English
Top

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

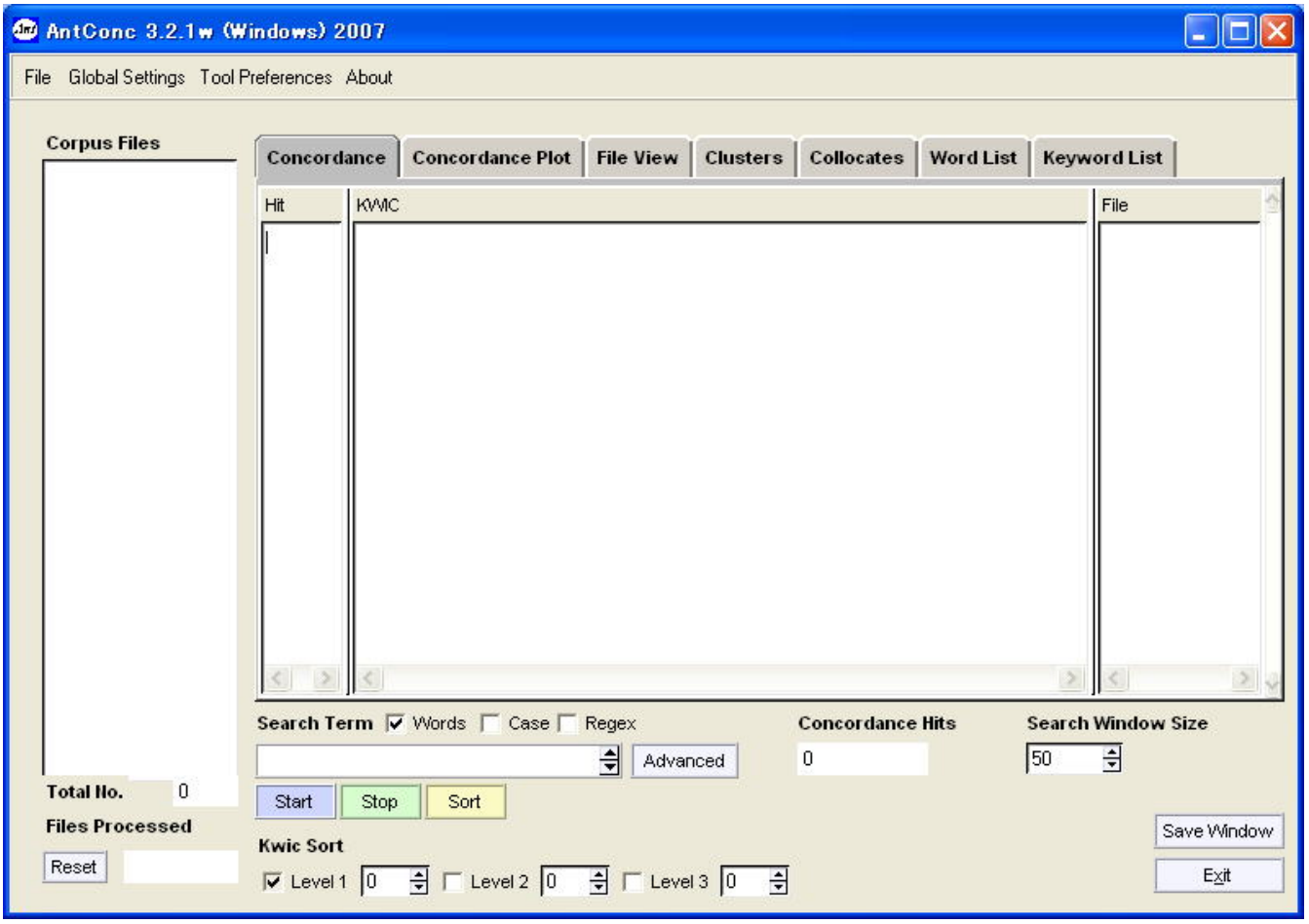
1．はじめに

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．



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。

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

Antconcにはさまざまな機能があり，それらは画面上部のタブを押して選ぶようになっています。
デフォルトでは左端の「Concordance」が選ばれています。

「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．

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。

