

### 神戸大学 国際コミュニケーションセンター

## 石川慎一郎研究室

Dr. Shin Ishikawa, Kobe University



# コーパス言語学入門 AntConcを使ってみよう

2007/4/15 Antconc最新バージョンV3.2.1にあわせて記述を全面改訂。 クラスター分析,共起語リスト分析に関する解説を新設。

#### Contents

1. はじめに
 2. 調べたいデータを設定しよう
 3. KWIC検索を試そう
 4. Collocatesで共起語を一気に調べる
 5. 特定語を含むクラスターを探そう
 6. コーパスから語彙表を作ろう
 7. コーパスから特徴語を抽出しよう
 8. データの中での語の位置関係を見る
 9. おわりに

#### 1. はじめに

Antconcは早稲田大のLAURENCE ANTHONY氏が開発されたコンコーダンス・ソフトウェアです。Antconcは優れた高機能ソフトウェアで,初心者であっても直観的に使用することが出来ます。またフリーウェアなので気軽に使うことが出来るのも魅力です。

ここでは、三つの簡単な調査例を通してその基本的な使い方をご紹介します

### 2. 調べたいデータを設定しよう

ダウンロードした実行ファイルをダブルクリックすると、下のような画面が出ます.

🐼 AntConc 3.2.1w (W	indows) 2007							
File Global Settings Tool P	references About							
Corpus Files	Concordance	Concordance Plot	File View	Clusters	Collocates	Word List	Keyword L	ist
	Hit KVVIC				47. 47.		File	N.
							> <	28
	Search Term 🔽	Words   Case	Regex		Concordance	Hits	Search Wind	ow Size
Total No. 0	Stort Ston	Cost		cea	U.	1	50 <b>T</b>	
Files Processed	Start Stop	Sur						Save Window
Reset	Level 1 0	😫 🗆 Level 2 🛛	🛨 🗖 Level	3 🛛 🜩				E <u>×</u> it

コンコーダンスソフトでは最初に使うテキストファイルを「設定」してやる必要があります.

画面上部のツールバーに「File」「Global Settings」などの文字が見えますね. ここでは、左端のFileを開きます.

様々なプルダウンメニューが表示されますが,重要なのは下の二つです.

■ Open File (s) は1つ(または複数の)テキストファイルを開くときに、
 ■ Open Dirは複数のファイルをおさめたフォルダ(Directory)ごと調べたいときに使います.

ここでは「Open Dir」を選択してみましょう.

すると「フォルダの参照」ウィンドウが開いて自分のパソコン内部を検索できます。 ここでは、イギリス英語100万語を集めたFLOBコーパスを検索対象に設定しましょう。

フォルダの参照	? 🔀
Open a Directory of ".txt Files" C:¥Bun Lang¥CPS¥ICAME¥91 US <u>U</u> K¥FLOB(91英)	
FLOB(91英)	
■ Calign EJ_align EXAM CAME CAME CAME CAME CAME CAME FLOB(91英) CAME FLOB(91英) CAME FLOB(91英) CAME FLOB(91英) CAME FLOB(91英) CAME C	
OK ¥	ャンセル

上記の画面で,調べたいテキストファイルを納めたフォルダがある場所を選択し,「OK」を押せば検索対象フォルダが設定されます。 メイン画面の左端にある「Corpus Files」のボックスに,フォルダ内に含まれるテクストファイルの一覧が表示されているはずです。

3. KWIC検索を試そう

「Concordance」とは、検索対象語を含むコンコーダンスラインを一括表示する機能のことです。

「Concordance」検索はKWIC(クイック)検索とも呼ばれます。

KWICとはKey Word in Contextの略で,検索対象語(NODEと呼びます)がどのような前後関係(コンテキスト)の中で使われているかを一覧表示します. Ant Concの一番下にあるSearch Termsのボックスに英単語を入力してみましょう.

ここではgreenという単語を例として入れてみます. その後でボックスの右側のStartボタンを押せば完了です.

すると次のような結果画面が出ます.

