケーション検索の例を示す．Word/phrase だけでなく，Collocates を指定することで，該当する例をすべて採取するだけでなく，「どのような語句がどれくらいの頻度で共起しているか」を明確にすることができる．以下にコロケーション検索のいくつかの例を示す．「数字/数字」という指定は「左に何語/右に何語」を共起範囲として設定するかである．

(1) Word/phrase: `[thick]`

Collocates: `[nn*]` 0/4

thick（変化形含む）に名詞が後続 ⇒ hair, skin, layer, glasses, smoke

		CONTEXT	FREQ	
1	☐	HAIR	1496	
2	☐	SKIN	667	
3	☐	LAYER	507	
4	☐	SMOKE	452	
5	☐	GLASSES	443	

(2) Word/phrase: `[look] into`

Collocates: `[n*]` 0/6

look+into の後に名詞 ⇒ eyes, face, mirror, future, matter

		CONTEXT	FREQ	
1	☐	EYES	2713	
2	☐	FACE	486	
3	☐	MIRROR	426	
4	☐	FUTURE	314	
5	☐	MATTER	240	

(3) Word/phrase: `[feel] like`

Collocates: `[vvg*]` 0/4

feel の後に動名詞が続くパターン ⇒ going, getting, talking, going, trying

		CONTEXT	FREQ	
1	☐	GOING	2001	
2	☐	GETTING	891	
3	☐	TALKING	474	
4	☐	GON	410	
5	☐	TRYING	397	

(4) Word/phrase: `[memory]`

Collocates: `[j*]` 2/0

memory（変化形含む）に形容詞が先行するパターン ⇒ fond, recent, good, short-term, bad

		CONTEXT	FREQ	
1	☐	[FOND]	1158	
2	☐	[RECENT]	1061	
3	☐	[GOOD]	895	
4	☐	[SHORT-TERM]	823	
5	☐	[BAD]	702	

(5) Word/phrase: `beautiful`

Collocates: `[n*]` 0/4

beautiful に名詞が後続

⇒ woman, thing, girl, day, place

		CONTEXT	FREQ	
1	☐	[WOMAN]	5273	
2	☐	[THING]	2531	
3	☐	[GIRL]	2384	
4	☐	[DAY]	2250	
5	☐	[PLACE]	1300	

(6) Word/phrase: `smile.[n*]`

Collocates: `[j*]` 5/5

名詞 smile の左右 5 語以内の形容詞

FREQUENCY でソート ⇒ big, little, small, wide, warm

RELEVANCE でソート ⇒ rueful, wry, toothy, beatific, gap-toothed

		CONTEXT	FREQ	
1	☐	BIG	1546	
2	☐	LITTLE	813	
3	☐	SMALL	680	
4	☐	WIDE	515	
5	☐	WARM	460	

		CONTEXT	FREQ	ALL	%	MI	
1	☐	RUEFUL	129	524	24.62	8.97	
2	☐	TOOTHY	114	507	22.49	8.84	
3	☐	WRY	394	1779	22.15	8.82	
4	☐	BEATIFIC	70	379	18.47	8.56	
5	☐	GAP-TOOTHED	33	189	17.46	8.48	

練習問題 9

次のコロケーションについて調べてみよう．

(a) 名詞 happening の直前に共起する形容詞
(b) at last の左右それぞれ 4 語以内に共起する語彙動詞
(c) finally の左右それぞれ 4 語以内に共起する語彙動詞

ヒント 形容詞は `[j*]`，語彙動詞は `[vv*]`

4.1 相互情報量（MI）スコアについて

English-Corpora.org コーパスのコロケーション検索では, 頻度（frequency）の他に関連度（relevance）を使用した結果のふるい分けやソートが可能になっている．具体的には Sort/Limit で SORTING を RELEVANCE にすることで関連度ソートに切り替わり，同時に MINIMUM に数値を入力することで頻度や関連度の下限を設定できる．では関連度とは具体的には何を意味するのか．English-Corpora.org のコーパスでは相互情報量（Mutual Information, MI）を関連度として扱っている．

