COCA を利用した言語データの採取と統計処理の基本

同志社大学 グローバル・コミュニケーション学部
長谷部 陽一郎
yhasebe@mail.doshisha.ac.jp

1 はじめに

この文書では次の 3 つのことを目的とする.

1. 大規模英語コーパス COCA と BYU コーパス群の概要について知る
2. COCA のさまざまな機能とその使い方を学ぶ
3. COCA で得られたデータに簡単な統計的検定を施す方法を学ぶ

2 COCA と BYU コーパス群

2.1 COCA の概要

Corpus of Contemporary American English (COCA) は，Brigham Young University の Mark Davies が開発しているオンラインでアクセス可能な BYU コーパス群（`http://corpus.byu.edu/`） 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.

- URL: `http://corpus.byu.edu/coca/`
- 収録語数：4 億 5 千万語
- 言語：アメリカ英語
- 期間：1990 年〜2012 年

4 億 5 千万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．また 1990 年から 2012 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 COCA の機能と特徴

COCA でできること：

- 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（これらを組み合わせることもできる）を行う．
- 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（例：little/small, democrats/republicans, men/women）のコロケーションを比較する．
- 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（ただし今回のワークショップでは，ユーザー・リストを実際に使う方法については扱わない）．

なお，COCA には ICE-GB コーパス（British component of International Corpus of English）のような統語解析は施されていないが，すべての語に品詞情報が付与されているので，ある程度は意味や統語的性質を問うような検索が可能になっている．

2.3 その他の BYU コーパス

Mark Davies が開発またはインターフェイスを提供しているコーパスとしては，COCA の他に次のようなものがある．またここに記載したオンライン・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（`http://www.ngrams.info/`）．

2.3.1 Corpus of Historical American English (COHA)

- URL: `http://corpus.byu.edu/coha/`
- 収録語数：4 億語
- 言語：アメリカ英語
- 期間：1810 年〜2009 年

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．

2.3.2 TIME Magazine Corpus（TIME）

- URL: http://corpus.byu.edu/time/
- 収録語数：1 億語
- 言語：アメリカ英語
- 期間：1923 年〜2006 年

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．

2.3.3 Corpus of American Soap Operas（SOAP）

- URL: http://corpus2.byu.edu/soap/
- 収録語数：1 億語
- 言語：アメリカ英語
- 期間：2001 年〜2012 年

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．

2.3.4 British National Corpus (BYU-BNC)

- URL: http://corpus.byu.edu/bnc/
- 収録語数：1 億語
- 言語：イギリス英語
- 期間：1980 年代〜2000 年代

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．BYU Corpora 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 BYU コーパスとデータ形式の互換性がある．

BYU-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．またレジスター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．

2.3.5 Strathy Corpus（STRATHY）

- URL: `http://corpus2.byu.edu/can/`
- 収録語数：5 千万語
- 言語：カナダ英語
- 期間：1970 年代〜2000 年代

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．BYU-BNC と同様，他の BYU Corpora と共通したデータ・フォーマットを採用している．

2.3.6 Global Web-Based English (GloWbE)

- URL: `http://corpus2.byu.edu/glowbe/`
- 収録語数：19 億語
- 言語：20 カ国の英語
- 期間：2012 年〜2013 年

英語使用国 20 カ国の 18 億のウェブページから採取した 190 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．

練習問題 1

COCA，COHA，TIME，SOAP，BYU-BNC，STRATHY，GloWbE の各コーパスの URL にアクセスし，適当な検索文字列を入れて試してみよう．

3 COCA の機能と使い方

ここでは実際に COCA を使用する際に役立つ様々な検索の方法や手順をみていく．基本的な操作方法は COCA 以外の BYU コーパス群でも有効である．

3.1 検索シンタックス

COCA 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」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 1　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
[user:list]	ユーザー・リスト	[userlist:clothes]	tie, shirt, blouse

練習問題 2

次のような語句・構文を検索してみよう．

(a) 形容詞 + performance
(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*]