### 🐲 AntConc 3.2.1w (Windows) 2007

#### File Global Settings Tool Preferences About

Corpus Files	Concordance Concordance Plot File View Clusters Collocates Word List Keyword List
FLOB_A.TXT FLOB_B.TXT	Hit KWC File
FLOB_C.TXT FLOB_D.TXT FLOB_E.TXT FLOB_F.TXT FLOB_G.TXT FLOB_H.TXT FLOB_J.TXT FLOB_K.TXT FLOB_L.TXT FLOB_M.TXT FLOB_N.TXT FLOB_P.TXT FLOB_R.TXT	1       passion within the A04 66 Green Party        FLOB_A.TXT         2       n communism for the A04 80 Green Party, there with A07 165 <quote_>"Tee to green I played as weither and the persuaded him to join Theo Green's stable in Synthese stitions, from never wearing green to always wear:       FLOB_A.TXT         4       6 persuaded him to join Theo Green's stable in Synthese stitions, from never wearing green to always wear:       FLOB_A.TXT         6       A27 6 development in the green belt. A27 7 '       FLOB_A.TXT         7       n, British ambassador Andrew Green was summoned to response to always wear:       FLOB_A.TXT         8       1 46 &gt;Now he has given a green light for more et Union A33 81 at Croxley Green last night aft.       FLOB_A.TXT         9       et Union A33 81 at Croxley Green last night aft.       FLOB_A.TXT         11       banking and bowling A37 81 green, Coun Malcolm 1       FLOB_A.TXT         12       at Band on the Wall, to the Green Room's review in grout A41 36 on the 36th green at The Masters       FLOB_A.TXT         8       I More Case Regex       Concordance Hits       Search Window Size</quote_>
Total No. 15 Files Processed Reset	green

greenを含む用例を数多く(Concordance Hitsのボックスを見ると141例!)取得できましたが, このままでは,greenの次にどういう名詞が来やすいのかを調べるのはめんどうです。

画面の下部に注目すると, Kwic Sortというパネルが見つかります。

3つの並べ替え基準を設定することができますが, デフォルトでは第1基準のみが選択されています。

 Example 1
 0

 Level 1
 0

-----

これは出てきた結果を見やすく並べ替えるためのものです.

3つの数字に注目してみましょう。 デフォルトでは「0」になっていますが、その右上にある▲(上矢印ボタン)をクリックすると、最初の「0」は「1R」「2R」・・・のように変わっていきます。 いっぽう、「0」の右下にある▼(下矢印ボタン)をクリックすると、最初の「0」は「1L」「2L」・・・のように変わります。

これらは中心語から見て,右(左)側何語目を基準に並べかえを行うかを設定する機能です。

下記の例で考えてみましょう。

greenを中心に考えると,すぐ左隣は[1L](左=Leftの1番目),右隣りは「1R」(右=Rightの1番目)となり, 以下,「2L」「3L」・・・,「2R」「3R」・・・のように変化していきます。

 Now
 he
 has
 given
 a
 green
 light
 for
 more
 to
 go
 ...

 [5L]
 [4L]
 [3L]
 [2L]
 [1L]
 [1R]
 [2R]
 [3R]
 [4R]
 [5R]

ここでは, greenの次にくる語を調べたいので, 第1基準を「1R」にします。 また,同じ1R語が複数ある場合も考えられるので, 第2基準(Level 2)にもチェックを入れて,ここは「2R」にしましょう。

これで,まずgreenの右隣語を第一基準に,次に右隣り2つ目の語を第2基準として 141のコンコーダンス行を並べ替えることになります。

ここで黄色の「Sort」ボタンを押します.

#### 🐲 AntConc 3.2.1w (Windows) 2007

### 

#### File Global Settings Tool Preferences About

Corpus Files	Concordance Concordance Plot   File View   Clusters   Collocates   Word List   Keyword	rd List
FLOB_B.TXT	Hit KWIC F	File 👌
FLOB_C.TXT FLOB_D.TXT FLOB_F.TXT FLOB_F.TXT FLOB_G.TXT FLOB_H.TXT FLOB_J.TXT FLOB_K.TXT FLOB_L.TXT FLOB_M.TXT FLOB_N.TXT FLOB_P.TXT FLOB_R.TXT	27       white room with dark green curtains. She was star         28       he room with PO6 79 green curtains with Rod, the         29       ing a sandwich at the Green Dragon in G10 62 Lor         30       aring a pink and lime green dress, made L12 70 (100)         31       be it red (peppers), green EO2 107 (apples) or (100)         32       3 <p_>She had on pale green eye shadow which went 100         33       he guessed, startling green eyes, a strong P21 100         34       ro with dark N17 220 green eyes and pale red hai:         35       te_&gt;"I have."<quote></quote> Green eyes glinted in the mode 100         36       -brown face and large green eyes. M06 134 She shows 100         37       <p_><quote_>"You have green eyes? seemed to ask and 100         38       21 197 but her lovely green eyes? She P23 170 lor         39       rm light in a pair of green eyes? She P23 170 lor         39       mode face Regex         Concordance Hits       Search Words</quote_></p_></p_>	FLOB_P.TXT FLOB_P.TXT FLOB_G.TXT FLOB_L.TXT FLOB_L.TXT FLOB_L.TXT FLOB_P.TXT FLOB_P.TXT FLOB_P.TXT FLOB_P.TXT FLOB_P.TXT FLOB_P.TXT FLOB_P.TXT FLOB_P.TXT FLOB_P.TXT
Total No. 15 Files Processed Reset	Start     Stop     Sort       Kwic Sort       ✓     Level 1       IR     ✓       Level 2     2R       ✓     Level 3	Save Window

