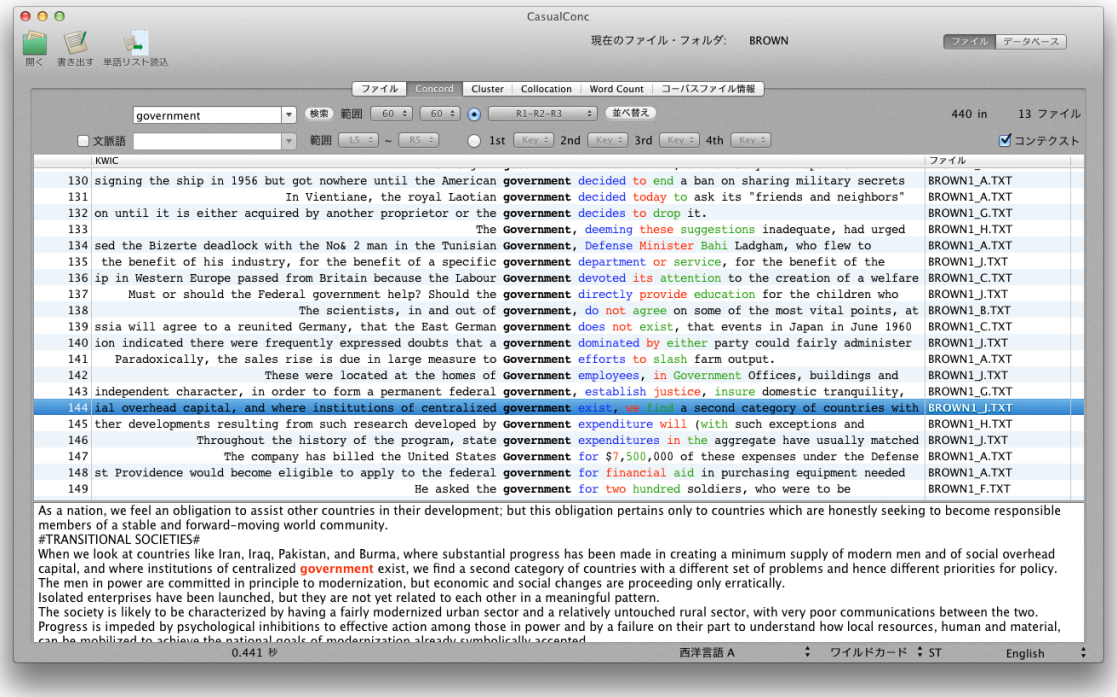
外国語教育メディア学会 (LET) 関西支部 メソドロジー研究部会 2011年度報告論集
今尾 康裕 (pp. 121–178)

Mac OS X 用コンコーダンサー CasualConc
—基本的な使い方と用例検索ツールとしての応用例—

今尾　康裕
大阪大学



Keywords: コンコーダンサー，コーパス，Mac，用例検索

1. CasualConcとは

CasualConc は Mac OS X 専用のテキスト分析ツールで，フリーウェアとして公開しています。[1] 元々は，WordSmith Tools や Txtana などの Windows 用コンコーダンサーや Mac でも動作する AntConc などを英語の論文を書く際の用例検索ツールとして試していたのですが，Mac では使えなかったり Mac 版の動作がいまいちだったりで，どれもこれもしっくりこないということで，ちょうど Ruby を使い始めた頃に Ruby で OS X 用のアプリケーションが開発できる環境が整ったのもあり，一念発起して開発を始めました。

開発には，RubyCocoa という Mac のアプリケーションを構築する「Cocoa」というフレームワークと日本発のオブジェクト指向スクリプト言語である「Ruby」をつなぐブリッジアプリケーションを使っています。具体的には，文字列処理を行うコアの部分はほぼ純粋な Ruby スクリプトで書き，GUI に関わる部分に RubyCocoa を使っていて，この GUI に関わる部分が Mac OS X に依存しているため，Mac OS X 専用のアプリケーションとなっています。

基本的な機能として，Windows 用などの多くのコンコーダンサーと同様に，テキストを含むファイルを読み込んで行う，KWIC（Keyword In Context）検索，単語クラスター検索，コロケーション検索，単語リスト・n-gram 作成などがあります。最初に起動した状態では，文字コードが UTF-8 もしくは ASCII のプレインテキストが扱えますが，設定を変えることで，これら 2 つ以外の文字コードが使われているプレインテキストや，PDF，HTML，Web Archive，リッチテキスト，MS Word 文書ファイルなどにも対応します。

テキストデータを扱う方法には，テキストファイルをそのまま扱う「ファイルモード」とテキストファイルからテキストを抜き出して SQLite データベースファイルを作成し，高速検索を可能にした「データベースモード」があります。デフォルトでは，どちらのモードも改行文字で区切られたパラグラフを単位とした検索をします。特に「データベースモード」ではパラグラフごとに一つのデータベースエントリーとなっていて，検索する文字列を含むパラグラフを SQLite で絞り込んでから処理をするため，多くの場合はファイルモードよりも検索速度が速くなります。

他のコンコーダンサーにはない CasualConc の特徴として第一にあげられるのが， Mac OS X 標準の GUI を使って Mac で native に動作するということです。そこで，ただ「Mac で動く」というだけではなく，Mac 用のアプリケーションらしい見た目や，Mac の扱いに慣れた人がある程度直感的に使えることを心がけて開発しています。さらに，現在の最新 β バージョン（2012 年 3 月末の執筆時点で 1.9.2）では，英語のインターフェイスの他に日本語も用意してあり，OS の言語環境が日本語であれば，表示はほぼ日本語になっています。本稿で使用するスクリーンショットは，すべて日本語環境で作成した日本語インターフェイスのものになります。

CasualConc のインストールは簡単で，CasualConc のサイトからディスクイメージをダウンロードして Finder 上で開き，中に入っている CasualConc を「Copy to Applications」のアイコンにドラッグ & ドロップして，アプリケーションフォルダに移動させるだけです（図 1.1）。ディスクイメージ

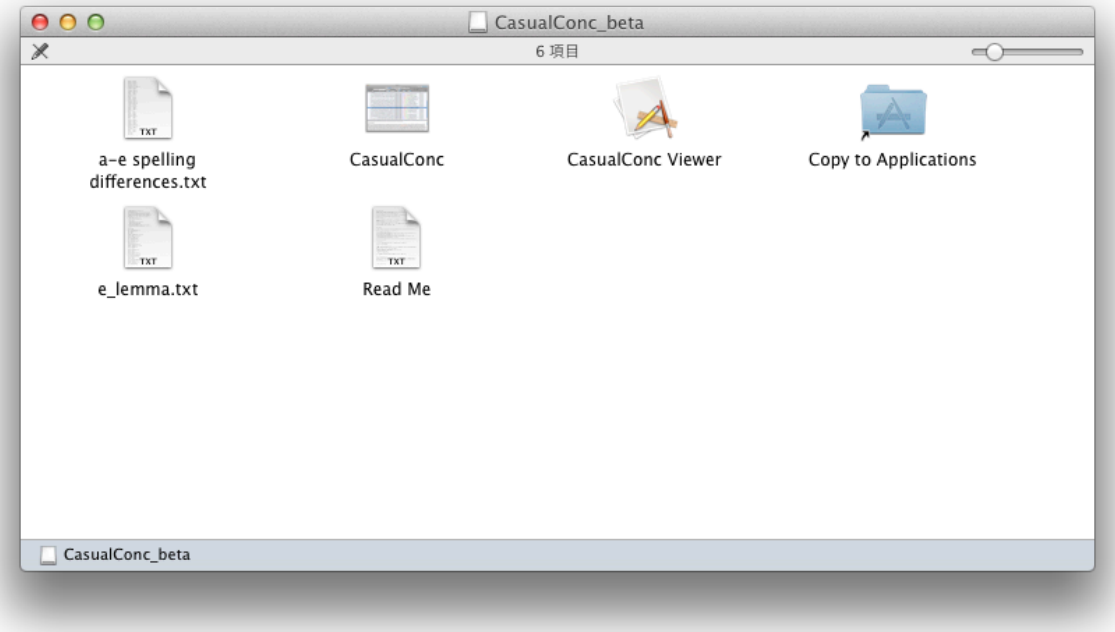


図 1.1 CasualConc のディスクイメージの中身

の中には，CasualConc 本体以外にもいくつかファイルが入っていますが，実際の使用例とともに本稿で説明していきます。

CasualConc のウェブサイトでは，CasualConc 以外にも，言語分析・教育に利用できるアプリケーションをフリーウェアとして公開しています。CasualConc を基にしたパラレルコンコーダンサーである CasualPConc/CasualMultiPConc，PDF や HTML ファイルなどからテキストを抜き出してプレインテキストファイルを作成する CasualTextractor，テキストへのタグ付けを補助する CasualTagger，MeCab を使って日本語形態素分析処理をする CasualMecab，メディアファイルの音声書き起しを補助する CasualTranscriber，IPA 記号を入力するための IPATypist があります。

本稿では，各ツールの機能を内部でどのような処理が行われているかにも少しずつ触れながら紹介した後，他のアプリケーションなどとも連携しつつ，用例検索ツール（特に英語の学術論文向け）として使うことを想定した利用例と言語研究の予備研究としての利用例を示します。 現在，サイトの情報を最新バージョン用に書き直している余裕がなく，サイトにある「使い方」の説明は古いバージョン（1.0.x）用のままですが，ここでは，最新バージョン（1.9.2，20120329）を基にして説明していきます。

2. 基本的な機能

CasualConc はファイルに含まれるテキストデータを読み込んで分析するアプリケーションなので，テキストファイルを用意する必要があります。使用に最も適したファイル形式はプレインテキスト形式で，ダウンロードして最初に起動した状態では，文字コードが UTF-8 もしくは ASCII のプレインテキストファイル（.txt）が扱えるように設定してあります。プレインテキストファイルでは，これら以外の文字コードも扱えますが，Mac OS の文字コードの自動判別の精度が高くないために自動判別の機能を使っていないので，文字コードはファイルごとに手動で設定する必要があります。このため，一度に読み込むファイルは，事前にすべて同じ文字コードに統一しておくことをお薦めします。また，プレインテキストファイル以外のテキストを含むファイルも扱えますが，テキストを抜き出す処理に多少の時間がかかるため，ファイル数が多い場合はあらかじめプレインテキストファイルに変換しておいた方がいいでしょう。ファイル形式の変換にはいろいろなフリーウェア，アップルスクリプト，Automator ワークフローなどがありますが，後述する CasualTextractor でもできるので試してみてください。

2.1. テキストファイルの扱い

最初に CasualConc を起動すると，Concord ツールが選ばれています（図 2.1）。この状態では，分析するためのテキストが選ばれていないので，まずは，分析対象となるテキストファイルを選択します。前述の通り，デフォルトでは文字コードが UTF-8 もしくは ASCII のプレインテキストファイルだけが読み込めるようになっていますので注意してください。

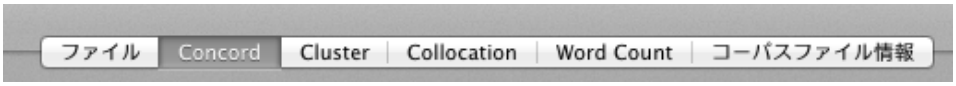


図 2.1 ツールタブ

図 2.2 「開く」アイコン

この状態でファイルを選択するには，メインウインドウの左上にあるアイコンのうち「開く」をクリックするか（図 2.2），メインメニューのファイルから「開く...」を選択します（図 2.3）。

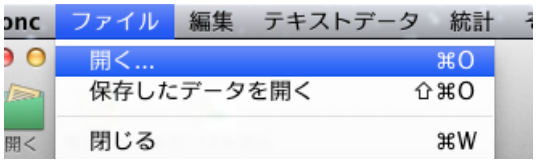


図 2.3 メインメニュー「開く...」

どちらの場合でも，OS X 標準のファイル選択パネルが現れるので，他の標準的な Mac 用アプリケーションでファイルを選択する際と同様に，分析対象に加えたいファイル，もしくは，そのファイルが含まれるフォルダを選択してください。ここでは，通常の Mac でのファイル操作と同じく，Command キーを押しながらファイル・フォルダを選択して複数を選択したり，Shift キーを押して連続する複数のファイル・フォルダを選択することができるので，一つのフォルダの中で特定の複数のファイルを読み込む，もしくは，いくつかあるフォルダのうち，特定のフォルダを選択するなどの操作もできます。ここでフォルダを選択した場合は，フォルダの中に入っているファイルのうち，読み込む設定にしてあるファイル形式のファイルすべてが読み込まれます。また，この方法で読み込む場合は，すべてのファイルが読み込むたびに新しく読み込んだものと置き換えられます。

このようにファイルやフォルダを選んですぐに Concord ツールでの KWIC 検索やその他のツールでの分析もできますが，CasualConc では，ただファイルを選んで分析するだけでなく，分析対象のファイルをもっと高度に扱うこともできます。

ファイルを扱うモードには，大きく分けて「テキストファイル処理」と「コーパスモード」の 2 つの区分があります（表 2.1）。「テキストファイル処理」には，テキストファイルをそのまま扱う「ファイルモード」と，ファイルからテキストを読み込んでデータベースファイルを作り，それを利用することで高速検索を可能とした「データベースモード」があります。「コーパスモード」には，難しいことを考えずに，読み込んだテキストファイルをすべて扱う，もしくは，一つのデータベースファイルだけを扱う「シンプルモード」と，複数のテキストファイルをグループ（コーパス）としてまとめて管理したり，複数のデータベースファイルを管理して，切り替えて利用したり，複数のコーパス・データベースファイルの横断処理を可能とする「アドバンストモード」があります。

ここでは，ファイルビューでの操作を中心に，この 4 つのモードそれぞれについて簡単に説明します。

表 2.1
テキストを扱う 4 つのモード

	テキストファイル処理	
コーパスモード	シンプルファイルモード	シンプルデータベースモード
	アドバンストファイルモード	アドバンストデータベースモード

2.1.1. シンプルファイルモード

最初に CasualConc を立ち上げた状態では，「シンプルファイルモード」になっています。前述のように，それぞれツールを選んだ状態からファイルを読み込んで検索・分析ができますが，ツールタブの「ファイル」をクリックすると，上部にファイルリストテーブル，下部にファイル内容を確認するプレビューテキストボックスが配置された「ファイルビュー」に切り替わります（図 2.4）。このファイルビューでは，テーブルとプレビューボックスは高さが変更できるので，テーブル上のファイルを一覧したいときなどは，プレビューボックスの上端あたりにカーソルを持っていってクリックし，クリックしたままでマウスやトラックパッド上の指を上下に動かすことで，サイズを変更できます。