COCA ではワイルドカードを用いた検索が可能になっている．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 2　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．

(a) 接頭辞 un と接尾辞 able を共に含む語
(b) under から始まる語
(c) holic で終わる語

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 3　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

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だけを抽出するような場合に役立つ．例えば表 4 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 4　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 5　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
`[[user:list]]`	ユーザーリスト＋レンマ	`[[userlist:clothes]]`	tie, tying, socks, socked, shirt
`[[=word]].[pos]`	同義語＋レンマ＋品詞	`[[=clean]].[v*]`	mop, scrubs, polishing
`[[user:list]].[pos]`	リスト＋レンマ＋品詞	`[[userlist:clothes]].[n*]`	tie, sock

練習問題 6

次のような語句を検索してみよう.

(a) help（動詞・レンマ）の同義語
(b) advise（動詞・レンマ）の同義語

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.

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．下記にいくつかの例を示す.

表 6　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*] [userlist:clothes].[n*]`	put on her hat, putting on my pants

練習問題 7

興味のある英語の語句や構文を自由に検索してみよう．また，COCA 以外の BYU コーパス群でも同様の検索文字列を入力して試してみよう．

3.2　CLAWS7 タグセット

ここでは, COCA の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．COCA 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．
- COCA のシステム上では `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 7　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 8　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*]	副詞	kindly, else, namely, very
[np*]	固有名詞	IBM, Andes, Smith, Sunday, October
[nn*]	普通名詞	sheep, cod, headquarters, book, girls

表 9　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 10　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 11　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 12　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

Goldberg（2011）では，COCA から 2 種の構文データを採取している．しかし，現在，COCA で複数語の検索を行う際には少なくとも 1 つのスロットが頻度 1 千万語以下のものでなければならないため，通常のオンライン・インターフェイスで検索を遂行することはできない．そこで，ここでは動詞（[v*]）の部分を [give] に変えて検索してみよう．

3.3 コロケーション検索の基礎

COCA の画面左側のパネルで，WORDS フォームにメインの検索文字列を入力すると通常の検索ができる．これに加えて，COLLOCATION フォームにも文字列を入力すると，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 WORDS 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，中心語句の最も右側のスロットを起点として（中心語句に含まれる他の語も数えた上で）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．

(1) WORDS: `[thick]`
COLLOCATION: `[nn*]` 0/4
thick（変化形含む）に名詞が後続 ⇒ glasses, smoke

(2) WORDS: `laugh.[n*]`
COLLOCATION: `[j*]` 5/5
名詞 laugh の左右 5 語以内の形容詞
FREQUENCY でソート ⇒ good, little, big
RELEVANCE でソート ⇒ hearty, scornful

(3) WORDS: `look into`
COLLOCATION: `[nn*]` 0/6
look+into の後に名詞 ⇒ eyes, future

(4) WORDS: `work|job`
COLLOCATION: `hard|tough|difficult` 4/0
work ないしは job の前に hard，tough, または difficult が共起 ⇒ hard, tough, difficult

(5) WORDS: `[feel] like`
COLLOCATION: `[vvg*]` 0/4
feel の後に動名詞が続くパターン ⇒ crying, taking

(6) WORDS: `[=gorgeous]`
COLLOCATION: `[n*]` 0/4
gorgeous の同義語に名詞が後続
GROUP BY WORDS を選択 ⇒ woman, face
GROUP BY BOTH WORDS を選択 ⇒ attractive woman, beautiful day

(7) WORDS: `[=beautiful]`
COLLOCATION: `[=face].[n*]` 5/0
beautiful の同義語に名詞 face の同義語が先行
GROUP BY WORDS を選択 ⇒ happy, delighted
GROUP BY BOTH WORDS を選択 ⇒ happy//child, delighted//boy

WORDS だけでなく，COLLOCATION を指定することで，該当する文字列を含む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．