```
==================================================
Kwic Sort
?Level 1 [ 0 ]　□Level 2 [ 0 ]　□Level 3 [ 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。

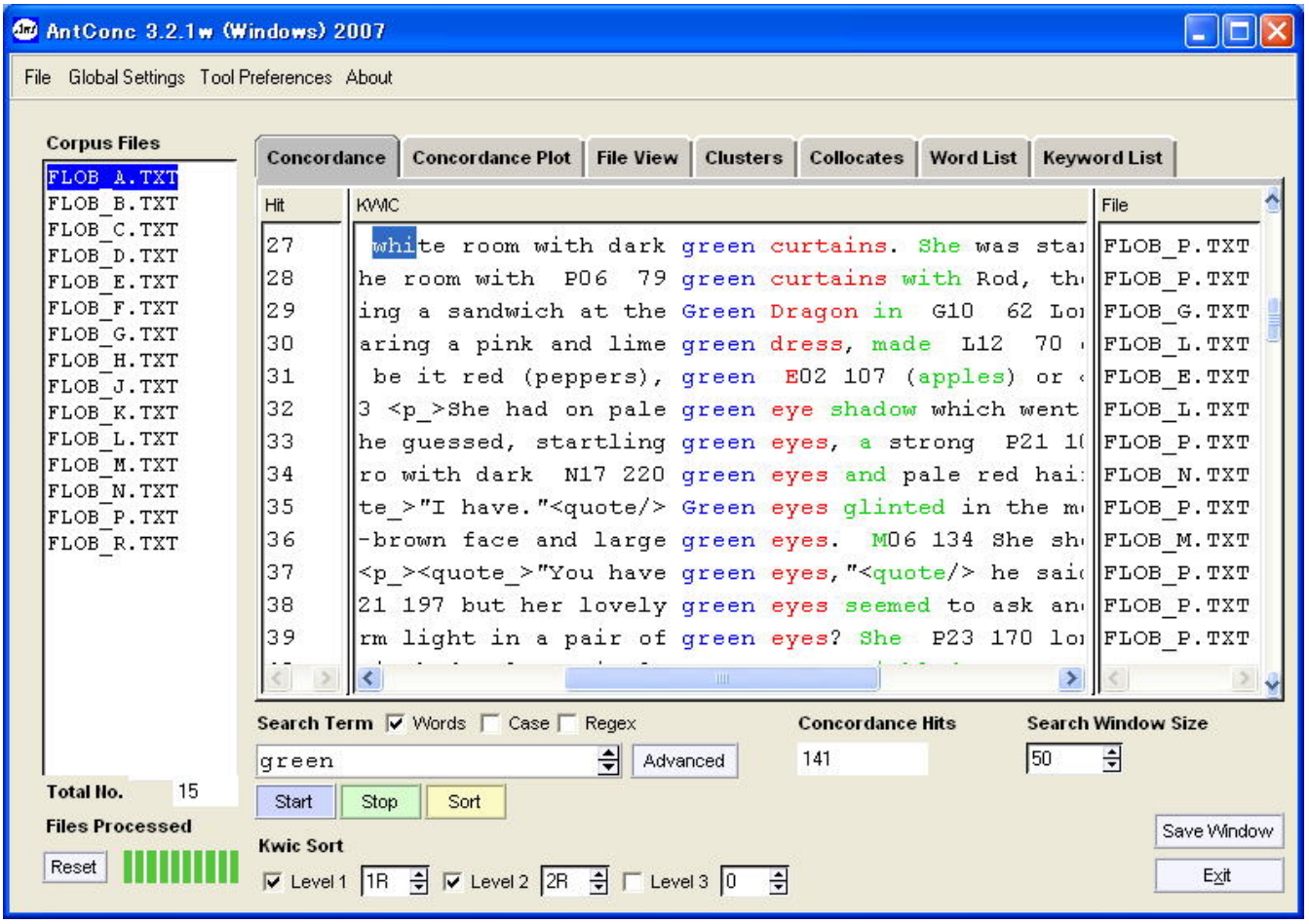
greenを中心に考えると，すぐ左隣は[1L]（左=Leftの1番目），右隣りは「1R」（右=Rightの1番目）となり，
以下，「2L」「3L」・・・，「2R」「3R」・・・のように変化していきます。

```
================================================================================
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.