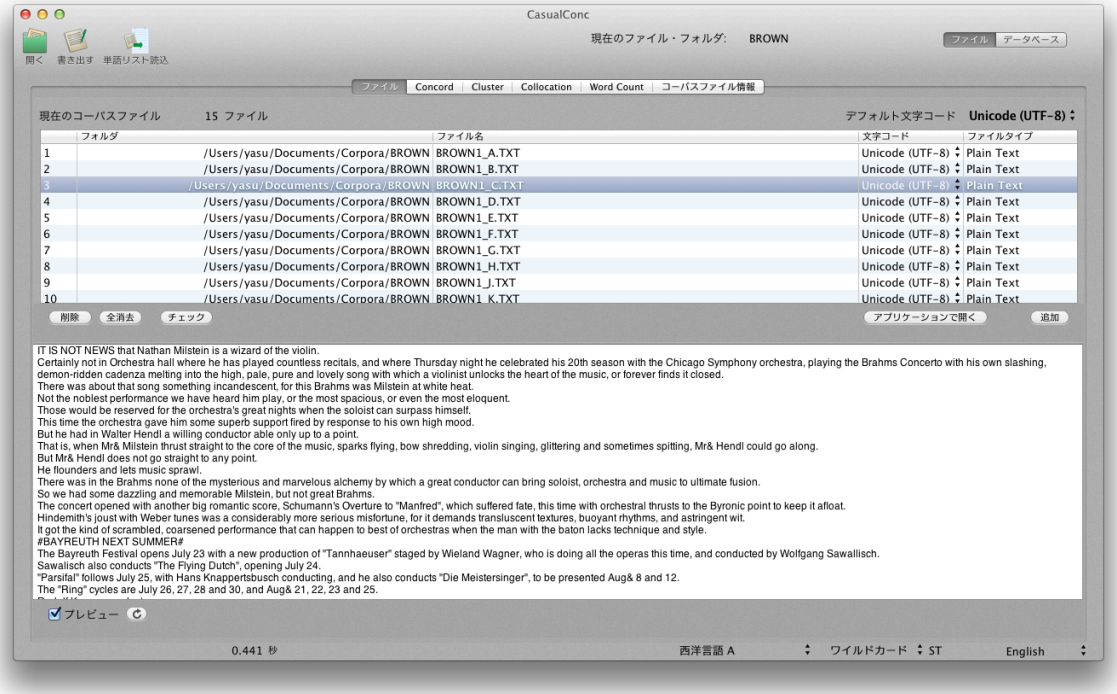


図 2.4 シンプルファイルモードの「ファイルビュー」

初期状態でファイルを読み込むと，どのツールが表示されている状態でファイルを読み込んでも，読み込まれたファイルはファイルビュー上部のファイルリストテーブルに表示されます。前述の通り，メニューやアイコンの「開く」からファイルやフォルダを選んだ場合は，毎回すべてのファイルがが新たに読み込んだファイルに置き換えられますが，このファイルビューでは，ファイルリストテーブル右下の「追加」ボタンをクリックすることで，新たに選択したファイルを既に読み込んだファイルリストに追加できます。また，Mac らしく，ファイルリストテーブルにファイルやフォルダをドラッグ & ドロップすることでもファイルを追加できます。

図 2.5 デフォルト文字コード設定

プレインテキストファイルをテーブルに追加する際は，

テーブル右上の「デフォルト文字コード」で選択された文字コードが設定されます（図 2.5）。最初に起動した状態では UTF-8 が選択されていますが，ファイルをテーブルに追加する前にデフォルト文字コードを変更しておけば，選んだ文字コードで追加できます。文字コードの設定は，ファイルを追加した後でもテーブル上で変更できますが，一つ一つ変更することになるため，読み込むファイルがすべて同じ文字コードの場合は，事前にデフォルト文字コードを変更して読み込んでください。また，デフォルト文字コードは CasualConc を終了しても保持されるので，読み込んだ後に，普段よく使う文字コードに戻しておくように心がけてください。また，このリストは英語や日本語を扱う上で代用的なものを含んでいますが，このリストにない文字コードが必要な場合は，要望を出していただければ，可能な限り対応します。ただ，CasualConc は，このように UTF-8 や ASCII 以外の文字コードにも対応していますが，不要な混乱を避けるため，普段からよく使うファイルは文字コードを統一しておき新規に作る場合は UTF-8 にしておくことをお薦めします。

ファイルリストテーブル上でファイルを選択した場合，ウインドウ左下の「プレビュー」チェックボックスにチェックが入っていると，ウインドウ下部のプレビューボックスにファイルの内容が表示されます。ここで内容が表示されない場合は，文字コードの設定が間違っている可能性があるので，テーブル上で選択したファイルの文字コードを変更して，「プレビュー」チェックボックスの右にある「リフレッシュボタン」をクリックし，プレビューを更新して確認します。また，テーブル上のファイルを選択した後に，環境設定で指定した外部アプリケーションでファイルを開いたり，テーブル上からファイルを削除することもできます（図 2.6，表 2.2)。さらに，メインメニューの「テキストデータ」から「選択されたファイルを Finder で表示」を選択すると，Finder 上で選択したファイルが含まれるフォルダが開き，ファイルを表示させることもできます。

図 2.6 シンプルファイルモード操作ボタン

表 2.2
シンプルファイルモード操作ボタンの動作

ボタン	動作
追加	テーブルにファイルを追加します。
アプリケーションで開く	選択したファイルを指定したアプリケーションで開きます。
削除	選択したファイルをテーブルから削除します。
全消去	テーブル上のすべてのファイルを削除します。
チェック	テーブル上のファイルが存在するかを確認します。

プレインテキストファイル以外のファイルを読み込むには，メインメニューの「CasualConc」から「環境設定」を選び（図 2.7），環境設定ウインドウで「ファイル」を選んで，読み込みたいファイルタイプのチェックボックスにチェックを入れます（図 2.8）。ファイルタイプおよび拡張子が指定してあるもののうち XML 以外は，それぞれのファイルタイプを認識してテキストを抜き出しますが，XML

図 2.7 メニューから環境設定

は基本的にプレインテキストファイルなので，そのように扱われます。また，「上記以外の拡張子」を選んで拡張子を指定した場合や「すべての拡張子」を選んだ場合も，あらかじめ設定のないファイル形式のものはプレインテキストファイルとして読み込みを試みます。つまり，どんな形式のファイルでも開けるからこのような設定があるわけではなく，一般的に入手できるコーパスの中に，プレインテキストファイルでありながら拡張子が付いていないファイルや，特殊な拡張子が付けられているファイルが存在するため，そのようなファイルに対処するために用意されています。

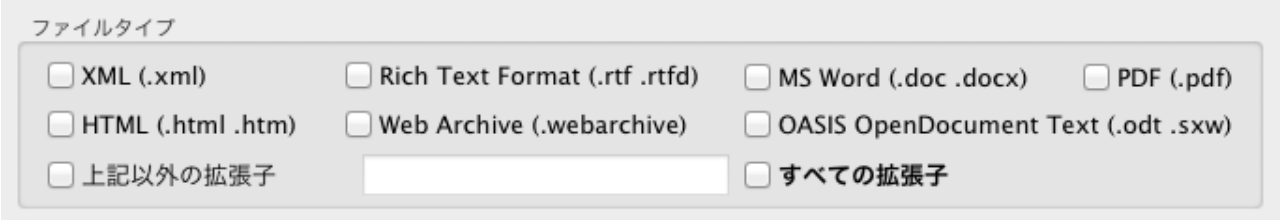


図 2.8 ファイルタイプ設定

ファイル環境設定では，この他にも，選択したファイルを開くためのアプリケーションの指定や，デフォルトのフォルダを設定することもでき，その設定は「シンプルファイルモード」だけではなく，他のモードでも反映されます。

2.1.2. シンプルデータベースモード

データベースモードに切り替えるには，メインウインドウ右上のボタンで「データベース」を選択します（図 2.9）。ファイルビューの外観はファイルモードとほとんど変わりませんが，ファイルリストテーブルは，分析対象となるファイルではなく，データベースに加えるファイルのリストとなり，右下にデータベースファイル作成およびテキストファイル追加のボタンが現れます（図 2.10）。また，右上の「開く」アイコンやメニューの「開く...」を選ぶと，使用するデータベースファイルの選択になるので，テーブルにファイルを追加する場合は，テーブル右下の「追加」ボタンをクリックするかファイルをテーブルにドラッグ & ドロップしてください。

図 2.9 モード設定

図 2.10 データベース用ボタン

ファイルリストテーブルにファイルを追加したら，ウインドウ右下の「新規DBファイル」ボタンをクリックして，名前をつけてデータベースファイルを作成し，保存します。最新のβ版では，データベースファイルが作成すると，そのファイルが選択されてすぐに使えるようになります。また，データベースファイルが選択されている状態で，ファイルリストテーブルにファイルを読み込んで「DBに追

加」ボタンをクリックすると，選択されているデータベースファイルにテーブル上のファイルを追加することができます。

少し技術的なことになりますが，CasualConc のデータベースファイルは SQLite ファイルとなっています。データベース作成のプロセスは，まず，テキストファイルを一つずつ読み込んで改行文字ごとに分割し，それぞれを一つのエントリーとしてデータベースに追加していきます。元のテキストファイルが段落ごとに改行されている場合は，一つの段落が一つのエントリーとなります。一文ごとに一つの区切りとしたい場合は，あらかじめ文ごとに改行したファイルを作成しておいてください。ファイルの読み込みに関しては，後述する様々な設定がデータベースファイル作成時に適用されます。

データベースモードの利点としては，指定した文字列を検索する Concord，Cluster，Collocation において，検索語を含むエントリー（段落，文）をあらかじめ SQLite の高速な検索で絞り込んでおき，絞り込まれたエントリーのみをテキスト処理するので，多くの場合で処理時間が大幅に短縮できます。ただし，すべての文字列を処理する Word Count での単語・n-gram リスト作成や， 機能語などコーパスに頻出する文字列を検索する場合は，ファイルモードよりも処理時間がかかる場合があります。また，あらかじめデータベースファイル作成時に後述する設定に従ってテキストの読み込み処理がされるため，データベースモードでの検索時には一部設定が反映されません。

データベースモードはあくまでも用例検索などを目的とした高速検索を念頭に置いており，正確さが多少犠牲になっています。簡単な検索であればファイルモードと結果が異なることはありませんが，ワイルドカード文字を使った検索や正規表現での高度な検索などでは，検索漏れが起こりうる可能性が否定できません。また，一度データベースファイルを作った後にテキストファイルを追加する場合などは，読み込み時の設定が異なると，出来上がったデータベースファイルを検索する際にデータの一貫性がとれなくなる場合があります。そのため，データベースモードは，用例検索や研究などの初期段階で大まかな傾向を見るのには適していますが，正確性が求められる分析をする際は，ファイルモードで設定を確認しながら行ってください。

2.1.3. アドバンストモード

アドバンストコーパスモードに切り替えるには，「環境設定」の「一般」にある「コーパスモード」で「アドバンスト」を選択します（図 2.11）。アドバンストモードになると，ファイルビューが左右に分割され，右側にシンプルモードと同様の追加用ファイルリストテーブルとプレビューボックスがあり，左側は上がコーパス・データベースを管理するテーブルで，下が選択したコーパス・データベースに含まれるファイルを確認するテーブルになっています（図 2.12）。

図 2.11 コーパスモード設定

ファイルモード・データベースモードともに，右上のテーブルにファイルを読み込んでから，左上のコーパス・データベースリストテーブル右下にある「新規コーパス」（ファイルモード），「新規データベース」（データベースモード）ボタンをクリックしてコーパス・データベースファイルを作成し，テーブルにコーパス・データベースを追加して管理します。

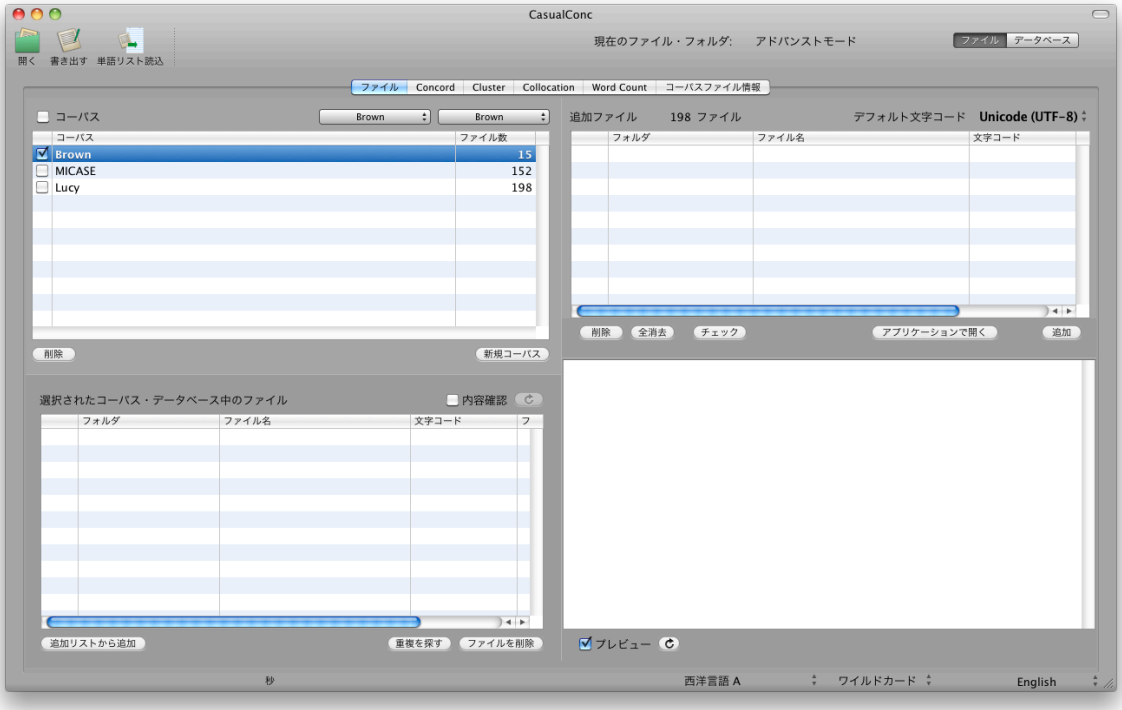


図 2.12 アドバンストモードファイルビュー

それぞれのコーパス・データベースファイルに含まれるテキストファイルを確認するには，左下テーブルの右上にある「内容確認」チェックボックス（図 2.13）にチェックを入れてから，左上のコーパス・データベースファイルを選択します。これで，左下のテーブルファイルリストを表示させることができます。既にコーパス・データベースファイルが選択されている場合は，チェックボックスにチェックを入れても表示されないので，その右横にある「リフレッシュ」ボタン（図 2.13）をクリックして表示させます。また，この内容ファイルリスト上のファイルを選択した際に右下のプレビューテキストボックスの「プレビュー」チェックボックスにチェックが入っていれば，右上の追加ファイルリストテーブルと同様に，ファイルの内容を確認できます。