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

3.4 コロケーション検索の応用

BYU corpora 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．

(8) V + NP + into + v-ing
WORDS: `into [v?g*]`（→ 動詞・ing 形）
COLLOCATES: `[vv*]`（→ 語彙動詞） 4/0
e.g. talked her into staying
⇓
`[vv*] her into [v?g*]`
`[vv*] the people into [v?g*]`
`[vv*] my best friend into [v?g*]`

(9) what|all RELATIVE-CLAUSE do BE V
WORDS: `do [be] [v*]`

COLLOCATES: `what|all` 8/0
e.g. what|all he wants to do is complain

⇓

```
what|all he wants to do [be] [v*]
what|all they expected Fred to do [be] [v*]
what|all any of these crazy people can do [be] [v*]
what|all your best friend can possibly hope to do [be] [v*]
```

(10) expect NP to V
WORDS: `[expect] [a*]|[d*]|[n*]|[p*]`
COLLOCATES: `[v?i*]` 0/8

⇓

`[expect] them to [v?i*]` (them = `[p*]` ← 代名詞)
`[expect] Bill Clinton to [v?i*]` (Bill = `[np*]` ← 固有名詞)
`[expect] those six people to [v?i*]` (those = `[d*]` ← 指示詞)
`[expect] the people in Florida to [v?i*]` (the = `[a*]` ← 冠詞)

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．移動使役構文とは，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！

(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．結果構文とは，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！

3.5 検索結果の表示

COCA の検索結果表示のモードには，LIST, CHART, KWIC, COMPARE という 4 つの種類がある．

3.5.1 LIST 表示

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，画面左上の DISPLAY 設定であらかじめ KWIC を選んでおく必要がある．

画面左の SECTIONS の SHOW にチェックを入れて，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 に含まれるセクションには 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 SORTING AND LIMITS 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，SORTING AND LIMITS 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．

練習問題 11

global warming という表現の 1990 年から 1999 年までの生起頻度と 2000 年から 2009 年までの生起頻度を比較してみよう．また，greenhouse effect という表現についても同様に調べてみよう．

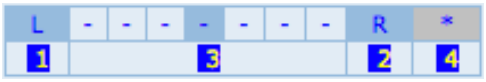
3.5.2 CHART 表示

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．

3.5.3 KWIC 表示

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 表示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，画面左の DISPLAY 設定で，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（= WORDS 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（= WORDS 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．

3.5.4 COMPARE 表示

画面左上の DISPLAY 設定で COMPARE 表示モードを選択すると，SEARCH STRING の WORDS のボックスが 1 つ増え，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，SORTING AND LIMITS 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．

(11) WORDS: `hot` vs. `warm`
COLLOCATES: `[nn*]`
⇒ tub, tips, shower vs. glow, embrace, person

(12) WORDS: `boy` vs. `girl`
COLLOCATES: `[j*]`
⇒ growing, rude vs. sexy, working

(13) WORDS: `utter.[j*]` vs. `sheer.[nn*]`
⇒ silence, despair vs. beauty, joy

(14) WORDS: `ground.[n*]` vs. `floor.[n*]`
COLLOCATES: `[j*]`
⇒ common, solid vs. concrete, dirty

練習問題 12

動詞 rob（レンマ）と動詞 steal（レンマ）がそれぞれどのような目的語（名詞）と結びつきやすいかを調べてみよう．（その際，MINIMUM の値を適宜調整すること．）

3.6 COCA 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．

3.6.1 # 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.

3.6.2 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
  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．

3.6.3 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.

3.6.4 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.

4 頻度の比較と有意差検定

以下では，COCA から得られた語句の頻度情報に対して，それらが「統計的に有為」(statistically significant) であるかどうかを確かめる方法を学ぶ．扱うのは，言語データを扱う様々な分野の研究（理論言語学研究，言語発達研究，第二言語習得研究，etc.）でもっともよく使われる手法の 1 つ「カイ 2 乗検定」(chi square test) である.

4.1 but と however

ここでは，Lindquist (2009) にならって，COCA における but と however の生起頻度の経年変化を題材に，カイ 2 乗検定を実際に行っていくことにする．これら 2 つの語は相補的な関係にあり，一方の生起頻度が増加すると他方は減少すると考えられる．そこで 2 語の頻度を COCA の 1990-1994 という期間と 2005-2009 という期間でそれぞれ調査してみたところ，次のようなデータが得られた.