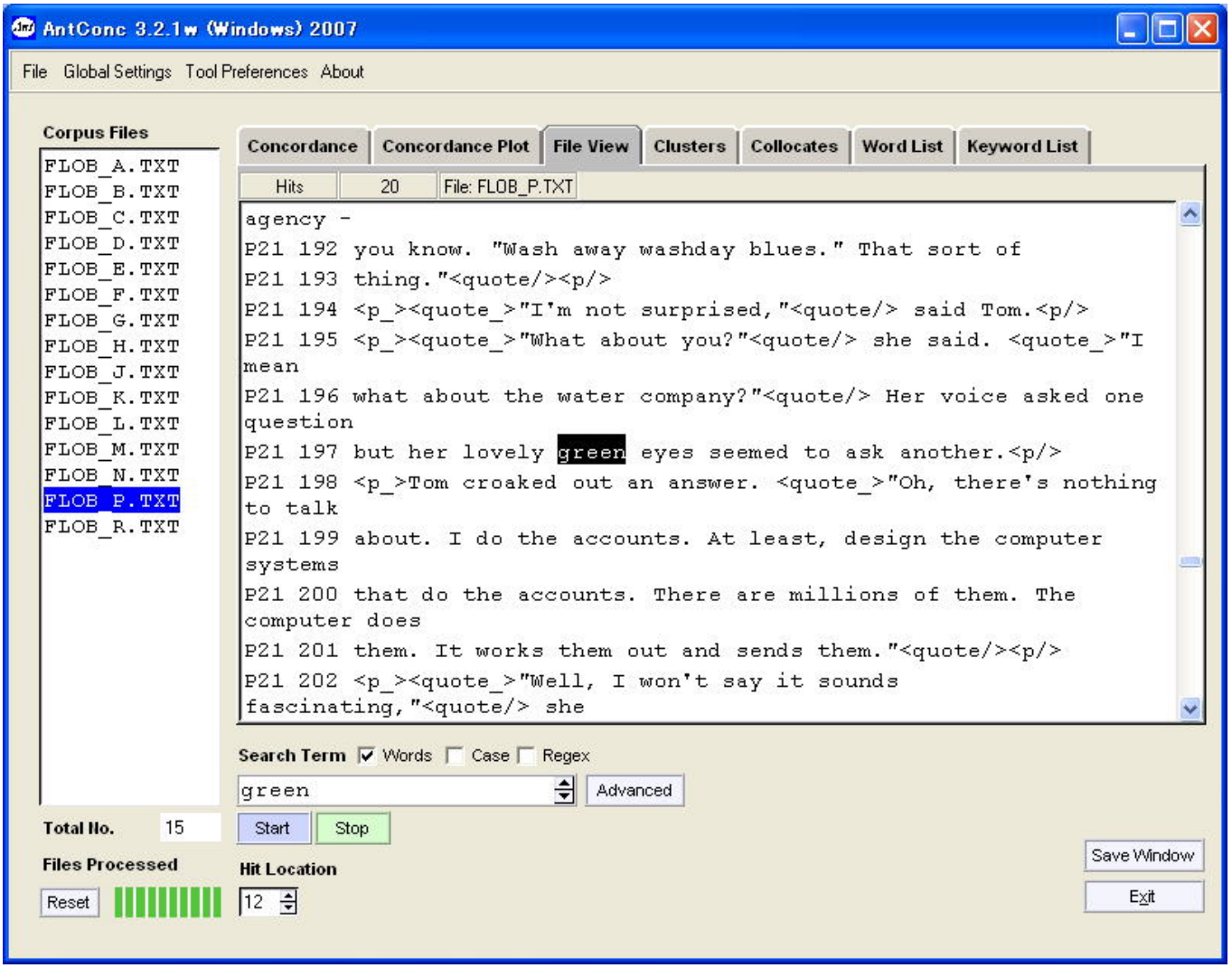


こうすることで，greenの右側にどういう単語が来る傾向にあるかが目で見て確認できます.
右側の上下方向移動タブを動かして全体をざっと一覧すると，
eyesやpolitics, Parkなどが目立ちますね.

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。



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。

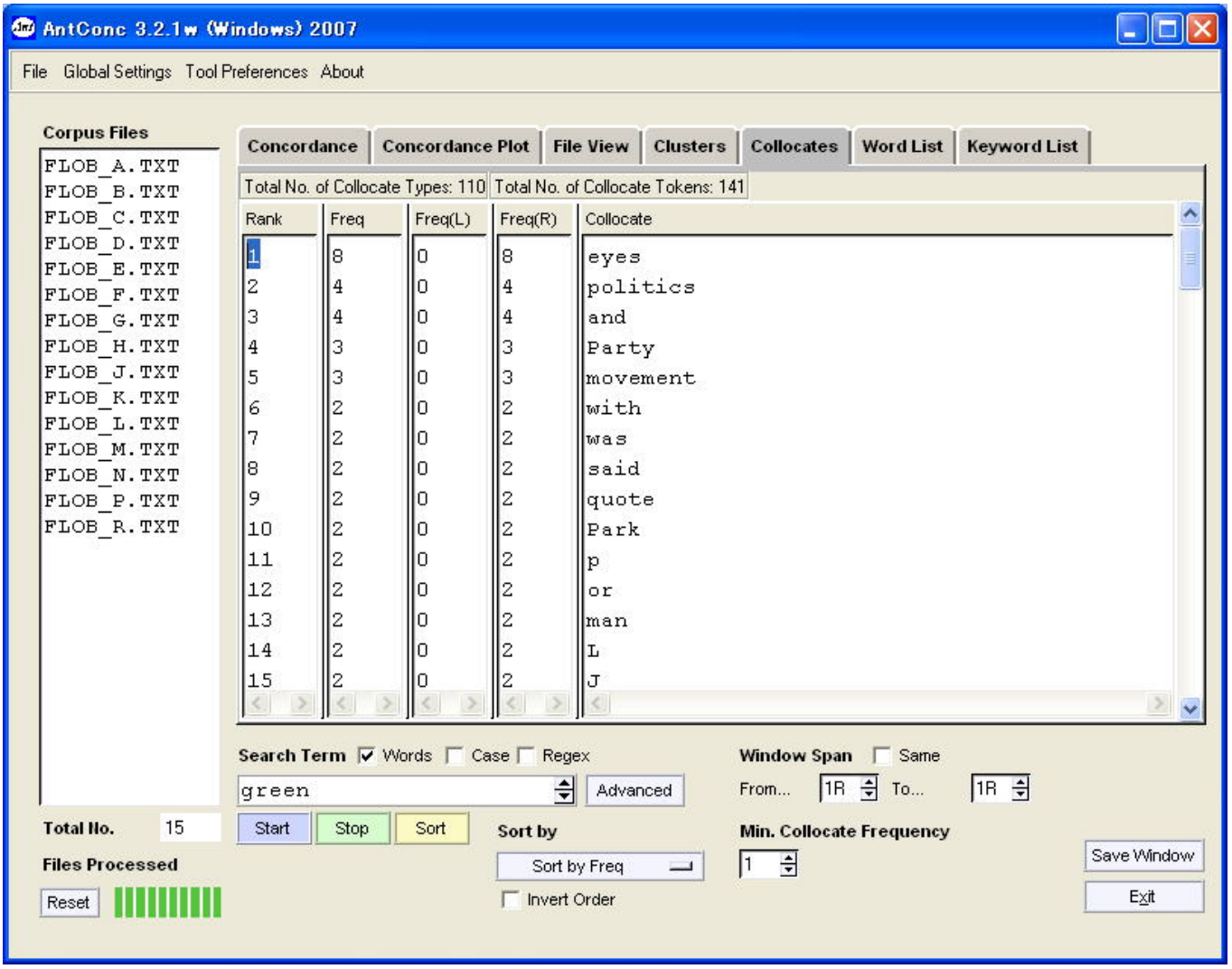
今回はgreenの右隣りに来る語だけを知りたいので，
Window Spanを狭めて下記のように設定します。