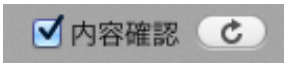


図 2.13 内容確認チェックボックス

このようにコーパス・データベースファイルに含まれるテキストファイルを表示させると，ファイルの削除・追加などの編集ができます（図 2.14，表 2.3）。ファイル追加の手順としては，作成時と同じように，右上の追加テーブルにファイルを読み込んでから，「追加リストから追加」ボタンをクリックする方法と，ファイル・フォルダを直接内容リストテーブルにドラッグ & ドロップする方法があります。削除はテーブル上で削除したいファイルを選んでから「ファイルを削除」をクリックしてください。「重複を探す」をクリックすると，同じファイルが 2 度以上読み込まれている場合にそのファイルが選択された状態になるので，「ファイルを削除」をクリックして削除します。

図 2.14 内容ファイルリスト編集用ボタン

表 2.3
内容ファイルリスト操作ボタンの動作

ボタン	動作
追加リストから追加	右上のテーブルに読み込まれたテキストファイルを選択されたコーパス・データベースファイルに追加します。
ファイルを削除	選択されたファイルをコーパス・データベースファイルから削除します。
重複を探す	ファイルリストで重複する物を探して選択された状態にします。

アドバンストモードの使うことの利点の一つに，コーパス・データベースファイルの管理があります。シンプルモードでは，読み込んだファイルをすべて処理する，あるいは，一度に一つのデータベースファイルを扱うことしかできないので，その都度読み込み直す必要がありますが，アドバンストモードでは，テキストファイルをまとめてコーパスとして管理する，複数のコーパスやデータベースファイルを横断的に検索する，検索時やリスト作成時に使うコーパス・データベースファイルを切り替えるなどの操作が可能となります。

実際に分析に使用する際は，コーパス・データベースファイルが登録されている左上のテーブルで，使用したいコーパス・データベースファイルの左側のチェックボックスにチェックを入れます（図 2.15）。ここで選択されたコーパス・データベースファイルは，テーブル右上のポップアップメニュー（図 2.16）と Concord および Collocation ツールの左上のポップアップメニューに追加されます。ここにある，2 つのポップアップメニューは，それぞれ Cluster と Word Count ツールの左右のテーブルに対応しています。「全て」を選択すると，チェックボックスにチェックが入ったコーパス・データベースファイルに含まれるすべてのテキストを検索・分析対象にし，リストから一つを選択すると，その選択したコーパス・データベースファイルだけを対象にします。これにより，Concord，Collocation ツールでは簡単にコーパス・データベースファイルを切り替えることができ，Cluster，Word Count ツールでは左右で異なるコーパス・データベースファイルの分析ができます。また，テーブルの左上にあるチェックボックスをクリックすることで，リスト上のすべてのチェックボックスにチェックを入れたり外したりできます。このテーブルで選択されたコーパス・データベースファイルの情報は，CasualConc を終了しても保持されますが，左上のチェックボックスを使ってすべてにチェックを入れたり外したりした場合は，その後にテーブル上でチェックボックスの操作をしない限り，その直前の状態が次回起動時に保持されます。

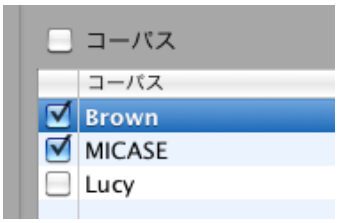


図 2.15 コーパスリスト

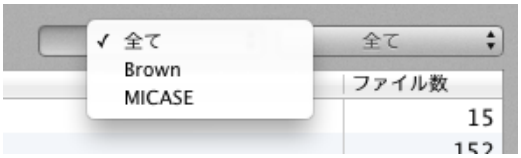


図 2.16 利用可能なコーパス選択

次に，それぞれのアドバンストモードについて少しだけ説明します。

2.1.4. アドバンストファイルモード

アドバンストファイルモードでは，テキストファイルをまとめてグループ（コーパス）を作って管理できます。内部での処理は，右上の追加ファイルリストテーブルに読み込まれたテキストファイルのファイルパス（ファイルが保存されている場所の情報）とそれぞれの文字コード情報を保持しているだけで，検索・分析処理ごとに元のファイルを直接読み込むため，検索・分析結果には処理時点での設定が反映されます。

前前述述の通り，左下の内容ファイルリストテーブルにファイルを表示させている状態で，リストそのものの編集はできますが，それぞれのファイルの文字コードの設定は変更できません。もし，間違った文字コードを設定したまま追加してしまった場合は，そのファイルを削除して，もう一度追加し直してください。

2.1.5. アドバンストデータベースモード

アドバンストデータベースモードでは，シンプルモードと同様の手順で作成したデータベースを左上のテーブルで管理します。データベースリストテーブルでは，右上の追加ファイルリストから新規のデータベースを作成してリストに追加するだけでなく，シンプルモードで作成したデータベースファイルの追加やリストからのファイルの削除，および，何らかの要因でデータベースファイルの概略データ（含有ファイル数など）に問題が生じた場合の修正などができます（図 2.17）。

図 2.17 データベース管理ボタン

2.1.6. ファイル読み込み時の設定

ファイルモードでテキストファイルを読み込んで処理する際，および，データベースファイルを作成する際に，ファイルに含まれる特定の文字を置換する設定があります。この機能の目的は，特に 1 バイト文字言語（英語など）のテキストを分析する際に，UTF-8 では 2 バイト以上が割り当てられている特定の記号などが，検索の際に記号ではなく文字として認識されてしまい，分析ノイズが入るのを防ぐことにあります。例えば，PDF や HTML などから抜き出したテキストで，引用符や二重引用符が ' ではなく ‘’，" ではなく “” など「curly quotes」になっている場合に，これらの記号を 1 バイトの記号に置き換えて，分析から除外するために用います。ただ，テキストファイルの文字数にもよりますが，多少の処理時間がかかるので，頻繁に利用するファイルでは，あらかじめ他のアプリケーションでこれらの文字列を置換しておくことをお薦めします。また，データベースファイルを作成する場合は，ファイル作成時にこの設定が適用されて，分析・検索処理時には適用されないので，あらかじめ考慮に入れておいてください。

図 2.18 文字置換処理設定

この機能はデフォルトではオフになっていますが，「環境設定」の「一般」にある「文字の置換処理」チェックボックスにチェックを入れることで使えるようになります（図 2.18）。

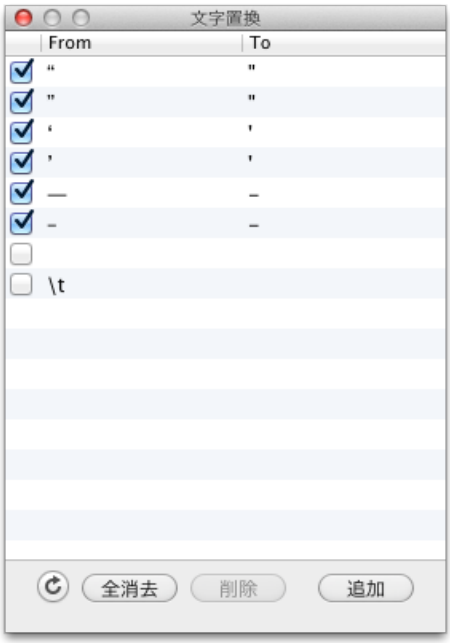


図 2.19 文字置換処理パネル

チェックを入れて「リストを表示」ボタンをクリックすると，置換文字列を設定するパネル（図 2.19）が表示され，このリストの左端のチェックボックスにチェックが入っている文字に対して置換処理が実行されます。 このパネルにはあらかじめ英文テキストで頻出する記号が登録されていますが，実際に使用する際はお使いのテキストファイルに含まれる文字にあわせて運用してください。

新規に置換したい文字を登録するには，「追加」ボタンをクリックすると新しい行が挿入されるので，その行をダブルクリックし，左側（From 列）に置換したい文字を，右側（To 列）に置換後の文字を入力します。「削除」ボタンをクリックすると，選択されている文字の行がテーブルから削除され，「全消去」をクリックすると，テーブル上のすべての行が消去されます。何らかの理由で元から登録されている文字を復活させたい場合は，左端の「再読み込み」ボタンをクリックします。

2.2. 各ツールの機能

次に各ツールの基本的な機能を説明していきますが，その前に，すべて，もしくは，複数のツールに影響する設定について説明します。

2.2.1. 共通設定

複数のツールに影響する設定は，ほとんどを「環境設定」の「一般」で行います。メインメニューから「環境設定」を選んで，環境設定ウインドウを表示させて設定してください。

2.2.1.1. 文脈処理

CasualConc では，読み込んだテキストを段落単位で処理するのがデフォルトとなっています。具体的には，改行文字で分割された単位が「段落」という扱いになり，Concord ツールの KWIC 検索での文脈で表示される範囲，Cluster，Collocation，n-gram リスト作成時の文脈語として扱われる範囲がこの単位となります。例えば，n-gram の場合は，段落の区切りまでの連続する n 個の単語の集まりで，段落を超えるものは含まれません。

文脈の範囲は，「環境設定」の「一般」で設定します（図 2.20）。「段落」の他にも「文」，「ファイル」が選べますが，それぞれの処理に時間がかかります。特に「文」を文脈の範囲に指定する場合は，指定した文字を基にして文ごとに分割して処理するため，処理時間がかかるだけでなく正確性にも欠けるので，文を区切りとした分析をする場合は，あらかじめ文ごとに改行したファイルを用意することをお薦

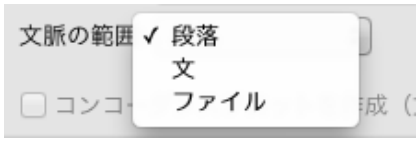


図 2.20 文脈の範囲設定

めします。どうしても使用したい場合は，デフォルトで指定してある文末記号もしくは任意の文末記号を入力し，「リストを表示」で現れるパネルで文末の例外を指定して（Mr. など）処理します。

「ファイル」に設定した場合も時間がかかりますが，ファイル中での特定の単語の位置が重要になる場合（コンコーダンスプロット作成など）や，段落の区切りとは関係なく改行されているテキストファイル（一定の文字数ごとに改行されている場合など）を扱う際には必要です。ただ，後者の場合は，処理時間を考えると，あらかじめテキストファイルを段落ごとの区切りに変更しておいた方がいいでしょう。

2.2.1.2. 単語の定義

CasualConc では，単語リストや n-gram リスト作成のみならず，Concord，Cluster，Collocation ツールでの文脈語も単語ごとに処理されるため，単語という単位が重要な意味を持っています。内部処理では正規表現を使っているので，\b\w+\b で表されるコンピューターが認識する 1 バイト文字の記号や空白文字以外の連続した文字列というのが単語の定義になります。ただ，これだけでは 1 バイト文字を多用する言語を分析する際に不都合が多くなるため，単語として扱える文字列をある程度柔軟に指定できるようになっています。

設定は，環境設定ウインドウを開き，「一般」の下の方にある「単語の一部として含める文字」で，それぞれ適用したい項目にチェックを入れます（図 2.21）。大きく分けると 4 つ項目がありますが，「Include Words」では，ピリオドを含んだり，常に一つの単語として扱われる文字列を設定します。例えば，論文などでく見かける e.g. などを指定すると，e と g の 2 つの文字としてではなく，e.g. という一つの単語として処理されるようになります。「指定した連語を単語として扱う」も，内部処理はほぼ同じなのですが，特定の検索時だけなど，常に適用するわけではない文字列を区別するために分けてあります。

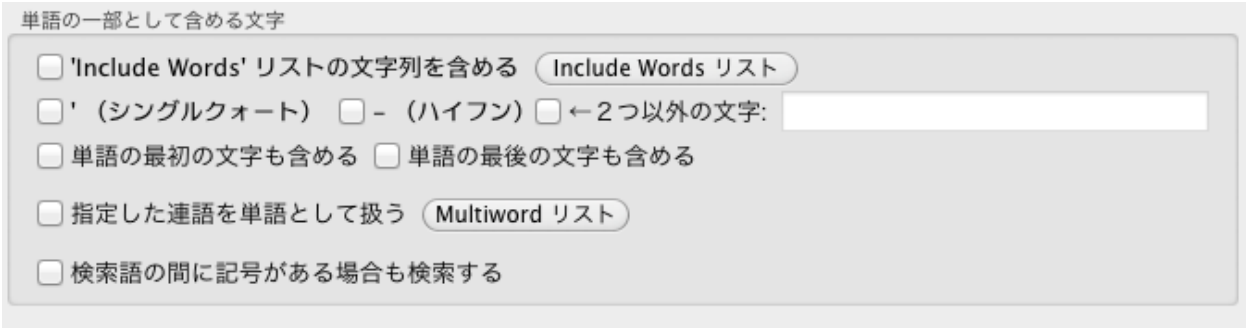


図 2.21 単語の設定

このどちらの場合も，右側の「〜リスト」ボタンをクリックするとパネルが現れるので，パネル上のリストに追加していきます（図 2.22）。このパネルの使い方は，後述する Stop Word/Skip Character とも共通で，メインメニューの「ウインドウ」から「Stop Word/Skip Character リストパネル」を選んでも表示させることができます。使い方は，まず左側テーブル下にあるテキストボックスにグループ名（言語名など）を入力して「追加」ボタンをクリックし，単語や文字処理の情報をまとめ