こうすることで, greenの右側にどういう単語が来る傾向にあるかが目で見て確認できます. 右側の上下方向移動タブを動かして全体をざっと一覧すると, eyesやpolitics, Parkなどが目立ちますね.

こうした画面をじっくり見ることで,greenの使い方を新たな視点から見ることが出来るでしょう.

なお,コンコーダンスラインだけで前後関係が充分に読み切れない場合は,検索対象語の上にカーソルを移動させましょう。

上記の38例目, green eyes seemed to ask…のgreenの上にカーソルを移動させてみます。 すると,カーソルが「指さしマーク」に変わりますので,その状態で左クリックすると,本文閲覧モードに切り替わり, コンコーダンスラインを元の文脈の中で確認することができます。 🐼 AntConc 3.2.1 w (Windows) 2007

#### File Global Settings Tool Preferences About

Corpus Files	Concordance	Concordance Plot	File View	Clusters	Collocates	Word List	Keyword List
FLOB_A.TXT							[
FLOB_B.TXT	Hits	20 File: FLUB_P	.1XI				
FLOB_C.TXT	agency -						<u>^</u>
FLOB_D.TXT	P21 192 y	ou know. "Was	sh away	washday	blues."	That so	rt of
FLOB_E.TXT	P21 193 t	hing." <quote <="" td=""><td>'&gt;</td><th></th><th></th><td></td><td></td></quote>	'>				
FLOB_F.TXT	P21 194 <	o > <quote>″I</quote>	'm not	surpris	ed." <quot< th=""><td>te/&gt; sai</td><td>d Tom.</td></quot<>	te/> sai	d Tom.
FLOB_G.TXT	P21 195 <	 > <quote>"M</quote>	Ibat abo	ut vou?	" <quote :<="" th=""><td>&gt; she sa</td><td>id <quote>"T</quote></td></quote>	> she sa	id <quote>"T</quote>
FLOB_H.TXT	mean	p	mae abo	ao you.	.quo 00, ,	Dire Da	ia. squbbo_r i
FLOB_J.TAT	D21 196 W	hat about the	wotor	company	2" <quote< th=""><td>/&gt; Har y</td><td>oice seked one</td></quote<>	/> Har y	oice seked one
FLOB L TAT	guestion	liac about chie	; Water	company	· squoce,	V HEL V	Oice asked one
FLOB M TYT	P21 107 h					k +	hav suls
FLOB N. TXT	P21 197 D	ut Her Tovery	green	eyes se	emed to a	ask anot	ner.
FLOB P. TXT	P21 198 <	p_>Tom croake	ed out a	n answe	r. <quoto< th=""><td>≞_&gt;"Oh,</td><td>there's nothing</td></quoto<>	≞_>"Oh,	there's nothing
FLOB R.TXT	to taik					2160174 <b>2</b> 01603	
	P21 199 a	bout. I do th	ie accou	nts. At	least, d	design t	he computer
	systems						
	P21 200 t	hat do the ac	counts.	There	are mill:	ions of	them. The
	computer	does					a ay na man
	P21 201 t	hem. It works	s them o	ut and	sends the	∋m.″ <quo< td=""><td>te/&gt;</td></quo<>	te/>
	P21 202 <	p_> <quote_>"W</quote_>	Jell, I	won't s	ay it sou	und s	
	fascinati	ng," <quote></quote>	she				× .
			-				
	Search Term	Vvords   Case	Regex				
	green		🗧 Advar	nced			
Total No. 15	Start Stop						
Files Processed	Hit Location						Save Window
Read							Exit
Reset	12 🗉						1 -2.