==========
Window Span　　□same

From [1R] to [1R]
==========

これで「1R」，すなわち検索対象語であるgreenの右隣りの位置だけ検索することになります。

以上の設定が終わると水色の「Start」ボタンを押しましょう。



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。

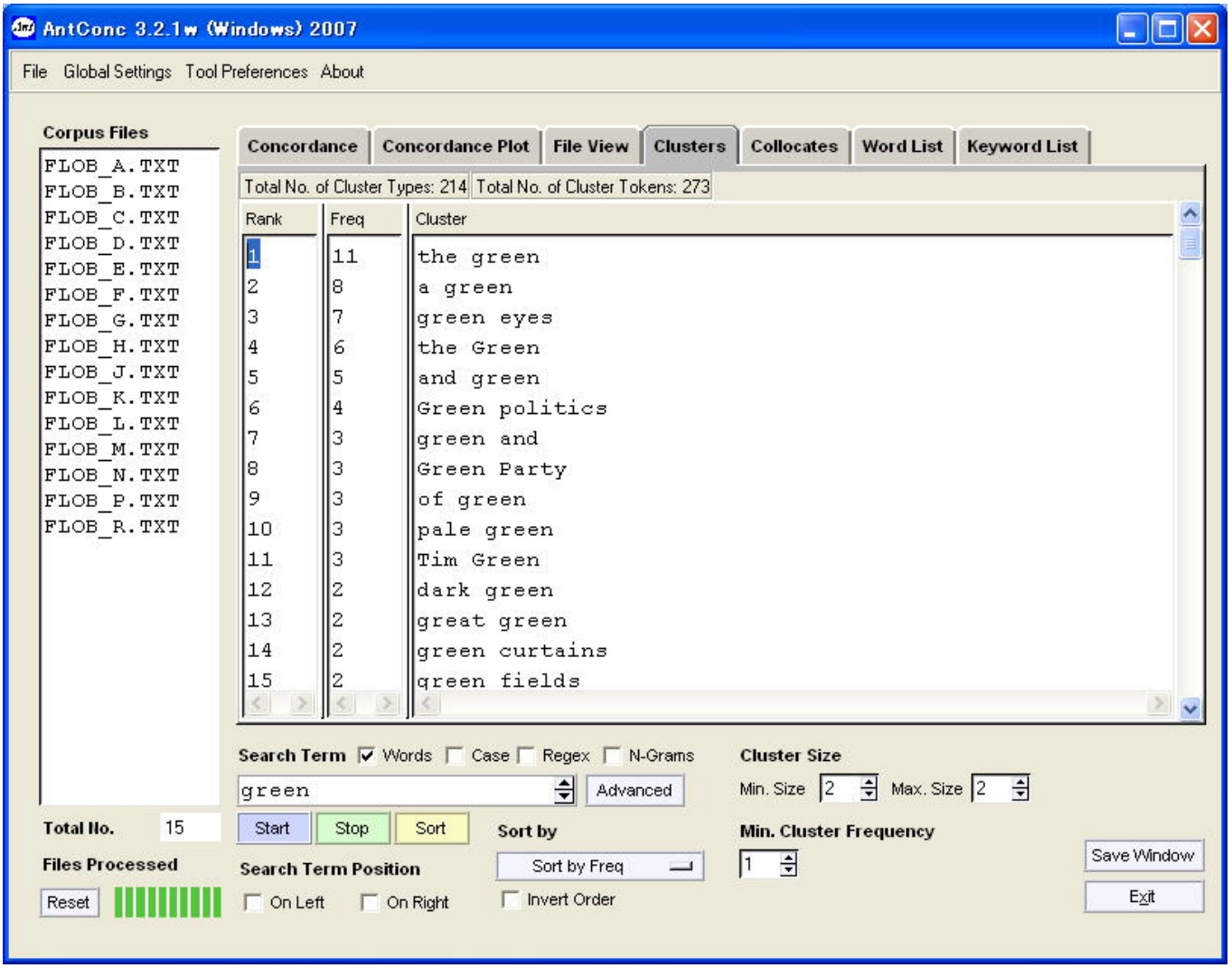
ついで，画面下部右側の「Cluster Size」パネルを設定します。

==========
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。



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

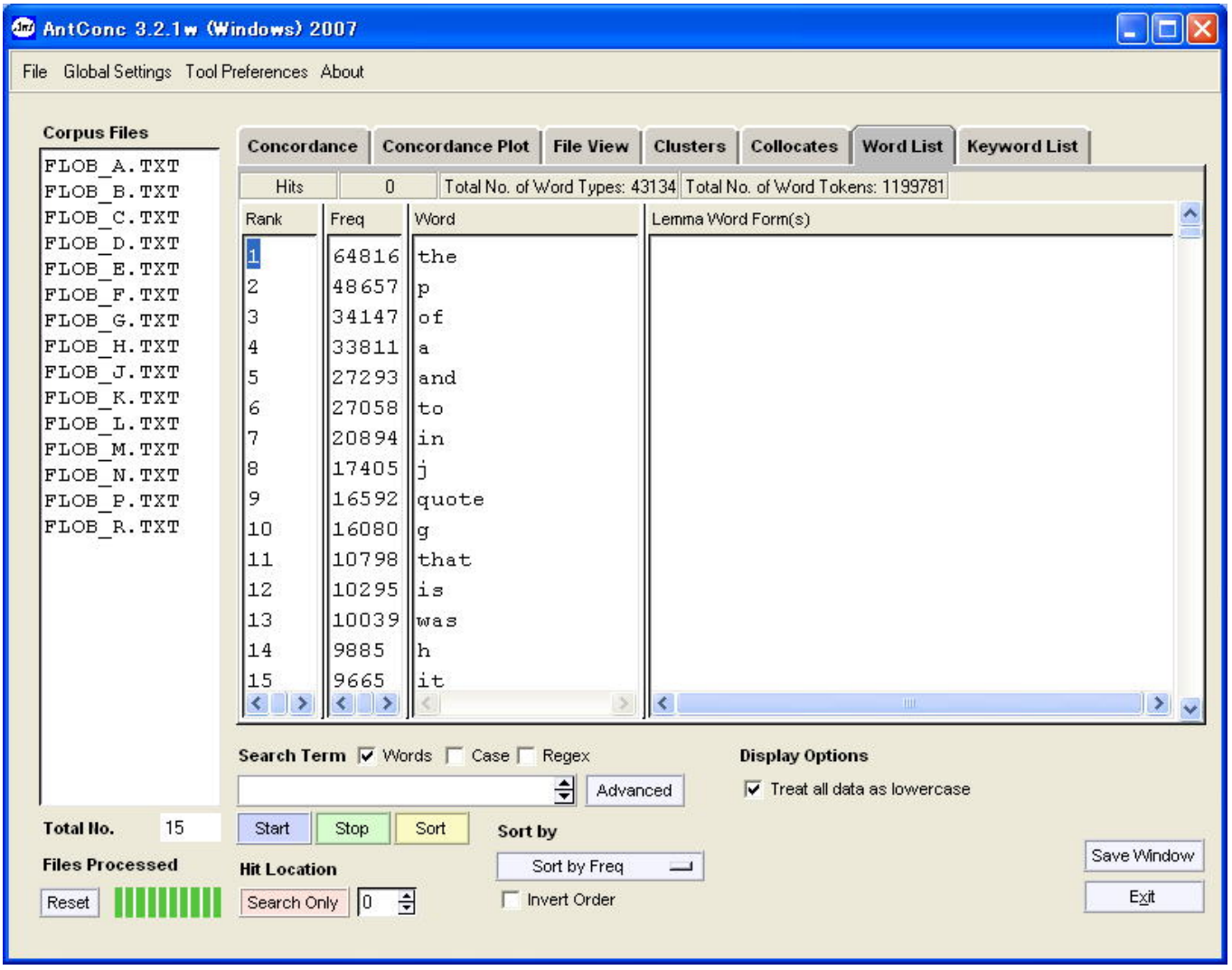
コーパスがあれば，そこに含まれるすべての語の頻度を計測し，語彙表を作ることが可能です。