表 13　1990-1994 と 2005-2009 における but と however の粗頻度値

	1990-1994	2005-2009	total
but	457,561	460,326	917,887
however	40,205	31,459	71,664
total	497,766	491,785	989,551

but は 1990-1994 に比べて 2005-2009 では 457,561 から 460,326 と数が若干増加しており，however のほうは 1990-1994 の 40,205 から 2005-2009 の 31,459 と若干減少している．

しかし，2 つの時代ではコーパスに含まれるテキストの量自体が異なるため，上記の観察から，「1990-1994 と 2005-2009 を比較すると，but の生起頻度は増加し，however の生起頻度は減少している」と即断することはできない．そこで，「100 万語あたりの生起頻度」（frequency per million words = PM）で比較してみることにする．

表 14　1990-1994 と 2005-2009 における but と however の PM 値

	1990-1994	2005-2009	total
but	4,399	4,511	8910
however	386	308	694
total	4785	4819	9604

これをみると，やはり上の観察は正しいことがわかる．すなわち，1990-1994 と 2005-2009 を比較すると，but の生起頻度は増加し，however の生起頻度は減少している．しかし，言葉の使用というものは様々な個別的・具体的条件に左右されるもので，コーパスは世の中の言語使用のほんの一部を切り取ったものに過ぎない．だとすれば，上の表 13 および表 14 が示す but と however の使用に関する 2 期間の差も今回「たまたま」生じたものである可能性がある．

ここでは「1990-1994 と 2005-2009 を比較すると，but の生起頻度は増加し，however の生起頻度は減少している」ということを明らかにしたい．そこで上記のような 2 語の頻度差が単なる偶然の産物ではないという確証を得るために，カイ 2 乗検定が役立つ．

4.2　帰無仮説について

カイ 2 乗検定を始めとする様々な統計的検定の手法は「帰無仮説」（null hypothesis）と呼ばれる考え方に基づいている．通常，研究者は異なる対象グループ間に何らかの「差がある」ということを証明しようとする．しかし，様々な程度があり得る「差」というものの存在を直接証明することは，実は困難である．一方で，「実際には差がない可能性」，言い換えると「差があったとしても偶然そうなった可能性」を求めることはさほど困難ではない．そこで，まずはグループ間で「差がない」可能性を検証する．その可能性（probability = p

値）が十分に低ければ，晴れて「差がある」と結論付けられるのである．

なお，p 値が十分に低いかどうかを判断するためには通常，0.05 ないしは 0.01 という数が用いられる．したがって，$p < 0.05$ もしくは $p < 0.01$ であれば，「帰無仮説を棄却」し，有為差の存在がみとめられることになる．これら 0.05，0.01 という数字は「有意水準」（significance level）と呼ばれ，言語学ではどちらかというと（より厳しい）0.01 が好まれるようである．

4.3 R を用いた統計処理

さて，カイ 2 乗検定で p 値を求めるには「期待値」と呼ばれる数値を求める必要がある．また，「自由度」と呼ばれる値や，期待値から導きだされる「カイ 2 乗値」も必要である．これらを算出する計算はそれほど複雑でなく，手計算が可能である．また，カイ 2 乗値と自由度の値を使って最終的に求めるべき p 値は，書籍やウェブで入手可能な「カイ 2 分布表」を用いて調べることができる．しかし，ここでは，フリーソフトウェアの「R」を用いたカイ 2 乗検定の方法を紹介する．

R は以下のサイトから Windows，Mac OSX，Linux の各 OS 向けパッケージがダウンロードできる．

`http://cran.r-project.org`

R のインストール方法などの詳細は，ウェブ上の情報の他，入門用の書籍が英語，日本語を問わず多数出版されているので，それらを参照されたい．また，言語研究に R を使うことに特化した入門書も存在する（Gries 2013）.