探していた文章が、「FLOB\_P.TXT」というファイルの中にあることがわかり、同時に前後をたっぷり読むことができます。 なお、自動的に上部のタブが「File View」に切り替わっていますので、もとの画面に戻したいときは「Concordance」タブをクリックすればOKです。

4. Collocatesで共起語を一気に調べる

上記では、「Concordance」機能を使ってgreenの右側に来る語を調べましたが、 このように詳しく用例を見るよりも、シンプルに、greenの右側に来る単語の頻度リストだけがほしい場合もあるでしょう。

こういう場合に便利なのが「Collocates」,つまり「共起語リスト作成機能」です。

まず, 画面上部のタブのうち, 右から3つ目の「Collocates」のタブを押します。

次に, 画面下部右側のWindow Span設定パネルを見てみましょう。

Window Span 🗆 same

From [1L] to [1R]

ここで言うWindow Spanとは, 共起語データを取りたい範囲のことです。 たとえば上記のデフォルトの設定だと,

(1))検索対象語(NODE)の左隣り位置

(2)検索対象語自身の位置

(3) 検索対象語の右隣り位置

上記の3つの位置に現れる単語の頻度を計算して表示してくれます。

今回はgreenの右隣りに来る語だけを知りたいので, Window Spanを狭めて下記のように設定します。

Window Span 🗆 same

From [1R] to [1R]

以上の設定が終わると水色の「Start」ボタンを押しましょう。

🖶 AntConc 3.2.1w (W	indows) 2(	007								
File Global Settings Tool P	references A	\bout								
Corpus Files FLOB_A.TXT FLOB_B.TXT FLOB_C.TXT FLOB_D.TXT FLOB_E.TXT	Concorda Total No. o Rank	nce Co f Collocate Freq 8	ncordance Types: 110 Freq(L) 0	Plot F Total No. Freq(R) 8	File View Cli of Collocate To Collocate eyes	usters kens: 14	Collocates	Word List	Keyword List	
FLOB_F.TXT FLOB_G.TXT FLOB_H.TXT FLOB_J.TXT FLOB_K.TXT FLOB_L.TXT FLOB_M.TXT FLOB_N.TXT FLOB_P.TXT FLOB_R.TXT	2 3 4 5 6 7 8 9 10 11 12 13	4 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2		4 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	polition and Party movemen with was said quote Park p or man	nt				
Total No. 15 Files Processed Reset	14 15 Search Ter green Start	2 2 m 🔽 Wo	0 0 rds T Ca Sort	2 2 Se Re Sort by Sort by	L J egex Advanced t by Freq rt Order	]	Window Spar From  1R Min. Collocat  1 = €	n Г Same ⊕ To… eFrequency	1R •	Save Window Exit

すると,見事,greenの右隣りに生起する語の頻度順リストが出力されました。

ふつう, greenと言うと, 「緑色」をイメージしますが, 実際の英語ではそうした直接的な意味合いよりも, 比ゆ的に「嫉妬」を表したり(※green eyesはしばしば「嫉妬の眼」の意), 「自然環境」を表したりする(※green politics/ Green Party/green movement)用例が多いことがわかります。

こうした単語のコアイメージはなかなか非母語話者には(母語話者にとってすら!) つかみにくいものですが, コーパスを駆使することで語の隠れた真の意味をあぶりだすことができるわけです。

せっかくの貴重なデータですから,これは保存しておきましょう。

File <Save Output to Text file (結果をテキストファイルとして保存する)とすると、 上記の結果がテキストファイルで保存されますので、 必要に応じてExcelなどで加工することも可能です。

5. 特定語を含むクラスターを探そう

実際の英語では,語は個々が独立して存在するのではなく,しばしば,複数の語があつまった小さなかたまり(cluster)の単位で存在していると言われます。

たとえば、「晴れた日」のことを英語では「a fine day」と言いますが、 母語話者は、これを言うために、そのつど、頭の中で <a + fine + day> という足し算的処理を行っているわけではなく、 頭の中に最初から「a fine day」というかたまりが入っているわけです。

語が, このように, よく使われる他の語と一体化してクラスターを構成しているとすれば, 単語集的な1語1語の単位ではなく, こうしたクラスターの単位で語を学習すれば, 英作やスピーキングにも役立つことでしょう。

特定語を含むクラスターを探すには、「Clusters」検索を行います。

まず,画面上部タブの真ん中にある「Clusters」を押します。

次にSearch Termボックスに検索したい語(ここではgreen)を入力します。

ついで,画面下部右側の「Cluster Size」パネルを設定します。

Cluster Size

Min. Size [2] Max Size [2]