（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。



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。

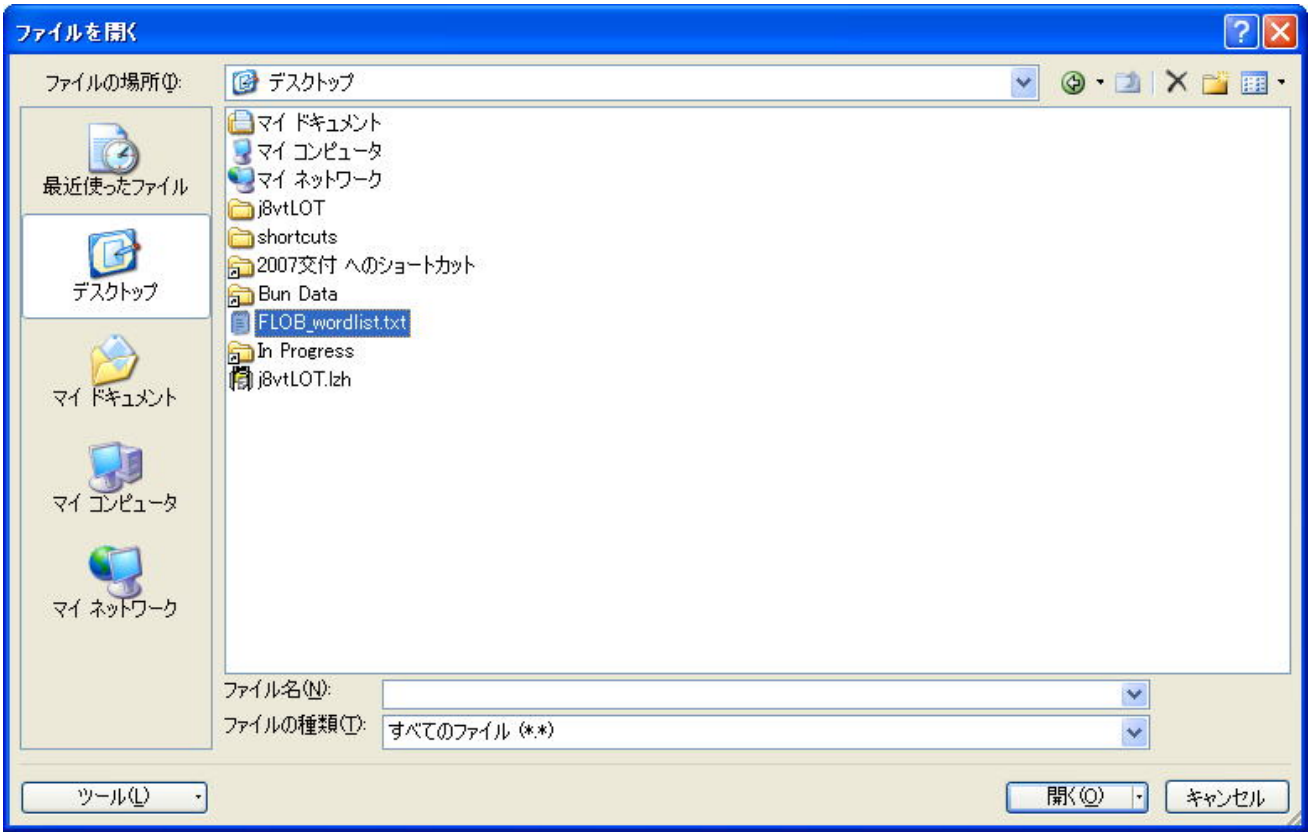
そこでFile＜Save output to Text file　（結果をテキストファイルで保存）を押し，
デスクトップなどにテキストファイルで保存します。