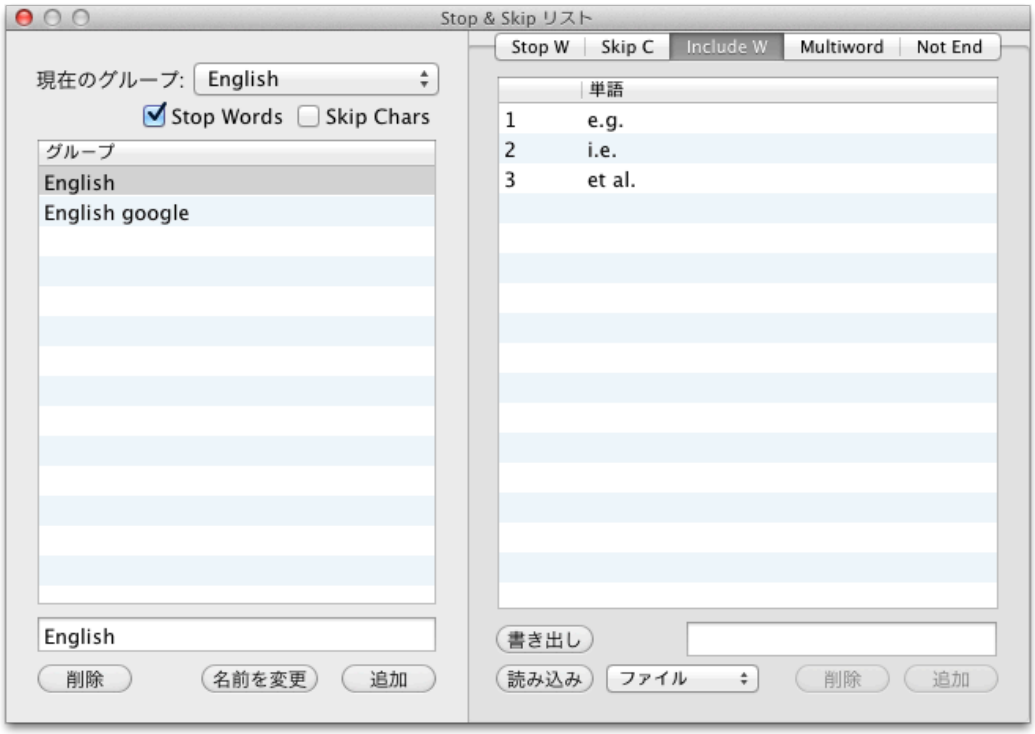


図 2.22 単語として扱う文字列の指定

て管理・運用するグループを作成します。ここでのグループは， Stop Word/Skip Character なども含めて扱われるので注意してください。次に，グループリストテーブルで，作成したグループを選択した後に，右側のテーブルで項目を追加していきます。環境設定の「Include Words リスト」，「Multiwords リスト」ボタンをクリックすると，それぞれ対応したタブが選択された状態でパネルが開きますが，開いた後に上部のタブで他の項目に変更することもできます。ただ，その場合は，適切なタブが選択されていることを確認してください。また，ひとつひとつ入力するだけでなく，一行一単語で入力されたテキストファイルを読み込んだり，他のグループのリストを読み込んだりすることもできます。別のグループで作ったリストを使いたい場合は，「読み込み」の右側のポップアップメニューで「リスト」を選んで，他のグループから読み込んでください。

次に，「シングルクォート」，「ハイフン」にチェックを入れると，半角の ' や - を単語の一部として扱うことができます。これは英語の所有格や短縮形，ハイフンでつないだ連語を一つの単語として扱う場合などを想定しています。「← 2 つ以外」にチェックを入れて，その右のテキストボックスに文字を入力すると，その文字を単語の一部として扱うこともできます。その下の「最初」と「最後」の文字も含めるにチェックを入れると ' や - もしくは指定した文字が単語の最初や最後にある場合も単語の一部として扱われますが，これは，言語によっては通常は記号として扱われる文字が単語の一部として扱われる場合などを想定しているので，通常はオフにしてください。英語などで使う場合は，引用や強調などで使われている記号との間に半角スペースを入れるなどの処理をしないと，単語に含まれないものも単語の一部として扱われてしまいます。

一番下の「検索語の間に記号がある場合も検索する」は，単語自体の扱いではないのですが，Concord などで連語を検索する際に，単語間にハイフンなどが入っている用例を検索結果に含める必

要があるときのため，また，Cluster や n-gram リストから Concord で検索する際に頻度が合わなくなることを防ぐための機能です。通常はオフで使ってください。

2.2.1.3. 検索モード

Concord，Cluster，Collocation および Word Count ツールの指定文字列検索モードで文字列の検索をする際に，入力された文字列をどのように扱うかを設定します。設定の変更は，環境設定パネルの「一般」もしくはメインウインドウ右下のポップアップメニューで選択します（図 2.23）。デフォルトでは「ワイルドカード」になっています。

図 2.23 検索モードの指定

「ワイルドカード」では，ワイルドカード文字（表 2.4）[2] を使って，柔軟な検索が可能になっています。内部処理としては，ワイルドカード文字を含めて正規表現に変換して検索しています。例えば，it (is|was) ? * that という文字列を検索すると，it is possible that，it was possible that，it is only natural that などが見つかります。ワイルドカード文字自体を検索したい場合は，表 2.4 の一番下に示すように，** で * を，?? で ? を，$! で ! を指定します。これ以外の記号は基本的に検索可能となっていますが，あくまでも文字列検索のモードなのでうまく検索にかからないとも限りません。このため，記号を含めた複雑な検索をしたい場合は，「正規表現」で検索したほうが確実です。

「文字」では，入力したすべての文字を正規表現のエスケープ処理し，記号を含めた指定した文字列そのものが検索できるようになっています。「正規表現」では，Ruby の正規表現（鬼車）が使えます。基本的には他の言語の正規表現と同じですが，細かい差異についてはインターネットなどで調べてください。「正規表現」では大文字小文字を区別するかどうか，複数行に渡る文字列にマッチさせる

表 2.4
ワイルドカード検索で特別な意味を持つ文字

ワイルドカード文字	意味
*	0 または 1 以上の連続した半角記号以外の任意の文字列
?	1 以上の連続した任意の半角記号以外の文字列
!	半角記号以外の任意の 1 文字
/	検索文字列の区切り（異なる単語・フレーズとして検索されます）
(A|B)	A もしくは B
** / ?? / $! / $$	* / ? / ! / $

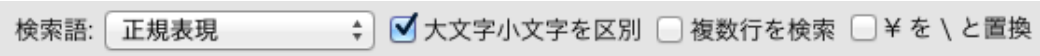


図 2.24 正規表現のオプション

かどうか，日本語キーボードでバックスラッシュ（\）の代わりに用いられる円記号（¥）をバックスラッシュに自動変換するかどうかのオプションがあります（図 2.24）。ただし，前述のように，文脈の範囲が「段落」になっていると，改行記号をまたいだ検索はできませんので，「ファイル」を選んで検索してください。「タグ」，「タグのみ」は，タグづけされたテキストを扱うモードで，タグで検索したり，単語リストの表示でタグと単語を別のコラムに表示したりできますが，ここでは詳しいことは省きます。

2.2.1.4. Lemma/キーワードグループ/異綴り処理

Lemma とは，辞書などの見出し語のように文法的な派生などを考慮しない語幹となる語 (Biber et al., 1998, p.25) で，例えば，lemma ‘see’ には see, saw, seen, sees, seeing が含まれます。CasualConc の lemma 機能は，他のコンコーダンサーと同じく，lemma とそれに含む単語を指定したリストを作るか，または既にあるリストを読み込んで，lemma ごとに単語を処理します。この lemma リスト作成には大変な労力が必要ですが，関西大学の染谷泰正先生に許可をいただいて，e-lemma ファイルを最新β版と一緒に配布していますの。この e-lemma ファイルを読み込むことで，手軽に lemma 処理が行えます。ただ，既存のどの lemma ファイルにも言えることですが，完全な lemma リストというのは存在しないので，必要に応じて修正，項目の追加等を行ってください。また，この e-lemma リストは品詞情報を含まないため，コーパスの概要を把握するのには十分ですが，詳しい分析を行う場合は，lemma を判断してタグづけする品詞タガーなどで処理した後に人手で修正したタグづけコーパスなどが必要になるかもしれません。

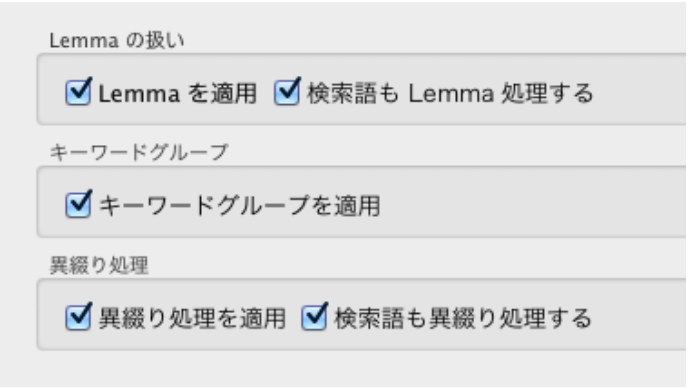


図 2.25 Lemma 等の設定

この「Lemma」機能を使うには，まず，「環境設定」の「Lemma」で「Lemma を適用」にチェックを入れます（図 2.25）。これで，Lemma パネルが表示できるので，メインメニューの「ウインドウ」から「Lemma リストパネル」を選んで開き，前述の Include Words の所と同様に，左側のテーブルでグループ（Language）を作成します（図 2.26）。ここで，左側のテーブルのグループを選んだ後，右側のテーブルの「読み込み」ボタンをクリックして，lemma ファイルを選択します。e-lemma ファイルを使う場合は，CasualConc のディスクイメージに入っている e-lemma ファイルを選択してください。このファイルは文字コードが UTF-8 に変換してあるので，デフォルトの設定のまま読み込めますが，これ以外のファイルから読み込む場合は文字コードなどを確認して指定してください。

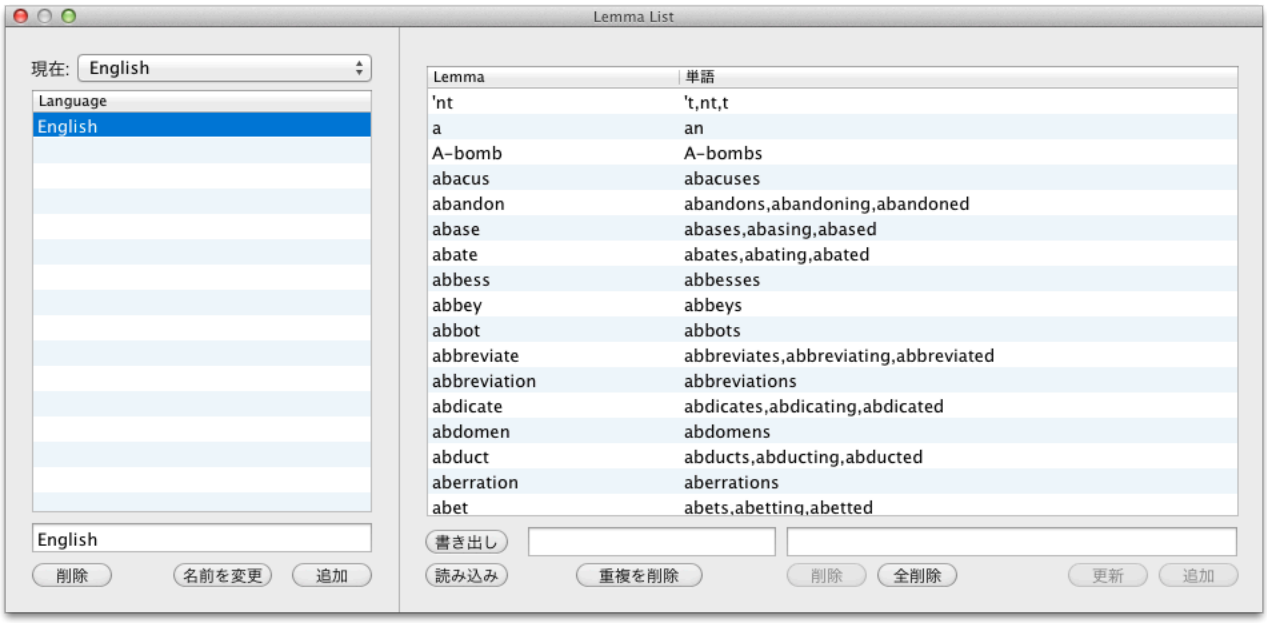


図 2.26 Lemma リストパネル

読み込んだリストは，テーブル上の lemma を選んで下部のテキストボックスで修正し「更新」ボタンをクリックしてパネル上で修正したり，テキストボックスに新たな lemma を入力して「追加」ボタンをクリックして追加できます。これで，Cluster，Collocation，Word Count のリスト作成時に lemma 処理が実行されます。複数のグループがある場合は，左上の「現在」のところで，使用したいグループを選択してください。また，図 2.26 にある環境設定のパネルで「検索語も Lemma 処理する」にチェックを入れると，検索語と同じ lemma の単語がすべて検索されます。例えば，e-lemma を使って see を検索すると，see，saw，seen，sees，seeing が検索されます。

「キーワードグループ」は，lemma 機能と同じような仕組みで処理されますが，任意の文字列を見出し語にして，それに登録されているすべての単語を検索する際に使うもので，検索語のみに適用されます。例えば，DAYS という見出し語に Monday，Tuesday... と曜日を登録しておけば，登録された単語を一度に検索できますが，lemma 機能と違い，見出し語は検索に含まれません。検索時には，見出し語に半角の @ を二つ重ねた @@ を付けて @@DAYS のようにします。このときの検索見出し語は大文字小文字の区別があるので，登録してある通りの文字を使ってください。リストの作成は，「環境設定」で「キーワードグループを適用」にチェックを入れた後に，「キーワードグループリストパネル」を開いて行ってください。

「異綴り処理」機能は，英語のイギリス綴りとアメリカ綴りなどを対応させたリストを作って読み込むことで，どちらも漏らさず検索したり，リスト作成時に同じ単語として扱ったりすることができます。これも，lemma と同様に，「環境設定」でチェックを入れた後，「異綴りリストパネル」を開いてリストを作るか，ファイルからリストを読み込みます。この異綴り処理も lemma 処理と同様に検索語にも対応しており，lemma 処理と異綴り処理は同時に適用することもできます。実際どのような機能なのかを知りたい方は，CasualConc のディスクイメージには，実験として作った「a-e spelling differences」という名前のイギリス綴りとアメリカ綴りを対応させたファイルがあるので試してみてください。このファイルも，e-lemma ファイルと同様に完全なものではなく，また，OS X のスペル辞

書機能などを使って半自動で作ったので，研究目的というよりも用例検索補助としての意味合いが強く，正確さは保証できません。

2.2.1.5. Stop Words/Skip Characters 処理

Stop words とは，英語の冠詞や前置詞などの機能語を中心とした，コーパスに頻繁に現れる単語のことで，単語リストやコロケーション単語リストなどで内容語を中心とした特徴語を探す際に障害になることもあります。そのため，CasualConc には，あらかじめ定義した stop words を分析から排除する機能が付いています。Stop words リストは自作するか，インターネットで検索すれば簡単に見つかります。[3]ファイルを入手したら，メインメニューの「ウインドウ」から「Stop Word/Skip Character リストパネル」を選んでパネルを表示し，上記の Include Words の場合と同じように左側のテーブルでグループを作って選択し，右側のテーブルにファイルからリストを読み込みます。ここでもこれまでと同様にリスト上の単語の編集ができます。