相互情報量とは，任意の語が与えられたときに，どの程度，その共起語が予測できるかという指標であり，以下の式で算出される．例えば，BNC における purple と color という 2 語の MI スコアを算出するなら：

$$MI = log((AB * sizeCorpus)/(A * B * span))/log(2)$$

- A = 中心語の頻度 (e.g. purple): 1262
- B = コロケーションとなる語の頻度 (e.g. color): 115
- AB = 両者のコロケーション頻度 (e.g. color near purple): 24
- $sizeCorpus$= コーパスのサイズ（すなわち今回の場合 BNC の収録語数）: 96,263,399
- $span$ = 語と語の間隔 (e.g. 中心語から左に 3 語分，右に 3 語分）: 6
- $log(2)$ = 2 の log_{10}: 0.30103

$$MI = log((24 * 96,263,399)/(1262 * 115 * 6))/0.30103 = 11.37$$

相互情報量は，いずれかまたは両方の語句の出現頻度が大きいと値が小さくなる性質がある（石川 2012; Gries 2013）．そのため，それ自体が高頻度語句であるような要素の組み合わせは比較的下位にランクされる．一方，頻度が低くてもほとんど常に共起するような要素の相互情報量は大きくなる．したがって，English-Corpora.org のコーパス上で関連度を利用して検索を行う際には，Sort/Limit で FREQUENCY の下限値を設けるなどして（出現頻度が著しく低いパターンが上位を占めたりしないよう）バランスを取ること

が必要である.

4.2 コロケーション検索の応用

English-Corpora.org のコーパスは統語解析されていないが，コロケーション検索と品詞タグを上手く使えば，名詞句や関係節といったものをある程度擬似的に表現できる.

(7) 構文 what|all RELATIVE-CLAUSE do BE V の用例を採取したい

Word/phrase: `do [be] [v*]`

Collocates: `what|all` 8/0

⇓

all you have to be able to do is take a bullet

all you folks have got to do is get a moose in your barn

all you have to do is remember to remember

(8) 構文 V + NP + into + v-ing の V として現れる動詞で頻度の高いものを知りたい

Word/phrase: `into [v?g*]`（→ 動詞・ing 形）

Collocates: `[vv*]`（→ 語彙動詞） 4/0

OPTION で GROUP BY を LEMMA に設定する.

⇓

`[GO]`, `[TALK]`, `[COME]`, `[TRICK]`, `[PUT]`, etc.

(9) 構文 expect NP to V の V として現れる動詞で関連度の高いものを知りたい

Word/phrase: `[expect] [a*]|[d*]|[n*]|[p*]`

Collocates: `[v?i*]` 0/3

SORTING を RELEVANCE に，MINIMUM の FREQUENCY を任意の数（例えば 50）に設定.

また OPTION で GROUP BY を LEMMA に設定する.

⇓

`[HAPPEN]`, `[WIN]`, `[FOLLOW]`, `[CONTINUE]`, `[BE]`, etc.

練習問題 10

次のような構文について調べてみよう

(a) 結果構文（resultative construction）の中で，形容詞 clean を含む例をできるだけ多く採取すると共に，どのような動詞が共起するのかを調査してみよう．結果構文とは，例えば，She wiped the table clean のような文である．

ヒント

動詞 + 冠詞 + 名詞 + clean
動詞を Collocates に指定するのがポイント！
検索オプションで GROUP BY を LEMMAS にすると良い．

(b) 移動使役構文（caused motion construction）の中で，不変化詞 off を含む例をできるだけ多く採取すると共に，どのような動詞が共起するのかを調査してみよう．移動使役構文とは，例えば，Jack sneezed the napkin off the table のような文である．

ヒント

動詞 + 冠詞 + 名詞 + off + 冠詞 + 名詞
動詞を Collocates に指定するのがポイント！

5 English-Corpora.org の検索モードと表示オプション

5.1 検索モード

ここでは，List, Chart, KWIC, Compare という 4 つの検索モードについて少し詳しくみていく．

5.1.1 List 検索モード

List 表示はマッチした文字列やコロケーションのリストを確認するためのモードである．リストアップされた語句をクリックすることで KWIC 表示に切り替わる．KWIC 表示の各行の左側をクリックすると，より詳細な前後文脈を見ることができる．ただし，より高機能な KWIC 表示を求めるときには，検索モードあらかじめ KWIC にしておく必要がある．