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，

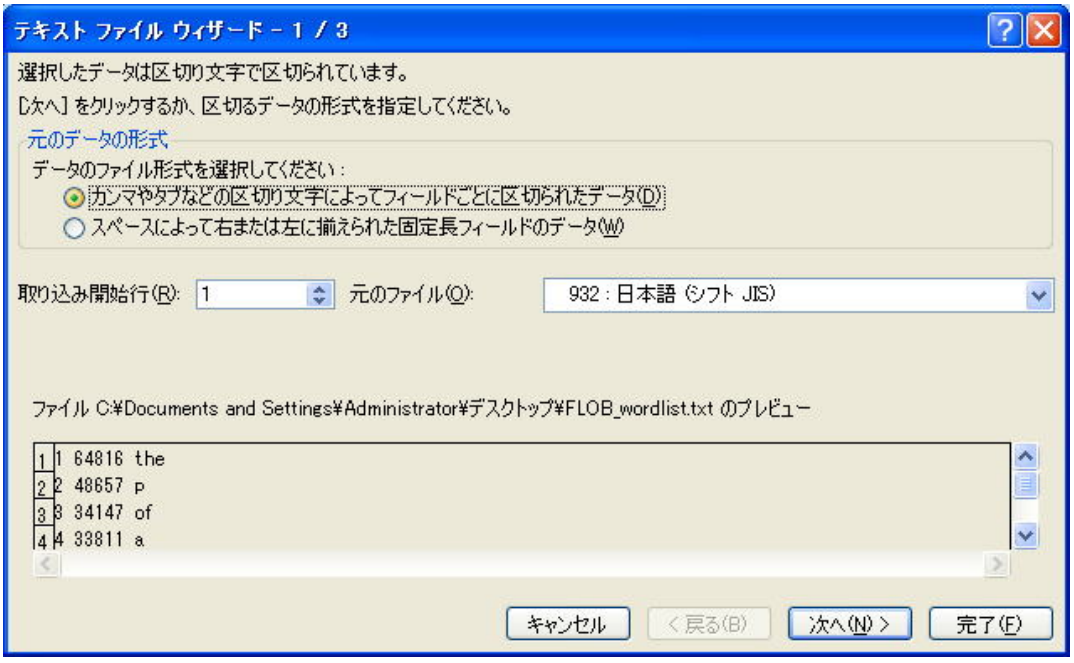
先にMS Excelを開けておいて，
Excelから「ファイルを開く」を設定します。

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。

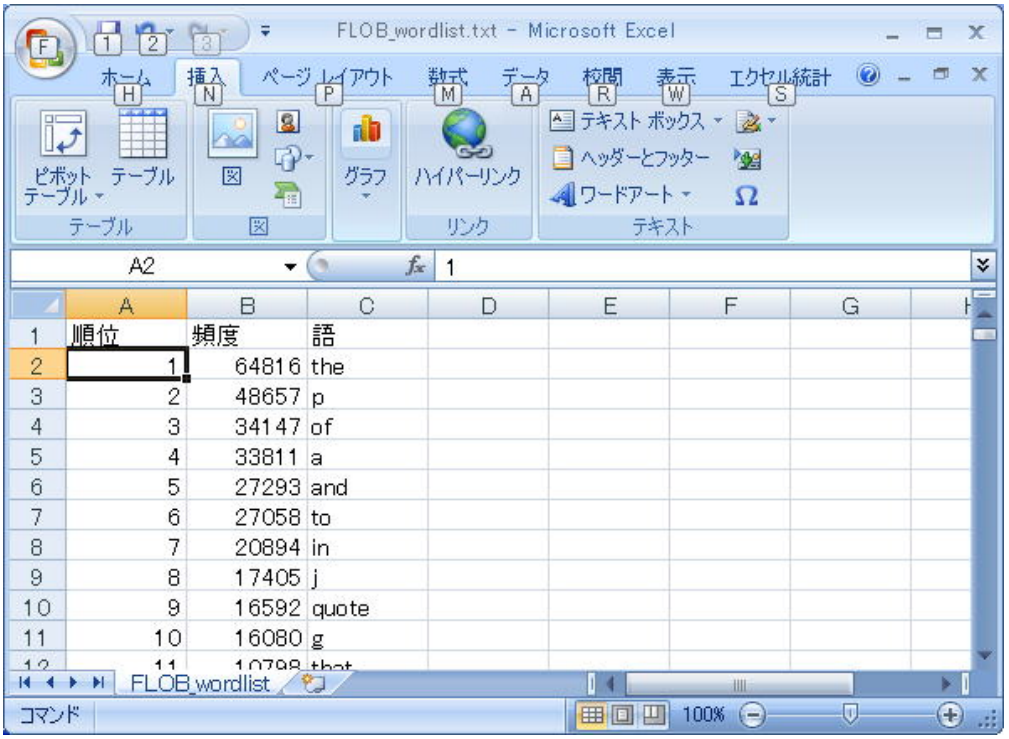
そこで「ファイルの種類」の右端の矢印ボタンを押して，「すべてのファイル」と設定変更すると，
保存したテキストファイルが見つかりますので，それを選んで「開く」を押します。



すると次のような画面が出てきます。



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。



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コーパスに10回，Bコーパスに20回出ていれば，
単語XはBコーパスの特徴語と言っていいでしょうか？