デフォルトでは2語以上2語以下, すなわちgreenを含むぴったり2語のクラスターだけを機械的に検索してくれます。

ここで水色の「Start」ボタンを押せば下記のような結果が得られます。

AntConc 3.2.1w (V	/indows) 2	2007							
File Global Settings 1001	Preferences	About							
Corpus Files	Concord	lance	Concordance Plot	File View	Clusters	Collocates	Word List	Keyword List	
FLOB_A.TXT	LOB_A.TXT				J				
FLOB_B.TXT	Total No.		riypes: 214 Total No	. or Liuster 1 o	kens: 273				
FLOB_C.TAT	Rank	Freq							
FLOB E. TXT	1	11	the green						
FLOB F.TXT	2	8	a green						
FLOB G.TXT	3	7	green eye	3					
FLOB_H.TXT	4	6	the Green	10)					
FLOB_J.TXT	5	5	and green						
FLOB_K.TXT	6	4	Green pol	itics					
FLOB_L.TXT	7	3	green and						
FLOB_M.TXT	8	3	Green Par	tv					
FLOB P. TXT	9	3	of green	of green					
FLOB R.TXT	10	3	pale gree	n					
	11	3	Tim Green						
	12	2	dark gree	n					
	12	2	arest are						
	14	2	great gre	taina					
	15	5	green cur	lains					
		4	> K	Ids					× a
	1	1)	- Iters						<b>Y</b>
	Search Te	erm 🔽	Words 🥅 Case 🥅	Regex 🥅 N	-Grams	Cluster Size			
	green			🔶 Advar	nced	Min. Size 2	😫 Max. Siz	e 2 🜩	
Total No. 15	Start	Stop	Sort Sort	by		Min. Cluster	Frequency		
Files Processed	Search Te	erm Po	sition	Sort by Freq		1 🕀			Save Window
Reset	☐ On Le	ft [	On Right	vert Order					E <u>×</u> it
			-					-	

ネイティブの頭の中にgreenがどのようなクラスターとして入っているか,かなり見えてきましたね。

なお,「Cluster size」パネルを設定すれば、2語クラスターだけでなく、3語クラスター、4語クラスター、5語クラスターなども検索可能です。 いろいろ設定を変えて試してみるといいでしょう。

「Search Term Position」は、「~ green」なのか、「green ~」なのかというように、 2語クラスターにおけるgreenの位置を限定するときに使います。 (デフォルトでは指定なしなので上記にはgreenが右に来ているものも左に来ているものも混ざって出ています。)

「Min. Cluster Frequency」は、出力結果に表示するクラスターの最低頻度を指定します。 デフォルトは1, すなわち1回でも出ていればすべて表示されますが、ヒット数が多いときは、これを「2」や「3」にして、 頻度の高いクラスターに絞って分析するのもいいでしょう。

以上のように, n語からなるクラスターのことを, 言語学では, 総称的に「n-gram(エヌグラム)」と呼ぶことがあります。 とくにnが2のときは「bigram(バイグラム)」, nが3のときは「trigram(トライグラム)」などと呼ばれます。

### 6. コーパスから語彙表を作ろう

コーパスがあれば、そこに含まれるすべての語の頻度を計測し、語彙表を作ることが可能です。

(File < Open Dirでデータは既に読み込んでいるとします。)

(1)まず, Antconcの上部のタブのうち, 右から2つ目の「Wordlist」を押します。

(2)次に,画面下部のDisplay optionsパネルにある「Treat all data as lowercase(すべて小文字として処理する)」にチェックを入れます。 (このチェックを入れないと,小文字のthisと文頭のThisは別の語としてカウントされます。チェックをいれて1語として数えるほうがふつうでしょう。)

(3) そして,水色の「Star」ボタンを押します。

📾 AntConc 3.2.1w (W	indows) 20	07							
File Global Settings Tool P	Preferences A	.bout							
Corpus Files	Concorda	nce Co	ncordance Plot	File View	Clusters	Collocates	Word List	Keyword List	
FLOB_A.TXT	Hite	0	Total No. of W	/ord Tupes: A	3134 Total N	o. of ₩ord Tok	ens: 1199781	2000 - 20	
FLOB_B.TXT	Renk	Fred	Mord	roid Types. 4		d Form(s)	.018. 1100101		~
FLOB D. TXT		печ	vvoru			u i onni(s)			=
FLOB E.TXT	1	64816	the						
FLOB F.TXT	2	48657	р						
FLOB G.TXT	3	34147	of						
FLOB_H.TXT	4	33811	a						
FLOB_J.TXT	5	27293	and						
FLOB_K.TXT	6	27058	to						
FLOB_L.TXT	7	20894	in						
FLOB_M.TXT		17405	4						
FLOB_N.TXT		16592	J						
FLOB P. TAT	10	1/00/2	quoce						
T DOD_K. IXI		10200	9						
		10798	that						
	12	10295	15						
	13	10039	was						
	14	9885	h						
	15	9665	it		-				
	< >	< >	5	>	<				
	Search Terr	m 🔽 Wo	rds 🥅 Case 🥅 I	Regex		Display Optio	ns		
	1				ced	✓ Treat all da	ata as lowerca:	se	
Total No. 15	Start	Stop	Sort o	<u> </u>				228 2	
	Start	Stop	Sort	y .					Save Window
Files Processed	Hit Location	n	S	ort by Freq					
Reset	Search Only	y 0 📫		/ert Order					E⊻it