Sections の左のチェックボックスをオンにすると，コーパス内の 2 つのセクションにおける語句の生起頻度を比較することができる．COCA に関して言えば，SPOKEN, FICTION, MAGAZINE, NEWSPAPER, ACADEMIC というジャンルの他, 1990 年から 2012 年までの各年があり, 1990-1994, 1995-1999, 2000-2004, 2005-2009, 2010-2012 といった時期のまとまりをセクションとすることも可能である．

結果は画面右にテーブルとして表示される．標準では頻度比に基づいたソート順になっているので，各セク

ションに「特徴的」な語が上位に来る．語の頻度比が 5.0 以上であれば該当行が緑，1.5 以上であれば黄緑で表示される．

可能なセクション比較検索の例

- ACADEMIC と FICTION における `de-*` 動詞
- SPOKEN と NEWSPAPER における動詞過去形 + `over`
- ACADEMIC と FICTION における`*ment`
- 2000-2009 と 1990-1999 において `green` と共起する名詞
- NEWS と SPOKEN における形容詞 + `track`
- ACADEMIC と FICTION における `chair` と共起する名詞

画面左の Sort/Limit の SORT BY によって，結果がどのようにソートされるかを指定できる．デフォルトでは FREQUENCY の降順であるが，RELEVANCE によるソートも可能である．RELEVANCE で用いられるのは，相互情報量（MI）スコアであり，これは 2 つの語がどれくらい「緊密に」関係しているかを示す．

また，Sort/Limit の MINIMUM を FREQUENCY ないしは MUTUAL INFO に設定して，検索結果に下限を設けることができる．MI スコアについては，通常，3.0 以上あれば当該の語句間に「強い結びつきがある」と考えられる．

5.1.2 Chart 検索モード

Chart 表示のモードでは，コーパスのセクションごとにマッチした語句が生起する総頻度を棒グラフで確認できる．各棒グラフをクリックすると，当該のセクションにおける語句の KWIC が表示される．

5.1.3 KWIC 検索モード

コンコーダンスを確認するのに最適なのが KWIC 表示である．Keyword in Context の形式で表示される他，このモードでは，中心語句と周辺語が品詞ごとに色分けされる．また，1 つあるいは複数のスロットを指定し，結果全体をソートできる．

KWIC 検索モードでソートを行う方法は次の通りである．まず，検索の前に，Sort/Limit で，ソートの基準（ALPHABETICAL または DATE/GENRE）指定する．その後，次に示す SORT 設定のコントローラーを利用してソートする対象となるスロットを指定する．

1. 中心語句（= Word/phrase ボックス内の文字列に対応する語句）の左方向の 3 語でソートする．
2. 中心語句（= Word/phrase ボックス内の文字列に対応する語句）の右方向の 3 語でソートする．

3. 中心語句を含む文字列の中で，3 つまでスロットを選んで自由にソート方法を決定する．
4. ソートのオプションをリセットする．

上記の設定後，RE-SORT をクリックするとソートが実行される．

5.1.4 Compare 検索モード

検索モードで Compare を選択すると，Word1，Word2 という 2 つの異なる語句について結果を比較できるようになる．

下は形容詞 small および little の直後に生起する名詞を比較した結果である．ここでは，Sort/Limit で MINIMUM の値を下記のように設定し，マッチさせるコロケーションに頻度の下限を設けている．

1) 中心語のうち共起頻度が大きい方とは 10 回以上の生起がある．2) 中心語のうち共起数が小さい方とも 4 回以上の生起がある．

WORD 1 (W1): **SMALL** 1 (.55) 3

	WORD 5	W1 6	W2 7	W1/W2 8	SCORE 9
1	BUSINESSES	1907	12	158.92	289.16
2	MINORITY	340	4	85.00	154.67
3	AMOUNT	1400	20	70.00	127.37
4	SIZE	454	9	50.44	91.79
5	BOWL	1267	28	45.25	82.34
6	NUMBER	1943	44	44.16	80.35
7	PORTION	486	13	37.38	68.03