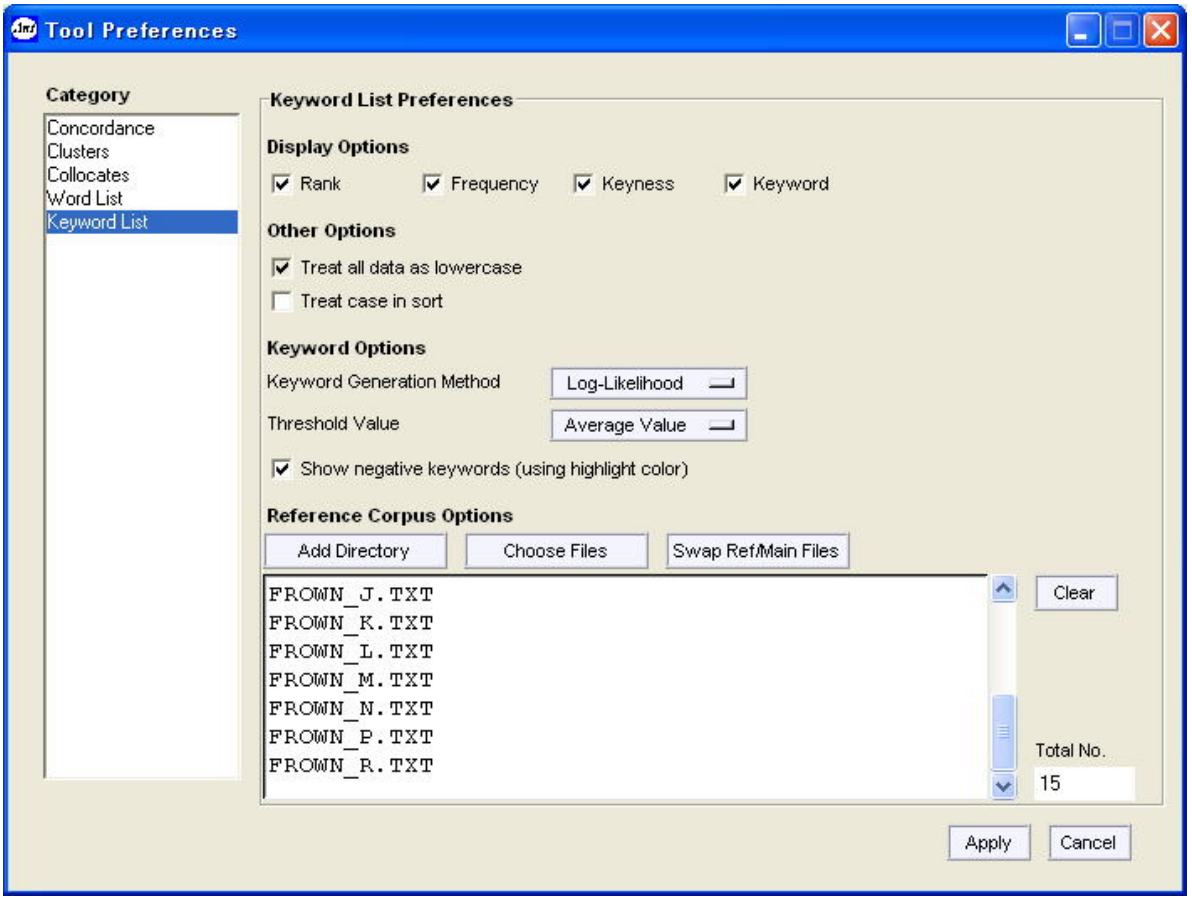
このように二つの値の間に意味のある差があるかどうかを考えるためには，
統計的な検定を行う必要があります。

ふつう，2値の有意差を見る場合は，「カイ二乗検定」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，
「カイ二乗検定」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。

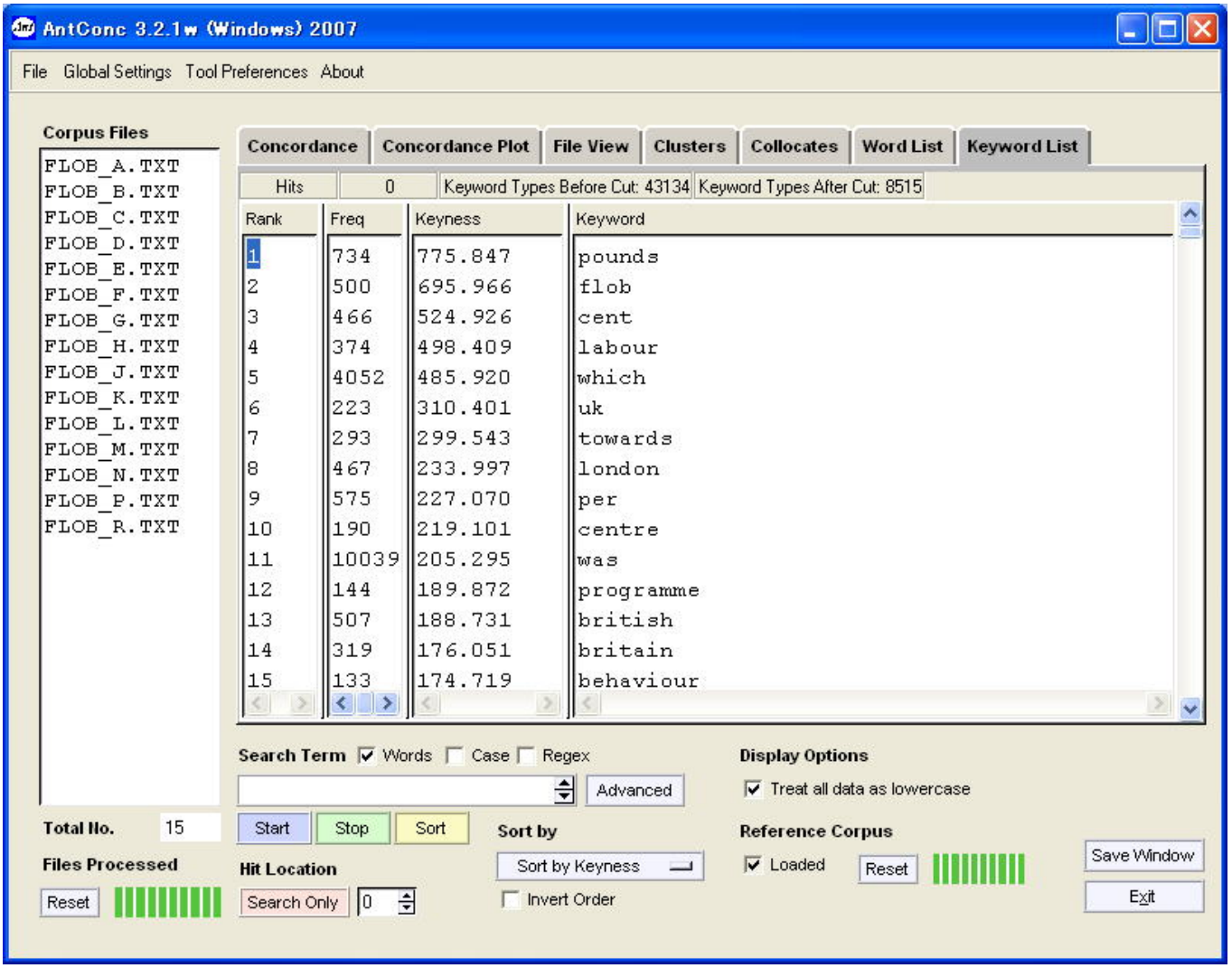
こうした計算は煩瑣なものですが，Antconcでは自動的に計算を行い，
統計的に意味のある特徴語だけをリストとして抽出してくれます。