Stop words を適用する場合は，左上の「現在のグループ」で使用したいグループを選択して，「Stop Words」にチェックを入れます（図 2.27）。また，「環境設定」の「その他」で stop word 処理を適用したいツールを選択することもできます（図 2.28）。ただし，Concord などでは，stop words 処理を適用させたままで検索した際に，検索語自体に stop words が含まれると，何も検索できなくなるので注意してください。これは，検索語でワイルドカード文字を使った場合に，そこに当てはまる単語に対して stop word 処理をすることを前提としているためです。

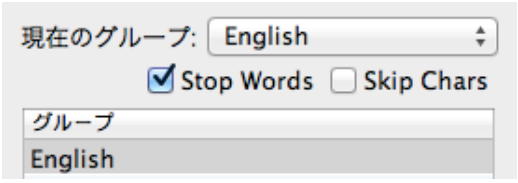


図 2.27 Stop Words チェックボックス

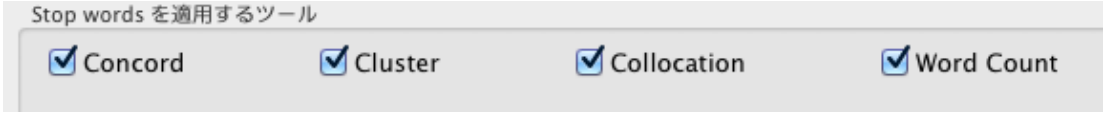


図 2.28 Stop Words のツールへの適用

Skip Characters は日本語などの 2 バイト文字で，記号が記号としてではなく文字として認識されてしまう言語を扱う場合を想定しています。例えば，句読点などを登録しておくと，文字としてカウントされなくなります。リストの作り方や使い方は，Stop Words の場合と同様です。

2.2.1.6. 結果の保存・書き出し

各ツールでの検索結果やリストなどのテーブル上の情報は，メインメニューの「ファイル」から「テーブルの結果を保存」を選んで，CasualConc で後に読み込むために保存するか，もしくは「テーブルの結果を書き出す」を選んで，CSV または Tab-delimited のテキストファイルとして書き出すことができます（図 2.29）。「保存」を選んだ場合は，それぞれのツールに対応した拡張子が付けられてファイルが保存され，同じメニューの「保存したデータを開く」で CasualConc に読み込むことができます。書き出されたファイルは，CasualConc で読み込むことを前提に，ツールによって SQLite

図 2.29 結果の保存および書き出し

データベースファイルか XML ファイルになっており，内部処理に必要な情報を含めて，CasualConc が保持している情報をほぼすべてそのまま保存します。「書き出す」を選んだ場合は，標準では拡張子「.csv」が付けられて CSV 形式のテキストファイルとして保存されますが，「.txt」に変更したい場合は，ファイル名に .txt を付けて保存してください。こちらの場合は，結果の情報を他で活用するためのもので，基本的には画面上で見えている情報だけが書き出されます。Cluster と Word Count には左右 2 つのテーブルがありますが，メニューで「右テーブル...」を選ぶことで，右側のテーブルの結果の保存・書き出しができます。

2.2.2. Concord

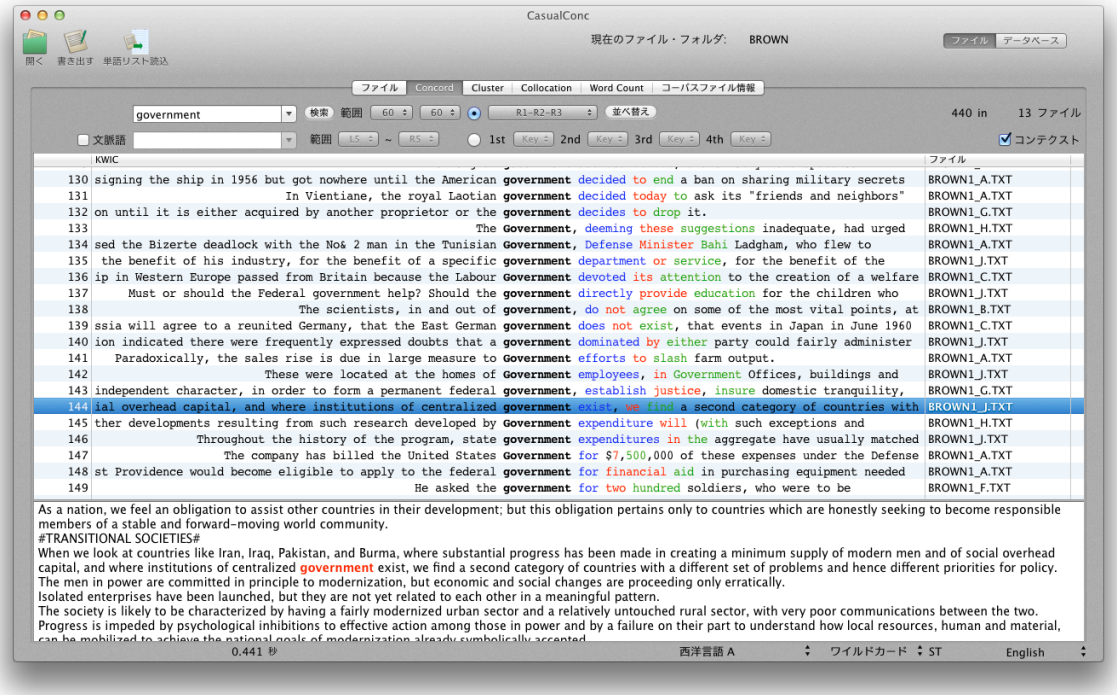


図 2.30 Concord ツール

Concord ツール（図 2.30）は，いわゆる KWIC (keyword in context) 検索のためのツールです。左上のテキストボックスに検索したい単語を入力して，「検索」ボタンをクリックするか，キーボードの Enter キーを押すと，指定されたテキストファイル，もしくは，データベースファイルで検索します。アドバンストコーパスモードでは，チェックが入ったコーパス・データベースファイルが検索テキストボックスの左にポップアップメニューとして現れて，そこで，すべてを横断的に検索するか，一つを選んで検索するかが選べます（図 2.31）。検索が終了すると，検索結果がテーブルに表示され，文脈を詳しく見た

図 2.31 コーパス選択

い行をテーブル上でクリックすると，元ファイルのテキストが検索語が強調された状態で下のコンテクストビューに表示されます。このコンテクストビューでは，Mac OS X の標準の機能として，コンテクストメニューから選択した語の意味を OS に標準の辞書アプリケーションで調べたり，「スピーチ」で読み上げさせたり，Google で検索したりできるだけでなく，Concord，Cluster，Collocation で検索することもできます。また，テーブル上の結果を選んだ際のコンテクストビューの表示は，右上の「コンテクスト」チェックボックスでオン・オフの切り替えができます。

検索結果は，文脈語の並べ替え指定の順に並べ替えられて表示されます。また，「並べ替え」ボタンをクリックすることで，検索結果を並べ替え直すこともできます（図 2.32）。デフォルトでは，上にある並べ替え順プリセットで指定しますが，左のラジオボタンの下側をクリックして，1 番目から 4 番目まで任意の並べ替え順を指定できます。この任意の並べ替えでは，指定した位置の文脈語のアルファベット順だけではなく，検索した語がファイルで現れる順番で行うこともでき，その場合は，1st で FN（ファイル名），2nd で POS（位置）を選択します。並べ替え順のリストに現れるそれぞれの選択肢がどのような意味を持っているかは，表 2.5 を参照してください。

図 2.32 Concord 並べ替えオプション

表 2.5
Concord ツールの文脈語範囲および並べ替え順指定

ラベル	意味
Key	結果の中央に位置する検索語
L1 ~ L5 (L10)	検索語の左 1 番目から 5 (10) 番目の位置の単語
R1 ~ R5 (R10)	検索語の右 1 番目から 5 (10) 番目の位置の単語
FN	ファイル名
POS	ファイル中の検索語の位置

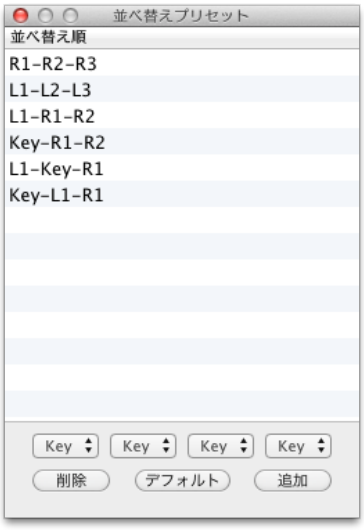


図 2.33 並べ替えリスト編集

最初に登録されたプリセット以外によく使う組み合わせがある場合は，「環境設定」の「Concord」中程右端にある「並べ替え」というボタンをクリックすると現れる「並べ替えプリセット」パネルで追加・編集することができます（図 2.33）。また，「環境設定」にある「広文脈モード」にチェックを入れると左右 10 番目の位置の単語まで指定することができるようになりますが，並べ替えに利用できる単語は，あくまでも画面上で表示されている範囲のものに限られるので，広文脈モードを利用する際などは，あらかじめ文脈語の表示範囲を広く取ってください。プリセットでは，左右 10 番目までを含めることができますが，通常の並べ替えモードでは左右 5 番目の単語までしか並べ替えに反映されないので注意してください。

図 2.34 Concord 入力系

検索語の左右に表示される文字の量を変更したい場合は，検索前に「範囲」のところで，検索語の左右に現れる文脈の文字数を変更します（図 2.34）。現行β版では，0 から 170 文字の間で 10 刻みで指定できます。文脈を前中心，後ろ中心などで表示させたい場合や，もっと広く見たい場合には値を変更してください。下の「文脈語」にチェックを入れると，範囲（L5: 左 5 つ目の単語，R5: 右 5 つ目の単語）を指定して，右のテキストボックスに入力した単語がその範囲で現れる結果だけに絞り込むことができます。また，「環境設定」の「Concord」中程にある「文脈語除外」にチェックを入れると，もう一つ下にテキストボックスが現れて，「除外」にチェックを入れてテキストボックスに単語を入力して検索すると，指定した範囲にその単語が現れたものが除外されて表示されます（表 2.6）。

表 2.6
文脈語絞り込み検索

通常	ıgress, the Federal **Government** assumed responsibilit which the national **government** assumed the responsi ponsibility of the **Government** at all levels to help ıt by declaring the **government** at Richmond in the ea the United States **Government** at the time. **Government** attorneys, Leavitt sa life, and when any **government** becomes subversive of consumers, with the **government** being no more than "a the operation of a **Government** Bid Center, which rec past year, 10,517 **government** bid invitations were which does have a **Government**-blessed monopoly.
文脈語絞り込み **(the)**	ıgress, the Federal **Government** assumed responsibilit which the national **government** assumed the responsi ponsibility of the **Government** at all levels to help ıt by declaring the **government** at Richmond in the ea the United States **Government** at the time. consumers, with the **government** being no more than "a the operation of a **Government** Bid Center, which rec past year, 10,517 **government** bid invitations were ot from the Federal **government** but from an exchange ıld not abolish the **government**, but would emphasize formed a new state **government** by declaring the gove
文脈語除外絞り込み **(the)**	less combat-tested **government** army in monsoon-shroud there was and such **government** as there was- passed s sent to all local **government** assessors or boards o or no government-to-**government** assistance. eir income by small **government** assistance, by tutori **Government** attorneys, Leavitt sa life, and when any **government** becomes subversive of , which does have a **Government**-blessed monopoly. lls, and in British **government** bonds and stocks. ions in longer-term **Government** bonds, they will certa en is clear: men of **government**, business men, lawyers

検索語，文脈語，除外語のいずれも，検索履歴が残り，テキストボックス右端の下向き三角をクリックするとそれまでに検索した語が表示されるので，そこから選んで検索することができます。履歴を残す数は，デフォルトでは 15 になっていますが，「環境設定」で変更することができます。ただし，検索語は他のツールとも共通なので「一般」で，文脈語・除外語は Concord のみに適用されるので「Concord」で指定します。

Concord ツールでは，テーブル上の検索結果を削除して編集し，必要な分だけ残して書き出したり，コピーしたりすることもできます。特定の検索結果の行を削除するには，テーブル上で削除したい行を選択して，Delete（Backspace）キーを押すか，コンテクストメニューから「選択したコンコーダンスラインを削除」を選択します（図 2.35）。一度目は警告が出ますが，問題がなければそのまま進めてください。この作業で結果の行を選択するごとにコンテクストビューにテキストを表示させたくない場合は右上の「コンテクスト」のチェックを外してください。

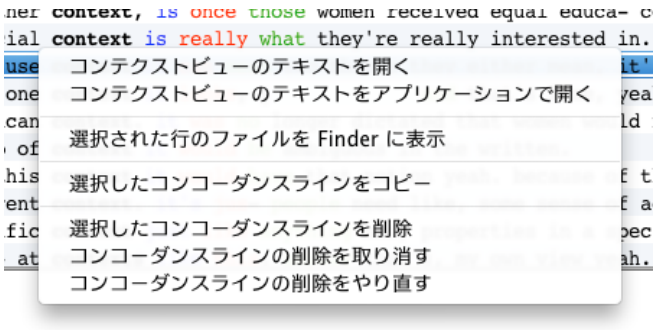


図 2.35 Concord コンテクストメニュー

選択した行をコピーするには，行を選択してから同様に右クリックでコンテクストメニューを表示して「選択したコンコーダンスラインをコピー」を選びます。「環境設定」の「Concord」にある「検索結果コピー時にスタイル情報を保持する」にチェックが入っていると，検索語の太字および文脈語の色や下線などの情報を保持したままリッチテキストとしてコピーされます。ただし，この情報は Microsoft Excel では保持されないので，Microsoft Word にいったんペーストしてから，再びコピーして Excel に貼付けてください。結果の書き出しでは，Concord ツールのみテーブル上で色がついている文脈語などのテキストのスタイル情報を保持したまま，リッチテキストファイルで保存することもできます。