するとtheを筆頭に, 高頻度語がリストされました。

画面の上に注目すると,

Total No of Word Types: 43134 Total No of Word Tokens 1199781

という情報が出ています。

これはコーパスがtokens(延べ語数)で言うと1199781語あり, 同一語の重複をのぞいたtypes(異なり語数)は43134語あることを示します。

(※FLOBは100万語と言われていますが,実際には,タグ等が含まれており[上記のpやquoteもそうです],100万語以上にカウントされます)

このようにしてできた語彙表は教育的にいろいろ活用できるものですが, このままでは使いにくいですね。

そこでFile <Save output to Text file (結果をテキストファイルで保存)を押し, デスクトップなどにテキストファイルで保存します。

保存されたテキストファイルをダブルクリックするとnotepadなどが開いてしまいますので,

先にMS Excelを開けておいて, Excelから「ファイルを開く」を設定します。

左側の「ファイルの場所」をデスクトップに設定しますが, デフォルトではせっかく保存したテキストファイルが見つかりません。

そこで「ファイルの種類」の右端の矢印ボタンを押して、「すべてのファイル」と設定変更すると、 保存したテキストファイルが見つかりますので、それを選んで「開く」を押します。

ファイルを聞く						? 🛛
ファイルの場所型:	🞯 デスクトップ			*	() • 🖄	X 📸 🖬 •
していたつ しょう	<ul> <li>□ マイ ドキュメント</li> <li>□ マイ コンピュータ</li> <li>□ マイ オットワーク</li> <li>□ j8vtLOT</li> </ul>					
デスクトップ	Constructs Sourt Aのシ Sourt Aのシ Sourt Anta El FLOB wordlist t	vョートカット xt				
مر المجمع الم	ja In Progress ∰ j8vtLOT.Izh					
२१ २७२७-७						
	 ファイル名(N): 「				~	
	ファイルの種類(工):	すべてのファイル (*.*)			~	
<u> </u>					開((()))	キャンセル

すると次のような画面が出てきます。

テキスト ファイル ウィザード - 1 / 3	? 🛛
選択したデータは区切り文字で区切られています。 D太へ]をクリックするか、区切るデータの形式を指定してください。 元のデータの形式 データのファイル形式を選択してください: ③力ンマやタブなどの区切り文字によってフィールドごとに区切られたデータ(D) ○スペースによって右または左に揃えられた固定長フィールドのデータ(W)	
取り込み開始行(E): 1 このファイル(Q): 932:日本語・ ファイル C:¥Documents and Settings¥Administrator¥デスクトップ¥FLOB_wordlist	ν7ト JIS) 🛛 💽
1 1 64816 the 2 2 48657 p 3 3 34147 of 4 4 33811 a	
キャンセル	₹る(B) <u>次へ(N) 完了(F)</u>

これはテキストファイルをExcelの標準的な形式に変換するためのウィザードです。 「次へ」「次へ」・・・と押してゆくと(※設定変更の必要はありません), うまくExcelのフォーマットになってExcelに読み込めます。

A	1 2	13D =	FLOB_wor	dlist.txt - N	licrosoft Exce	1	-	= X
	市금소	插入 ページ	יוליק אלי	教式 デー	々 校開 孝	武 I2년 N	統計 🔞 -	. 🗖 X
「 ビボ テープ	ア ット ット ・テーブル		- <sup>1</sup> 757 1	یں۔ روریانہ ا	<ul> <li>テキスト ボ</li> <li>ヘッダーとフ</li> <li>ヘッダーとフ</li> </ul>	ックス + 🗟 + ックス + 🏂 + ッター 🌌 - + Ω	,	
	テーブル			リンク	テキン	スト		
	A2	•	( fx	1				*
	A	В	С	D	E	F	G	1
1	順位	頻度	語					
2	1	64816	the					
3	2	48657	р					
4	3	34147	of					
5	4	33811	a					
6	5	27293	and					
7	6	27058	to					
8	7	20894	in					
9	8	17405	J					
10	9	16592	quote					
11	10	16080	g					
10	FLOE	10799 Wordlist	that					
עדב	<u>۲</u>					100% 😑	U	:

上記は、データを読み込み、1行目に見出しを追加した状態です。 ここまでくれば、自由に加工して使うことができますね。

現場の教師であれば,教科書コーパスから語彙表を作ったり,入試コーパスから大学別の語彙表を作ったりと, 研究目的に限らず,こうした語彙表作成の手法は応用の可能性が広いものです。

### 7. コーパスから特徴語を抽出しよう

二つのコーパスの構成語彙を比較すれば,一方に特徴的に頻出する語彙を特定することができます。

ここでは、100万語のイギリス英語コーパスであるFLOBと、同様の基準で作成されたアメリカ英語コーパスであるFROWNを比較し、 イギリス英語の特徴的語彙を探ってみましょう。

ところで、「特徴的に頻出する」というのは、簡単なようで、意外に難しい概念です。

たとえば単語Xが, Aコーバスに10回, Bコーバスに20回出ていれば, 単語XはBコーバスの特徴語と言っていいでしょうか?

このように二つの値の間に意味のある差があるかどうかを考えるためには, 統計的な検定を行う必要があります。

ふつう、2値の有意差を見る場合は、「カイニ乗検定」を行いますが、 コーパスの世界では、分母が異なるデータ(10万語コーパスと100万語コーパスなど)を比較する場合も多いので、 「カイニ乗検定」よりも安定性が高い「対数尤度比(loglikelihood ratio)」がしばしば使用されます。

こうした計算は煩瑣なものですが、Antconcでは自動的に計算を行い、 統計的に意味のある特徴語だけをリストとして抽出してくれます。

以下, 手順を確認してゆきましょう。

- (1) 先にFLOBから語彙表を作成しておく。
- (2)画面上部のTool Preference <Keyword Listを選ぶ。
- (3) Treat all data as lower case (すべて小文字とみなす) にチェック。
- (4) Show negative keywords(低頻度特徴語も表示する)にチェック。
- (5) Add Directoryを押し、比較対象のコーパスの入っているフォルダを選ぶ。

🚾 Tool Preference	25 0 28 19 19	
Category Concordance Clusters Collocates Word List Keyword List	Keyword List Preferences         Display Options         Rank       Frequency         Keyness       Keyword         Other Options         Treat all data as lowercase         Treat case in sort         Keyword Options         Keyword Generation Method         Log-Likelihood         Threshold Value         Average Value         Show negative keywords (using highlight color)         Reference Cornus Options	
	Add Directory Choose Files Swap Ref/Main Files          FROWN_J.TXT         FROWN_K.TXT         FROWN_L.TXT         FROWN_M.TXT         FROWN_N.TXT         FROWN_N.TXT         FROWN_P.TXT         FROWN_R.TXT         FROWN_R.TXT	Clear

(6)「Apply」を押す。
(7)メイン画面に戻って上部タブの右端「Key Word List」を押す。
(8)下部のReference Corpus パネルに「Loaded(対照コーパス設定済み)」にチェックが自動で入っていることを確認。
(9)水色の「Start」を押す。

🐼 AntConc 3.2.1w (Windows) 2007

#### File Global Settings Tool Preferences About

orpus Files	- Concor	dance Co	ncordance Plot	File View Clusters	Collocates	Word List K	eyword List	
LOB_A.TXT LOB_B.TXT	Hits	Hits 0 Keyword Types Before Cut: 43134 Ke				eyword Types After Cut: 8515		
LOB_C.TXT	Rank	Freq	Keyness	Keyword				1
OB_D.TXT	1	734	775 847	nounds				
OB_E.TXT	2	500	695 966	flob				
OB_F.TXT	3	466	524 926	cent				
Эв_G.ТАТ Эв_н түүт	4	374	498 409	labour				
OB J.TXT	-	4052	405 020	rabour				
DB K.TXT		4032	403.920	WILCH				
DB_L.TXT	2	223	310.401	Luk				
DB_M.TXT		293	299.543	towards				
OB_N.TXT	8	467	233.997	London				
DB_P.TXT	9	575	227.070	per				
OB_R.TXT	10	190	219.101	centre				
	11	10039	205.295	was				
	12	144	189.872	programme				
	13	507	188.731	british				
	14	319	176.051	britain				
	15	133	174.719	behaviour				
	5. 3	< >	<	5			3	2
	- <del>10</del> 					1000		
	Search 1	erm 🖌 Wo	ords   Case   R	egex	Display Optio	ns		
				Advanced	Treat all da	ata as lowercase		
al No. 15	Start	Stop	Sort Sort by	,	Reference Co	orpus		
s Processed	Hit Locat	tion	Sort	by Keyness 💷	🔽 Loaded	Reset	Save Wir	ndov
et IIII	Search			ert Order			Exit	_
xer IIIIIII	Search							_