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。



（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。



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

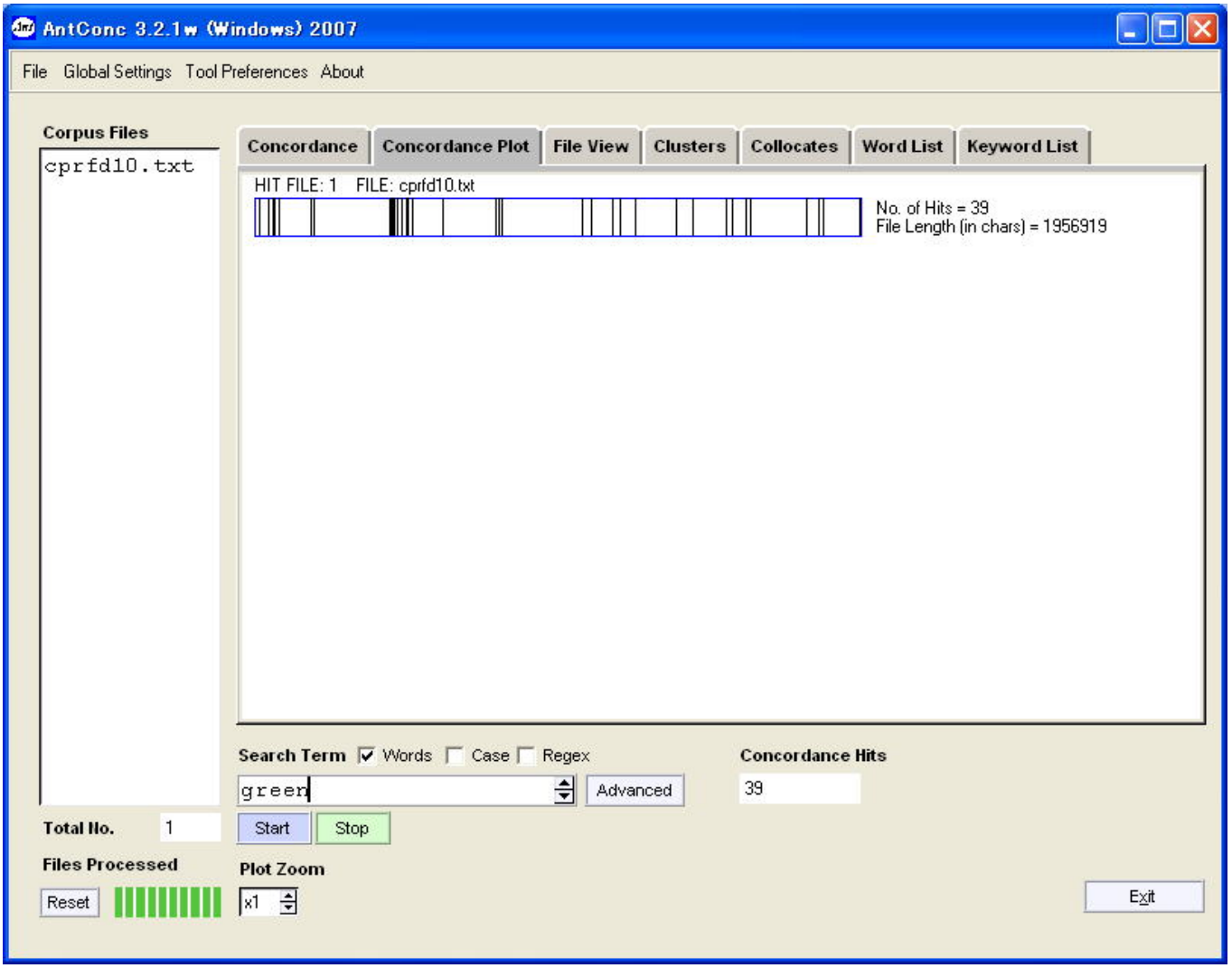
Antconcの「Concordance Plot」機能を使えば，データの中で特定の語が出現する位置を視覚的に把握することができます。

もっとも，FLOBのように，短いサンプルを寄せ集めたコーパ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。

すると下記のような結果が出力されます。



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