この他に，Concord ツール特有の設定で「言語モード」の指定があります（図 2.36）。この設定は「環境設定」の「Concord」およびメインウインドウの下部で行い，デフォルトでは「西洋言語 A」になっていますが，英語のテキストを扱う場合は通常変更する必要はありません。ギリシア語など通常のアルファベット以外の文字が多く扱われる言語を扱う場合のみ「西洋言語 B」にしてください。西洋言語にモードが 2 つある理由は，Ruby と Objective-C で文字列の扱いが異なるためで，Objective-C では文字ごとに処理されるのに対し，Ruby ではバイトごとに文字列を処理するので，[4] UTF-8 で装飾文字などのマルチバイト文字を扱う際に文字の途中で切れて Objective-C 側に渡す際にエラーが出るのを防ぐためです。日本語を扱う場合は，元のファイルでテキストが分かち書きされているかどうかでモードを変更してください。

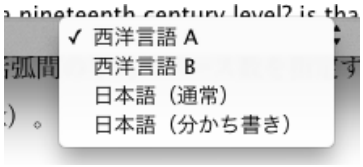


図 2.36 Concord 言語モード

この他にも，「環境設定」の「Concord」では並べ替えに使われた文脈語の色づけを変更したり，結果表示のフォントを変更するなど，様々な設定ができますが，教材を準備する際などに使える機能

図 2.37 Concord 検索語置換処理の設定

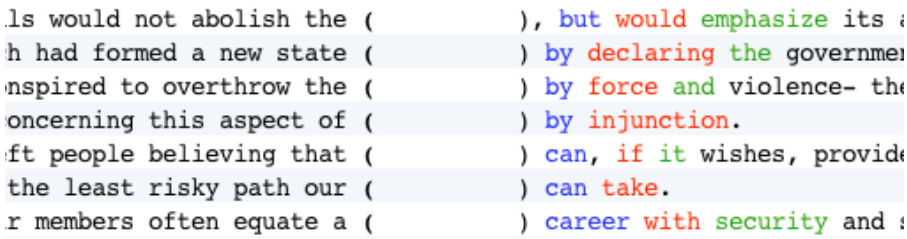


図 2.38 Concord 検索語置換処理の検索結果

として，「検索語の置換」処理があります（図 2.37）。これにチェックを入れて置換する括弧のタイプを選び，括弧間の半角スペース数を指定すると，検索結果の検索語部分が括弧で置き換わって表示されます（図 2.38）。

最後に，「環境設定」で文脈の範囲を「ファイル」にして，「コンコーダンスプロットを作成」にチェックを入れると，ツールタブの一番右に，「プロット」というタブが表示され，検索語がそれぞれのファイルの中でどこに現れるかを示すプロットが作成されます（図 2.39）。このプロットは，メインメニューの「ファイル」から「テーブルの結果を書き出す」を選んで PDF または JPEG ファイルとしてプロットごと，もしくは，まとめて保存できます。その際に，各プロット左上にあるチェックボックスにチェックを入れると，チェックを入れたプロットだけが書き出され，いずれにもチェックを入れないと，すべてが書き出されます。

図 2.39 コンコーダンスプロット

2.2.3. Cluster

Cluster ツール（図 2.40）では，上部のテキストボックスに検索したい単語を入力して検索ボタンをクリックし，検索語を含む 2 ~ 8 単語のクラスターリストを作ることができます。また，Concord ツールと Collocation ツールと共通の検索履歴も残せます。図 2.40 でわかるように，Cluster ツールには，左右 2 つのテーブルがあるので，それぞれで異なる語数のクラスターを検索して比べることができるほか，アドバンストコーパスモードでは，ファイルビューで左右異なるコーパス・データベースを指定して検索することで，コーパス・データベース間で同じ語数のクラスターリストを比較できます。

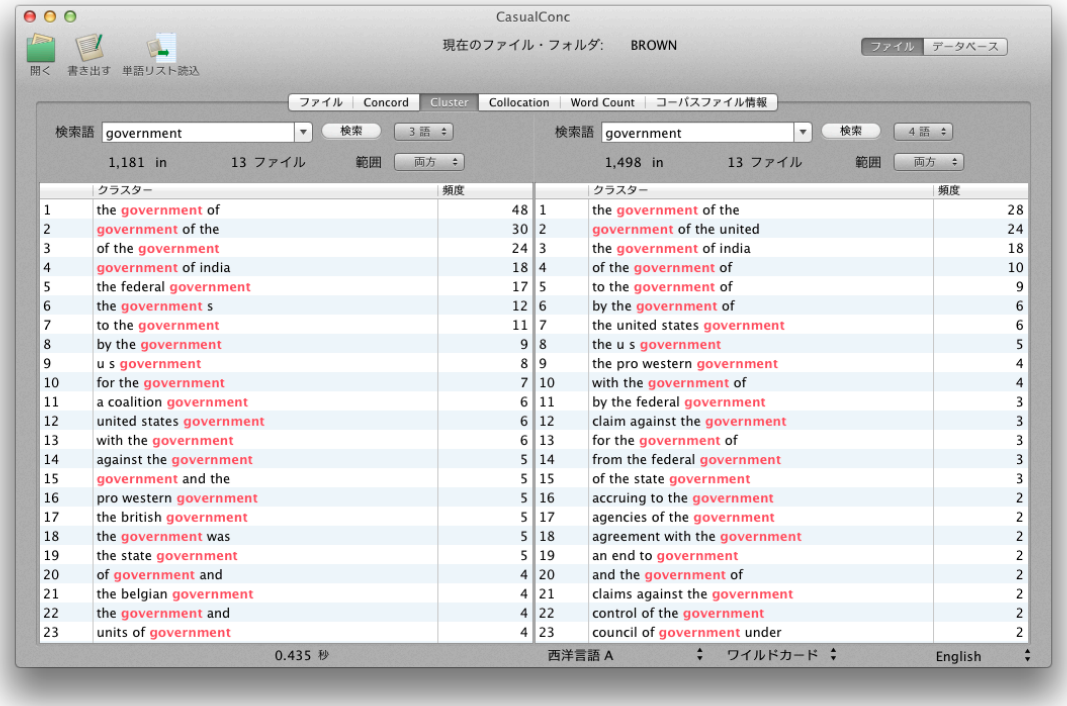


図 2.40 Cluster ツール

検索結果をテーブルで選択してから右クリックでコンテクストメニューを表示させて，選択されたクラスターを Concord ツールや Collocation ツールで検索することもできますが，その際は結果の頻度が異なる場合があります。これは，Cluster ツールでは単語間の記号は無視されますが Concord では無視されないためです。Concord で同じ条件で検索するには，前述の通り，「環境設定」の「一般」で一番下にある「検索語の間に記号がある場合も検索する」にチェックを入れてから検索してください。

Cluster ツールでも文脈語の範囲を指定できますが，Concord ツールと違い，「左のみ」を選択すると，検索語が一番最後に使われているクラスターのみが結果テーブルに表示され，「右のみ」を選択すると，検索語が一番最初に現れるクラスターのみが表示されます（図 2.41）。また，単語クラスター検索では，頻度が低いものは場合によってはあまり有用でないため，「環境設定」の「その他」で結果として表示する最低頻度を指定して，指定頻度未満のクラスターを結果から除外することができます。この他に，lemma 処理を適用した場合は，検索結果のクラスターに含まれるそれぞれの単語が lemma 処理され，頻度の右側に新たな列が生成されて，実際にテキスト中に現れたクラスターが頻度とともに表示されます（図 2.42）。

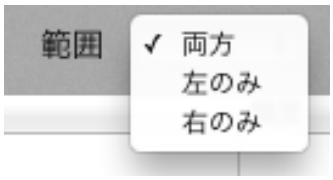


図 2.41 Cluster 範囲指定

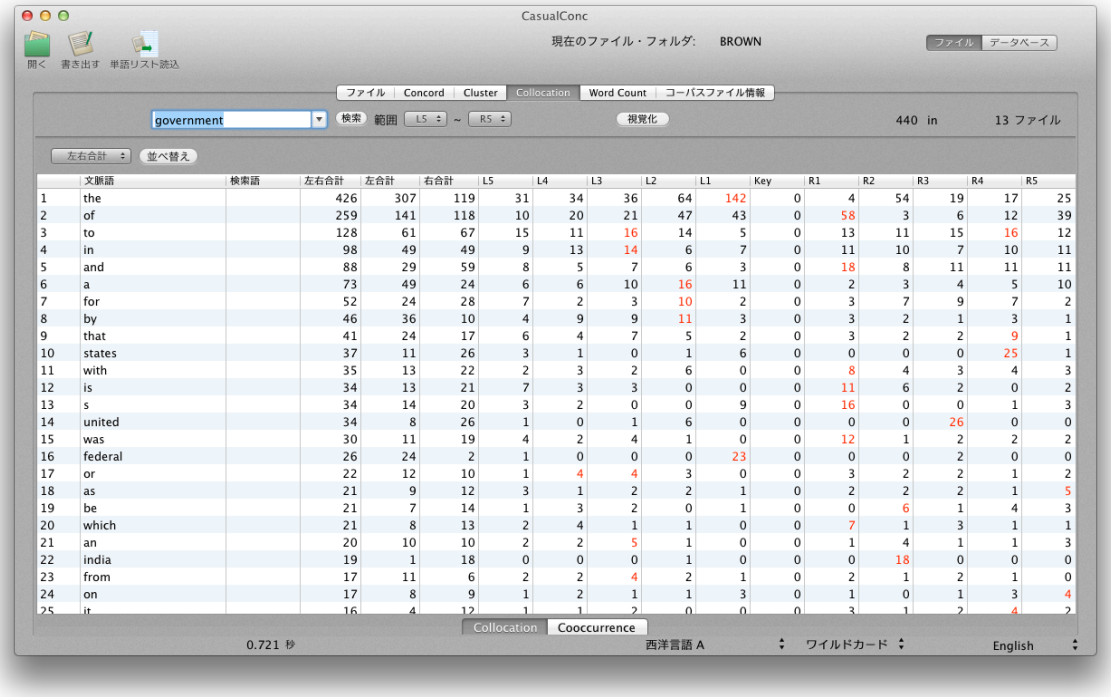
図 2.42 Cluster ツールでの lemma 処理

左テーブルの検索結果は，メインメニューの「ファイル」から「テーブルの結果を書き出す」を選び，右テーブルは「右側テーブルの...」を選んで書き出します。

2.2.4. Collocation/Cooccurrence

Collocation と Cooccurrence は独立したツールではなく，Collocation 検索時に Cooccurrence テーブルも同時に作成されます。

2.2.4.1. Collocation



	文脈語	検索語	左右合計	左合計	右合計	L5	L4	L3	L2	L1	Key	R1	R2	R3	R4	R5
1	the		426	307	119	31	34	36	64	142	0	4	54	19	17	25
2	of		259	141	118	10	20	21	47	43	0	58	3	6	12	39
3	to		128	61	67	15	11	16	14	5	0	13	11	15	16	12
4	in		98	49	49	9	13	14	6	7	0	11	10	7	10	11
5	and		88	29	59	8	5	7	6	3	0	18	8	11	11	11
6	a		73	49	24	6	6	10	16	11	0	2	3	4	5	10
7	for		52	24	28	7	2	3	10	2	0	3	7	9	7	2
8	by		46	36	10	4	9	9	11	3	0	3	2	1	3	1
9	that		41	24	17	6	4	7	5	2	0	3	2	2	9	1
10	states		37	11	26	3	1	0	1	6	0	0	0	0	25	1
11	with		35	13	22	2	3	2	6	0	0	8	4	3	4	3
12	is		34	13	21	7	3	3	0	0	0	11	6	2	0	2
13	s		34	14	20	3	2	0	0	9	0	16	0	0	1	3
14	united		34	8	26	1	0	1	6	0	0	0	0	26	0	0
15	was		30	11	19	4	2	4	1	0	0	12	1	2	2	2
16	federal		26	24	2	1	0	0	0	23	0	0	0	2	0	0
17	or		22	12	10	1	4	4	3	0	0	3	2	2	1	2
18	as		21	9	12	3	1	2	2	1	0	2	2	2	1	5
19	be		21	7	14	1	3	2	0	1	0	0	6	1	4	3
20	which		21	8	13	2	4	1	1	0	0	7	1	3	1	1
21	an		20	10	10	2	2	5	1	0	0	1	4	1	1	3
22	india		19	1	18	0	0	0	1	0	0	0	18	0	0	0
23	from		17	11	6	2	2	4	2	1	0	2	1	2	1	0
24	on		17	8	9	1	2	1	1	3	0	1	0	1	3	4
25	it		16	4	12	1	1	2	0	0	0	3	1	2	4	2

図 2.43 Collocation ツール

図 2.44 Collocation 並べ替えオプション

Collocation ツール（図 2.43）では，左上部のテキストボックスに検索したい単語を入力して検索し，検索語の前後の指定した範囲に現れる文脈語の頻度を集計したリストを作成します。検索履歴は Concord ツール，Cluster ツールと共通になっています。アドバンストコーパスモードでは，検索テキストボックスの左にコーパス・データベースを選択できるポップアップメニューが現れるので，検索ごとにコーパス・データベースを切り替えることもできます。表示される数値は L5 から R5 までの指定した範囲でそれぞれの位置に現れる単語の頻度で，数値が赤くなっているものは最頻値になります。また，結果表示は指定された範囲のものだけになります。例えば，範囲を L3 から L1 に指定すると，検索語の左側一語目から三語目までの位置に現れる単語の頻度が集計されて表示され，それ以外の範囲はテーブルから列ごと除外されます。検索結果を並べ替えるには，検索する前に並べ替えの基準とする列を一つ選ぶか，または，検索後に選んで「並べ替え」ボタンをクリックして行います（図 2.44）。

デフォルトの設定では，複数の語を検索したり，ワイルドカード文字を使って検索すると，それぞれの単語ごとに文脈語の頻度が集計されます。例えば，test/assessment で検索した場合には，test と assessment それぞれに，文訳語の頻度が集計されます。検索語すべてを一つの単語としてまとめて扱いたい場合は，「環境設定」の「その他」にある「キーワードを一つの単語として扱う」にチェックを入れてください。例えば，context/contexts という 2 つの単語を検索した場合，デフォルトでは，context と contexts それぞれで文脈語が集計されますが，一つの単語として扱えば，両方の単語に共通する文脈語の頻度は合計されて表示されます。