以上はイギリス英語の特徴的高頻度語のリストです。 下のほうまで見ていくと、色が変わるところがありますが、それらはnegative keywords、すなわち特徴的低頻度語を示しています。

イギリスの通貨単位poundや、イギリス風のスペリングのlabour, centre, programme, behaviourなどが、うまくイギリス英語の特徴語として抽出できていることを確認しましょう。

また, centとperの二つが上位に入っていますが, これは「パーセント」と書くときに, 米語はsolid compoundでpercentと書き, イギリス英語ではopen compoundでper centと分かち書きすることを反映したものです。

なお,775.847や,695.966などの数字が,それぞれの語がもつ対数尤度比指数であり, Antconcではそれを「Keyness」(特徴度指数)と呼んでいます。

このように二つのコーパスを比較して特徴語を抽出するという処理は意外に汎用性が高いものです。

最近では大学の英語教育でもESP (English for Specific Purposes)の必要性が叫ばれていますが、 たとえば、基礎的な英語語彙力をもった看護学生を対象に、看護英語の語彙表を作ろうとする場合は、 看護英語コーパスと、一般的な英語コーパスの2種類を用意し、 看護英語コーパスにおいて特徴的に頻出する語彙を選び出せば良い語彙表ができそうです。

また、日本人の英作文とネイティブの英作文の比較し、日本人が過剰使用(overuse)する傾向にある語彙をリストしてみるのも有益でしょう。

### 8. データの中での語の位置関係をみる

Antconcの「Concordance Plot」機能を使えば、データの中で特定の語が出現する位置を視覚的に把握することができます。

もっとも、FLOBのように、短いサンプルを寄せ集めたコーバスではこの機能はあまり意味を持ちませんが、 たとえば長編小説の中で特定の語やイメージの推移を見る場合にはこの機能は有益です。

ここでは、チャールズディケンズの長編作品『デイビッドカパーフィールド』において、 greenという語がどのような位置に出現するかを見てみましょう。

まず, File < Open Fileで, 『デイビッドカパーフィールド』のデータを読み込ませた後, 「Concordance Plot」のタブを押し, Search Termボックスにgreenを入力, その後, 水色の「Start」ボタンを押します。

### \_ 🗆 🔼

📾 AntConc 3.2.1w (Windows) 2007										
File Global Settings Tool Preferences About										
Corpus Files	Concordance         Concordance Plot         File View         Clusters         Collocates         Word List         Keyword List           HIT FILE: 1         FILE: oprid10.txt         No. of Hits = 33         No. of Hits = 37         File Length (in chars) = 1956919									
Total Ho. 1 Files Processed Reset	Search Term ▼ Words Case Regex       Concordance Hits         green       ▲ Advanced       39         Start       Stop         Plot Zoom       Exit									

上記の横長のバーは作品の全体を表します。 縦線が入っている部分がgreenの出現している箇所です。 縦線が重なったように見えるところはgreenが集中的に出現していることを示します。

上記を見ると、greenという語は、作品の冒頭部、ついで冒頭5分の1あたりに集中的に出ており、 その次は作品の結末部分に集中的に出ていることがわかります。 こうした直観的な把握は、文学作品のイメージの構成などを見る上で重要なヒントを与えてくれることでしょう。

### 9. おわりに

以上,実際の検索実例を披露しながら,Antconcの使い方を簡単に見てきました.

冒頭でも述べたようにAntconcは有償のコンコーダンス・ソフトの主要機能を殆ど網羅しており,しかも操作が簡単です. 初めてコンコーダンスを使う方にはとくに推薦できるソフトだと思います.

また、上記の検索実例をヒントにすれば、いろいろとおもしろい研究課題が広がることでしょう。

コーパスは、アイデアさえあれば、研究の可能性は無限に広がります。

みなさんも, ぜひ一度, おためしになってみてください。

Back to top

神戸大学石川慎一郎研究室 (C)2006-2008