WORD 2 (W2): **LITTLE** 2 (1.82) 4

	WORD	W2	W1	W2/W1	SCORE
1	BIT	15571	75	207.61	114.10
2	SISTER	818	4	204.50	112.39
3	BROTHER	975	5	195.00	107.17
4	BILL	361	5	72.20	39.68
5	GIRLS	1327	23	57.70	31.71
6	EXPERIENCE	225	4	56.25	30.91
7	MONEY	842	15	56.13	30.85

上記の図中で番号の付いた箇所は，それぞれ下のような意味を持つ．

[1-2] 検索語句

[3-4] 語の出現頻度比（little を 1 としたとき，small は 0.55 であり，small を 1 としたとき，little は 1.82 である．これらは頻度データ 145,028 vs 263,893 がもとにとなっている．）

[5] 1 のコロケーションをランク順にならべたもの

[6-7] W1 または W2 のコロケーション頻度

[8] 6 と 7 の比率

[9] 8 の 3 に対する比率（＝対立語に対して，コロケーション頻度が「何 %」であるか）

以下は 2 つの語句を比較した例である．

(10) Word/phrase: `hot` vs. `warm`
Collocates: `[nn*]`
⇒ tub, tips, shower vs. glow, embrace, person

(11) Word/phrase: `boy` vs. `girl`
Collocates: `[j*]`
⇒ growing, rude vs. sexy, working

(12) Word/phrase: `utter.[j*]` vs. `sheer.[nn*]`
⇒ silence, despair vs. beauty, joy

(13) Word/phrase: `ground.[n*]` vs. `floor.[n*]`
Collocates: `[j*]`
⇒ common, solid vs. concrete, dirty

5.2 English-Corpora.org コーパスの詳細オプション

画面左下の CLICK TO SEE OPTIONS をクリックすると 4 つのオプション項目が表れる．

5.2.1 # HITS

表示される検索ヒット件数を指定する．デフォルトは 100.

5.2.2 GROUP BY

検索結果のグループ化の方法を指定する．

- WORDS
  デフォルトの指定．語の形式によってグループ化して表示する．
- LEMMA
  結果がレンマでグループ化される（例えば swim, swimming, swam はすべて同じレンマのバリエーションと見なされる）

- NONE (SHOW POS)
  同じ形式の語が複数の品詞で現れているとき，それぞれを別の要素として扱う．通常は使わないオプションだが，KWIC で特定の品詞だけ表示したいときにはこれを選ぶ必要がある．
- BOTH WORDS
  コロケーション検索において有用．例えば，pretty の同義語と flower の同義語との共起を調べるとき，pretty flower, beautiful roses といった組み合わせをすべて列挙できる．
- BOTH LEMMA
  上記と同じことをレンマを単位に行う．

5.2.3 DISPLAY

頻度表示のフォーマットを指定する．

- RAW FREQ
  デフォルトの指定．コーパスの各セクションのトークン数を表示.
- PER/MIL
  100 万語あたりのトークン数を表示．異なるサイズのセクション間で比較を行う際に有用.
- RAW FREQ+
  RAW FREQ + PER/MIL の順で表示.
- PER/MIL+
  PER/MIL + RAW FREQ の順で表示.

5.2.4 SAVE LISTS

後に続く検索で使用できるよう，結果をユーザー・リストに保存できるようにする．例えば，beautiful の同義語検索の結果をもとに，別の語彙を加えたりして，オリジナルの [beautiful] リストを作成できる．デフォルトの指定は NO である．

参考文献

Davies, Mark (2010) "The Corpus of Contemporary American English as the first reliable monitor corpus of English," *Literary and Linguistic Computing*, Vol. 25, No. 4, pp. 447–464.

Goldberg, Adele E. (2011) "Corpus evidence of the viability of statistical preemption," *Cognitive Linguistics*, Vol. 22, No. 1, pp. 131–153.

Gries, Stefan Th (2013) *Statistics for Linguistics With R: A Practical Introduction*: Berlin: Mouton de Gruyter, 2nd edition.

Lindquist, Hans (2010) *Corpus Linguistics and the Description of English*: Edinburgh: Edinburgh University Press.

石川慎一郎 (2012) 『ベーシックコーパス言語学』, 東京：ひつじ書房.