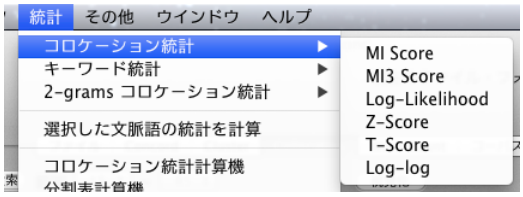
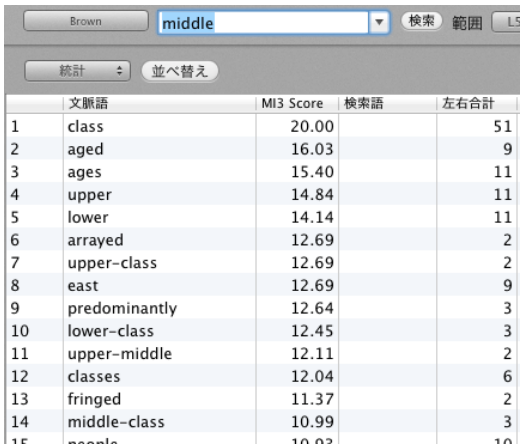


図 2.45 Collocation 統計メニュー



	文脈語	MI3 Score	検索語	左右合計
1	class	20.00		51
2	aged	16.03		9
3	ages	15.40		11
4	upper	14.84		11
5	lower	14.14		11
6	arrayed	12.69		2
7	upper-class	12.69		2
8	east	12.69		9
9	predominantly	12.64		3
10	lower-class	12.45		3
11	upper-middle	12.11		2
12	classes	12.04		6
13	fringed	11.37		2
14	middle-class	10.99		3

図 2.46 Collocation 統計

Collocation ツールには，2 つの特徴的な機能があります。一つは，Collocation 統計の計算です。後述の Word Count ツールの左テーブルで同じテキストの単語リストが作成してあれば，メインメニューの「統計」にある「コロケーション統計」から計算したい統計を選んで，文脈語ごとの Collocation 統計値を計算できます。計算できる Collocation 統計は，MI（mutual information），MI3，Log-Likelihood，Z-score，T-score，および Log-Log で（図 2.45，2.46），計算式は，主に BNCweb manual[5] にあるものを使っています。また，これとは別に，メインメニューの「統計」から「コロケーション統計計算機」もしくは「分割表計算機」を呼び出して，個々の文脈語のコロケーション統計の計算をすることもできます。

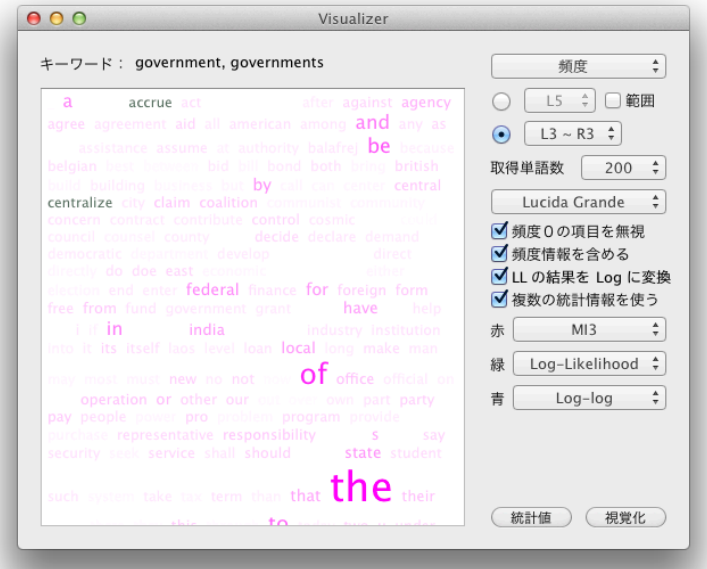


図 2.47 Collocation の視覚化

もう一つの特徴的な機能は，Collocation の視覚化処理です。この機能は実験的なものですが，Collocation 統計値を用いて，文字の大きさと色で Collocation の強さを視覚的に表示させます。この機能を使うには，単語リストを作成した後に，Collocation 検索をしてから「視覚化」ボタンをクリックして，Visualizer パネルを表示させます。デフォルトでは，統計値計算に使う単語が現れる範囲もしくは位置を指定しますが，「範囲」にクリックを入れると，検索語から指定した位置までの間に現れる頻度が利用されます。一番下の「複数の統計情報を使う」にチェックを入れる

と，赤，緑，青の三色それぞれに統計値を割り当てて，統計値の情報を色で表すこともできます。図 2.47 は 3 つの統計値と頻度情報を使って視覚化した例です。

これ以外にも，メインメニューの「その他」から，文脈語とその頻度を Word Count のテーブルにコピーしたり，コンテクストメニューから検索語と文脈語の組み合わせで Concord ツールで検索したりできます。この場合は，Concord での検索語は，Collocation で検索されたものが使われて，文脈語で絞り込んだ検索結果が表示されます。

2.2.4.2. Cooccurrence

L5	L4	L3	L2	L1	Key	R1	R2	R3	R4	R5
the	the	the	the	the	government	of	the	united	states	of
to	of	of	of	of		and	india	the	the	the
of	in	to	a	federal		s	to	to	to	to
in	to	in	to	a		to	in	and	of	and
and	by	a	by	s		was	and	for	and	in
for	a	by	for	state		in	for	in	in	a
is	and	and	u	coalition		is	be	of	that	as
a	at	that	and	in		with	is	will	for	on
that	government	an	in	local		which	were	a	a	an
by	or	from	united	british		agencies	would	its	be	be
their	that	or	with	foreign		are	an	people	government	had
was	which	was	against	states		has	its	were	it	have
2	be	claim	pro	central		itself	not	which	with	s
against	department	claims	that	our		the	office	with	assistance	with
as	i	control	units	to		will	with	agrees	by	all
s	is	does	agencies	western		assumed	a	all	on	at
states	were	for	council	belgian		at	have	as	use	d
agreement	with	is	form	and		by	made	berlin	25	for
an	all	operation	new	any		concerned	most	end	affairs	has
are	an	accruing	or	by		decided	no	entered	are	him
been	but	against	when	city		for	of	explained	corps	into
but	efforts	agencies	agency	cosmic		it	american	federal	had	is
figured	for	agreement	all	new		might	as	from	into	it
from	form	as	as	on		offices	buildings	i	its	its
has	from	be	end	other		or	by	industry	no	or

図 2.48 Cooccurrence ツール

Collocation 検索を実行すると，Cooccurrence テーブルには指定した範囲の各位置で現れる単語のリストが作成されます（図 2.48）。Cooccurrence テーブルを表示するには，ウインドウ下の「Cooccurrence」タブをクリックして切り替えます。デフォルトではそれぞれの位置に現れる単語が頻度順に並んでいますが，Word Count で単語リストを作成すれば，Collocation 統計値で並べ替えることもできます。並べ替えに使える統計値は，前述の Collocation の項で示した 6 種類です。

2.2.5. Word Count

Word Count ツール（図 2.49）では，単語・n-gram リストを作ることができます。n-gram は分野によって構成単位が異なりますが，ここでは，「環境設定」で指定した「文脈の範囲」を区切りとした文字列に現れる n 個の連続した「単語」になっています。この「単語」に何を含めるかは，前述の「単語の定義」にある設定で指定します。

Word Count の基本的な使い方は簡単で，左右どちらのテーブルでも，左上のポップアップメニューから「単語」もしくは「n-gram」（n = 2 ~ 5）を選択して「実行」をクリックするだけです。処理速

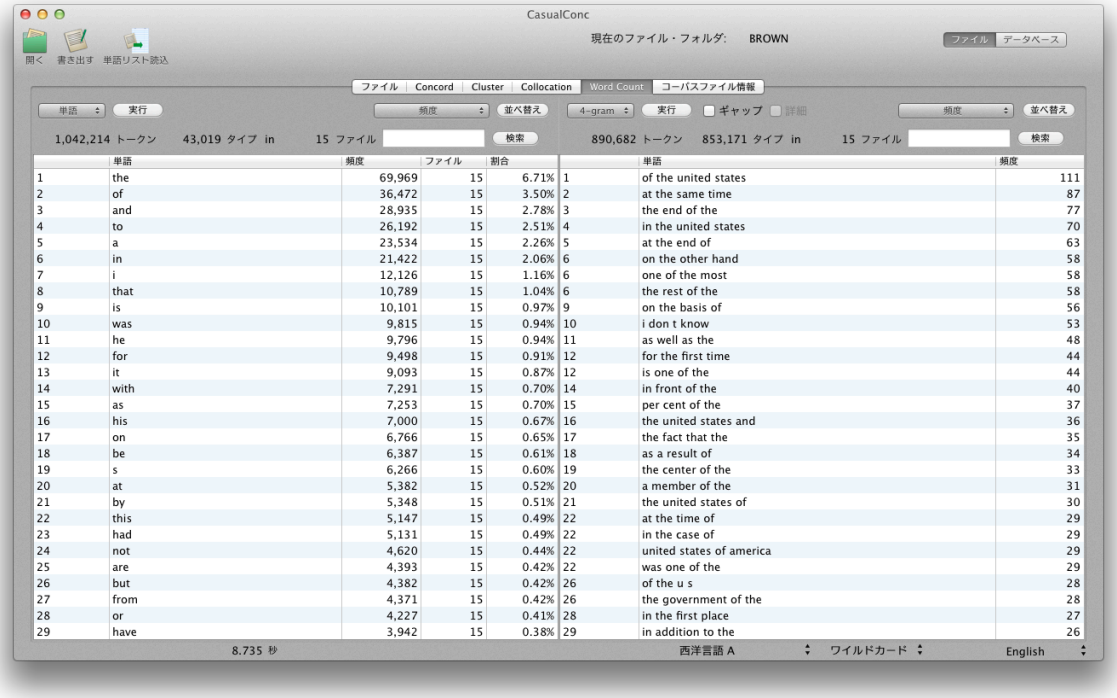


図 2.49 Word Count ツール

度は，お使いの Mac の CPU スピードと搭載メモリの量によって変わりますが，コーパス総単語数が多いほど，また，n-gram の n の値が大きくなるほど時間がかかります。コーパスサイズが 1000 万語を超えるような場合には，n-gram リストの作成にかなりの時間がかかるかもしれません。

結果の数値で，「トークン」は総語数（延べ語数），「タイプ」は異なり語数になり，「ファイル」の列はその単語・n-gram が現れるファイルの数，「割合」は総語数に対するその単語の頻度の割合になります。並べ替えは，リスト作成前，作成後のどちらでもできますが，リスト作成後の場合は，「並べ替え」ボタンをクリックして並べ替えてください。並べ替え順の詳細は表 2.7 に説明してあります。

テーブル上部右端下段のテキストボックスはテーブル上の単語・n-gram 検索用で，単語などを入力してからその右の「検索」ボタンをクリックすると，その単語がリスト上で見つかればその行が強調されて表示されます。この検索には，ワイルドカードモードであればワイルドカード文字が，正規表現

表 2.7
Word Count ツールの並べ替え順

ラベル	機能
頻度	頻度順
アルファベット順	アルファベット順
逆アルファベット順	単語の最後の文字からアルファベット順
単語の長さ	単語の長さが一番長いものが一番上に
単語の長さ（逆順）	単語の長さが一番短いものが一番上に
統計	Keyness 統計値の値が大きいものが上に
統計（逆順）	Keyness 統計値の値が小さいもの（負の大きいもの）が上に

モードであれば正規表現が使え，マッチするものが複数あれば，再び「検索」ボタンをクリックすることで，次にマッチする単語・n-gram が表示されます。

単語リストを作成する際に数値の扱いが問題になることもありますが，CasualConc では，「そのまま」扱う，「# として数える」，「無視」するという 3 つから選択ができ，「環境設定」の「一般」にある「数値の処理」で設定できます。「そのまま」では，すべての数値が別個の単語として頻度集計され，「# として数える」では，単独の数値（他の文字と組合わさっていないもの）もしくは，数値で始まる文字列の数値を # で置き換えて集計します。「無視」を選択すると，数値は無視されて単語リストに含まれません。また，大きなコーパスで頻度が少ない単語や特に n-gram をテーブル上ですべて扱うと処理が非常に重くなるので，「環境設定」の「その他」でテーブルに表示する最低頻度を設定できます。n-gram の場合は，1 回のみ現れるものは重要度が低いので，デフォルトで削除されます。

図 2.50 キーワード統計

Word Count ツールでも Collocation ツールと同じく統計値が計算できます。Word Count ツールで計算できる統計値は Keyness という 2 つの単語リストを比べたときに，そこで使われる単語の特異性・特徴性を示す値で，右側に比較するリファレンスコーパスの単語リスト，左側に Keyness を計算したいコーパスの単語リストを表示させた状態で計算します。右側のリストは，CasualConc で作成したものでも，後述するようにテキストファイルとして用意されている単語リストを読み込ませたものでもかまいません。計算方法は，2 つの対比させたい単語リストを左右のテーブルに表示した後に，メインメニューの「統計」から「キーワード統計」にある「Log-Likelihood」または「Chi-Square」を選択します（図 2.50）。問題がなければ，左側のテーブルに統計の列が表示され，そこに計算された統計値が表示されます。統計値のうち黒い数値は左側のテーブルのリストに特徴的な単語で，値が大きいほどその度合いが大きくなります。赤い数値（負）は右側のテーブルのリストに特徴的な単語になります。図 2.51 の例は，BNC（British National Corpus）の単語リストにをリファレンスコーパスにして，Brown コーパスの Keyness を計算したもので，統計値で並べ替えてあります。この Keyness の

	単語	頻度	ファイル	割合	Log-Likeli
1	the	69,969	15	6.78%	636202.41
2	he	9,796	15	0.95%	88970.779
3	be	6,387	15	0.62%	58022.727
4	are	4,393	15	0.43%	39925.181
5	s	6,288	15	0.61%	27004.203
6	me	1,186	15	0.11%	10767.726
7	_	1,044	9	0.10%	9492.709
8	before	1,016	15	0.10%	9222.288
9	t	2,253	15	0.22%	8368.943
10	american	599	14	0.06%	5431.716
11	general	509	13	0.05%	4613.705
12	every	492	15	0.05%	4459.198
13	president	432	13	0.04%	3882.609
14	present	403	15	0.04%	3650.352
15	f	1,182	12	0.11%	3522.475

図 2.51 Keyness 計算

	単語 - Brown	頻度	ファイル	割合	Log-Likeli
1	before	1,016	15	0.10%	9234.389
2	american	599	14	0.06%	5438.850
3	general	509	13	0.05%	4619.767
4	every	492	15	0.05%	4465.057
5	president	432	13	0.04%	3887.753
6	present	403	15	0.04%	3655.152
7	f	1,182	12	0.12%	3535.346
8	didn	403	14	0.04%	3514.581
9	re	345	15	0.03%	3106.635
10	experience	276	15	0.03%	2489.173
11	don	513	14	0.05%	2265.690
12	federal	250	11	0.02%	2263.110