では実際にやってみよう．R を起動して，表 14 の数値から次のように「クロス表」（contingency table）を構成する．

(15)
```
x <- matrix(c(4399, 4511, 386, 308), ncol=2, byrow=T)
colnames(x) <- c("1990-1994", "2005-2009")
rownames(x) <- c("but", "however")
chisq.test(x, correct=F)
```

1 行目では，4 つの観測値からクロス表を作成している．また，2 行目と 3 行目では表の行と列に名称を与えている．実際には，行と列に名前を与える必要はないが，これを行っておくと，クロス表自体を表示させたときに，それぞれの列や行で表されている内容がわかりやすくなる．最後の 4 行目では実際のカイ 2 乗値，自由度，および p 値の計算を行っている．

結果として表示されるのは次の文字列である．

(16)
```
Pearson's Chi-squared test
data:  x
X-squared = 10.0542, df = 1, p-value = 0.00152
```

これによると，入力したクロス表から導きだされるカイ 2 乗値（$\chi - squared$）は 10.0542 であり，自由度

(df) が 1 のときの p 値 ($p-value$) は 0.00152 である．これは有意水準 0.01 よりもずっと小さい数字なので，今回のデータが偶然の産物である可能性は無視できる程小さく，したがって，1990-1994 と 2005-2009 という 2 つの時期における but と however の頻度分布には「差がある」と見なすのが妥当と結論づけられる．

以上は Lindquist（2010）で示されているカイ 2 乗検定の手順を，基本的にそのまま実施したものである．ただし，1 つ注意しておくべき点がある．ここではコーパスから得られた粗頻度値を「100 万語あたりの頻度値」に変換した後にカイ 2 乗検定をおこなった．しかしながら，このような調整済頻度をカイ 2 乗検定の際に用いることは本来推奨されない．(Cf. 石川 2012)．ここでは，何らかの理由で R を使えず，手計算を行ったり，許容される数値に上限が設けられている別のソフトウェアで計算を行うような場合を想定して調整済頻度を用いた．

その他，今回のケースでは当てはまらないが，クロス表のいずれかのセルで期待値（＝当該の列の合計値を総頻度比にあわせて分配した値）が 5 未満になる場合もカイ 2 乗検定は使えないとされている．統計的検定にはこのように適用のための条件があるため，結論を急がず，手順を確かめながら丁寧に作業を行うことが必要である．

練習問題 13

Taylor（2012: 2）は，論理的には対等な関係にある success/failure という 2 つの名詞のうち，形容詞 total に後続しやすいのは failure の方であり，それが英語を母語とする話者の自然な感覚であると述べている．カイ 2 乗検定を使って，COCA のコーパス・データからこのことが裏付けられるか確認してみよう．

ヒント 1

次の 4 つの要素からクロス表を作ることから始めよう．
(1) total と success の共起頻度.
(2) total と failure の共起頻度.
(3) success の全体頻度から (1) を引いた値.
(4) failure の全体頻度から (2) を引いた値.

ヒント 2

R をインストールして実際に試せない場合は，ウェブ上でカイ 2 乗検定を実施できるサービスがいくつかあるので，それらを使うとよい．例えば次のサイトは非常に使いやすいインターフェイスを備えている．

GraphPad: Analyze a 2x2 contingency table
`http://graphpad.com/quickcalcs/contingency1.cfm`

ヒント 3

ほとんど問題の解答になってしまうが，R で p 値を求めると $2.2e-16$ という結果が得られる．これは 2.2×10^{-16} という意味であり，有意水準 0.01 よりはるかに小さい数である．本練習問題では，この数を実際に導き出すことを目標としてほしい．

参考文献

Davies, M. (2010). “The Corpus of Contemporary American English as the first reliable monitor corpus of English”. *Literary and Linguistic Computing*, **25**(4) pp.447–464.

Goldberg, A. E. (2011). “Corpus evidence of the viability of statistical preemption”. *Cognitive Linguistics*, **22**(1) pp.131–153.

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

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.

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