図 2.52 Keyness 計算（Stop words 処理適用）

計算などで，頻出する機能語などを除きたい場合は，前述（2.1.1.4）の Stop words 処理をして単語リストを作成してください（図 2.52）。

Word Count ツールでは，前述のように既存の単語リストを読み込むこともできます。まずは，メインウインドウ左上の「単語リスト読込」（図 2.53）をクリックするか，メインメニューの「ファイル」から「単語リストを読み込む」を選択します。読み込める単語リストの形式は，CasualConc から書き出した形式のものか，単語と頻度のみがカンマ（,）もしくはタブで区切られた形式（タイトル行の有無はどちらでも）のプレインテキストファイルだけになっています。[6] 読み込む際は，区切りの種類と，左右どちらのテーブルに読み込むかの指定を適切に行ってください。もし左右のテーブルを間違えて読み込んでしまった場合は，メインメニューの「その他」から「リストを移動」もしくは「リストをコピー」を選んで，単語リストをテーブル間で移動・コピーすることができます。

図 2.53 単語リスト読込アイコン

Lemma 処理や異綴り処理を Word Count ツールで適用すると，lemma のリストが作成され，各 lemma に含まれる単語が頻度とともに右端の列に表示されます。これで，それぞれの lemma においてどの単語が何回現れるかを確認できます（図 2.54）。Lemma 処理や異綴り処理，それに Stop words 処理は，Word Count で作成した単語リストのみならず，読み込んだリストに対しても適用す

	単語 - Brown	頻度	ファイル	割合	Lemmaed Words
1	the	69,969	15	6.82%	
2	be	37,935	15	3.70%	is (10101), was (9815), be (63
3	of	36,472	15	3.56%	
4	and	28,935	15	2.82%	
5	a	27,286	15	2.66%	a (23536), an (3750)
6	to	26,192	15	2.55%	
7	in	21,422	15	2.09%	
8	have	12,677	15	1.24%	had (5131), have (3942), has (
9	that	11,639	15	1.13%	that (10789), those (850)
10	he	9,796	15	0.96%	
11	for	9,498	15	0.93%	
12	it	9,093	15	0.89%	
13	with	7,291	15	0.71%	
14	as	7,253	15	0.71%	
15	his	7,000	15	0.68%	

図 2.54 Lemma 処理適用済み単語リスト

ることができます。読み込んだリストに適用するには，環境設定などで適用したい処理が使えるようになっていることを確認してから，メインメニューの「その他」の下の方にある「Lemma/異綴りを適用」もしくは「Stop Words を適用」を選んで適用してください。

この他にも，Word Count ツールには，用例検索などで有用となり得る少し特殊な機能がついています。まずは，n-gram の単語の一つが * に置き換わったリストを作る gapped n-gram です。リスト作成オプションのポップアップメニューで n-gram を選ぶと，「実行」ボタンの右側に「ギャップ」チェックボックスが現れるので，チェックを入れて「実行」をクリックすると gapped n-gram のリストが作成されます。さらに，「詳細」にチェックを入れて実行すると，* の位置に現れる単語が頻度とともに別の列で表示されます。ただし，この処理には多くのメモリと CPU パワーが必要となりますので，特に大きなコーパスで使う際は気をつけてください。

また，特定の文字列の頻度を集計することもできます。Concord ツールでは，検索語にマッチした文字列すべての頻度が集計され，Collocation ツールでも Key のところで検索語にマッチした文字列ごとの頻度が集計されるのですが，文脈語の情報が必要ない場合は，Word Count ツールでマッチした文字列だけのリストを手早く作れます。使い方は，「環境設定」の「その他」で「指定文字列検索モード」にチェックを入れて Word Count ツールで「単語」を選択すると「実行」ボタンの右側にテキストボックスが現れるので，そこに検索したい文字列を入力して「実行」をクリックするだけです。これで，検索文字列にマッチした文字列のリストを作ることができます。ワイルドカードモードでは，ワイルドカード文字が，正規表現モードでは正規表現が使えます。ただし，ワイルドカードモードで「文字列全体を検索」にチェックが入っていないときに (A|B) を使う場合は，最初の括弧の後に ?: を入れて，(?:A|B) としてください。これは，() で囲まれた文字列がある場合は，その場所に現れる単語のみでリストが作られるためです。これを応用して，() 内にワイルドカード文字を使うことで，共起する単語だけのリストが作成できます。例えば，Brown コーパスを使って it (?:is|was) (?) (*) (*) that を検索したところ，図 2.55 に示す結果となりました。この他にも，?ly を検索すると，ly で終わる単語のリストが作れます。

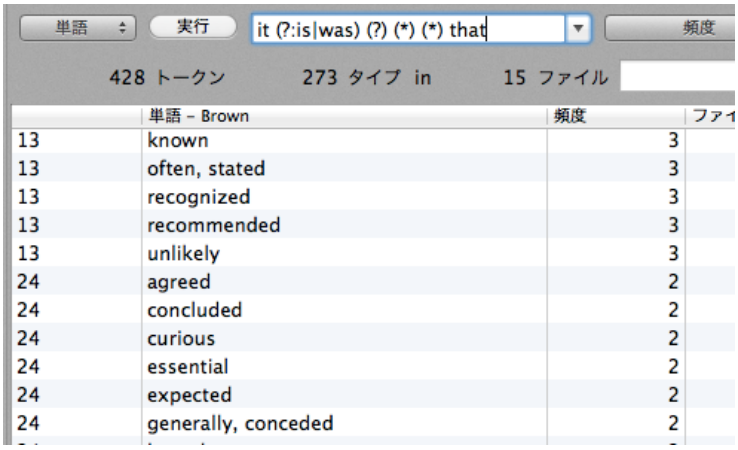
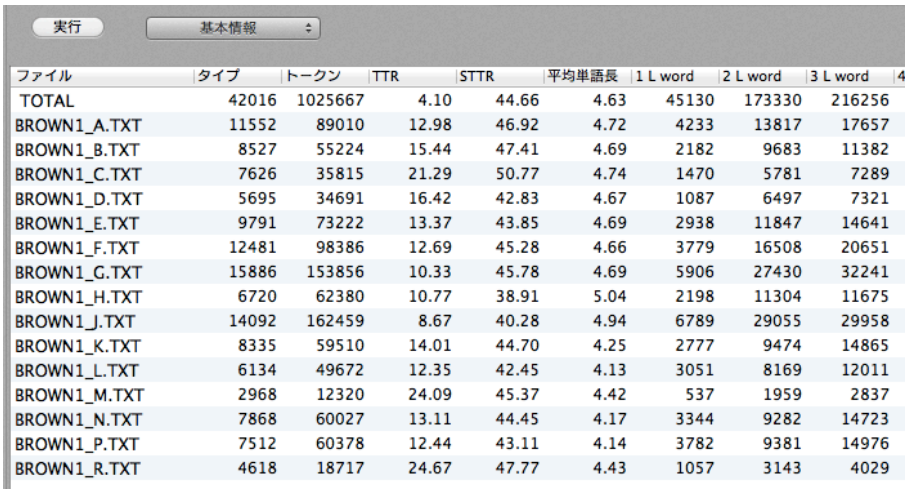


図 2.55 Word Count 指定文字列検索

最後に，それほど頻繁に使われないとは思いますが，「環境設定」の「その他」で「アドバンストモード」にチェックを入れると，最長 10-gram までのリストを作ることができるようになります。ただし，n の値が大きくなればなるほど，時間とメモリの消費量が大きくなるので気をつけてください。

2.2.6. コーパスファイル情報

コーパスファイル情報ツールは，コーパスのファイルに含まれる単語の頻度などの基本的な情報をファイルごとにまとめたり，その他，指定した単語などの頻度をファイルごとやコーパス・データベースファイルごとに集計する機能をまとめたものです。実験的な機能も含んでいて，本稿執筆時（2012 年 3 月末）の仕様から大きく変更することもありますので，気をつけてください。

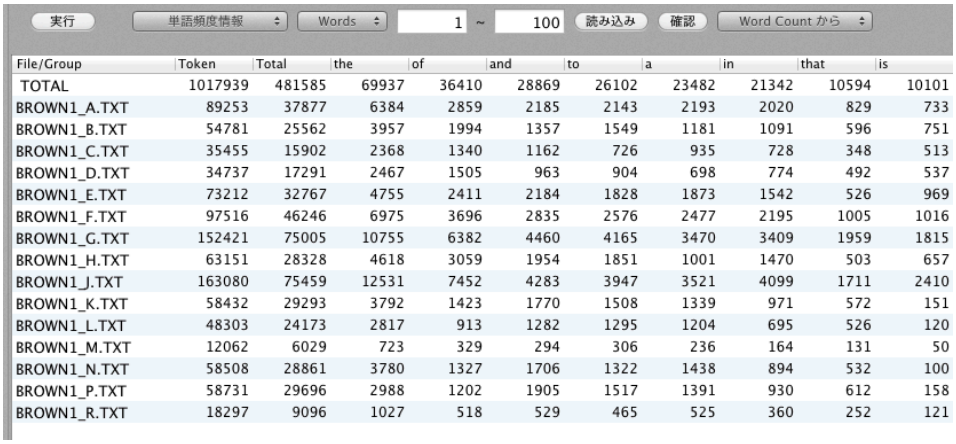

実行　基本情報

ファイル	タイプ	トークン	TTR	STTR	平均単語長	1 L word	2 L word	3 L word	4
TOTAL	42016	1025667	4.10	44.66	4.63	45130	173330	216256	
BROWN1_A.TXT	11552	89010	12.98	46.92	4.72	4233	13817	17657	
BROWN1_B.TXT	8527	55224	15.44	47.41	4.69	2182	9683	11382	
BROWN1_C.TXT	7626	35815	21.29	50.77	4.74	1470	5781	7289	
BROWN1_D.TXT	5695	34691	16.42	42.83	4.67	1087	6497	7321	
BROWN1_E.TXT	9791	73222	13.37	43.85	4.69	2938	11847	14641	
BROWN1_F.TXT	12481	98386	12.69	45.28	4.66	3779	16508	20651	
BROWN1_G.TXT	15886	153856	10.33	45.78	4.69	5906	27430	32241	
BROWN1_H.TXT	6720	62380	10.77	38.91	5.04	2198	11304	11675	
BROWN1_J.TXT	14092	162459	8.67	40.28	4.94	6789	29055	29958	
BROWN1_K.TXT	8335	59510	14.01	44.70	4.25	2777	9474	14865	
BROWN1_L.TXT	6134	49672	12.35	42.45	4.13	3051	8169	12011	
BROWN1_M.TXT	2968	12320	24.09	45.37	4.42	537	1959	2837	
BROWN1_N.TXT	7868	60027	13.11	44.45	4.17	3344	9282	14723	
BROWN1_P.TXT	7512	60378	12.44	43.11	4.14	3782	9381	14976	
BROWN1_R.TXT	4618	18717	24.67	47.77	4.43	1057	3143	4029	

図 2.56 ファイル情報の基本情報

「基本情報」（図　2.56）では，各ファイルに含まれる単語の総タイプ（異なり語），総トークン（延べ語），TTR（Type-Toke ratio），STTR（Standardized TTR: 1000 単語ごとの TTR の算術平均値），平均単語帳（単語に含まれる文字数の算術平均値），L1 ~ L15 word（文字数 1 から 15 までの単語の頻度）が集計されます。STTR の値で * が付いているものは，トークン数が 1000 単語に満たなかったファイルで，単純に TTR の値が計算されたものとなっています。文字長ごとの単語の頻度は，「環境設定」の「その他」にある「頻度情報」でタイプで数えるかトークンで数えるかを選べます。

「単語頻度情報」（図　2.57）は，各ファイルに含まれる単語数をファイルごとに集計するための機能です。まずは準備段階として集計する単語のリストを作ります。リストの作成は，右端にあるポップアップメニューから「Word　Count　から」を選んで「読み込み」ボタンをクリックし，Word　Count


実行　単語頻度情報　Words　1 ~ 100　読み込み　確認　Word Count から

File/Group	Token	Total	the	of	and	to	a	in	that	is
TOTAL	1017939	481585	69937	36410	28869	26102	23482	21342	10594	10101
BROWN1_A.TXT	89253	37877	6384	2859	2185	2143	2193	2020	829	733
BROWN1_B.TXT	54781	25562	3957	1994	1357	1549	1181	1091	596	751
BROWN1_C.TXT	35455	15902	2368	1340	1162	726	935	728	348	513
BROWN1_D.TXT	34737	17291	2467	1505	963	904	698	774	492	537
BROWN1_E.TXT	73212	32767	4755	2411	2184	1828	1873	1542	526	969
BROWN1_F.TXT	97516	46246	6975	3696	2835	2576	2477	2195	1005	1016
BROWN1_G.TXT	152421	75005	10755	6382	4460	4165	3470	3409	1959	1815
BROWN1_H.TXT	63151	28328	4618	3059	1954	1851	1001	1470	503	657
BROWN1_J.TXT	163080	75459	12531	7452	4283	3947	3521	4099	1711	2410
BROWN1_K.TXT	58432	29293	3792	1423	1770	1508	1339	971	572	151
BROWN1_L.TXT	48303	24173	2817	913	1282	1295	1204	695	526	120
BROWN1_M.TXT	12062	6029	723	329	294	306	236	164	131	50
BROWN1_N.TXT	58508	28861	3780	1327	1706	1322	1438	894	532	100
BROWN1_P.TXT	58731	29696	2988	1202	1905	1517	1391	930	612	158
BROWN1_R.TXT	18297	9096	1027	518	529	465	525	360	252	121

図 2.57 単語頻度情報

ツールの左テーブルに単語リストを取り込むか，「ファイルから」を選んで，一行一単語のプレインテキストファイルを読み込むか，「単語取り込みパネル」を選んで開いたパネルに一行一単語で入力して読み込みます（図 2.58）。読み込んだ単語リストは，「確認」ボタンをクリックして現れる単語リストパネルで確認・編集したら，2 つのテキストボックスに読み込んだ単語のうち何番目から何番目の単語を使うかを指定してから，「実行」をクリックしてリストを作成します。デフォルトでは，Word Count のリストから読み込む単語数が 300 に制限されているので，それを超える単語を扱いたい場合は，環境設定の「ファイル情報」で制限なしにするなり，制限値を変更してください。この制限は，不用意に大きなコーパスでこの集計表を作ると処理や表示に多くの時間とメモリーが必要となるため，それを防ぐ目的で設定しています。この他にも，「環境設定」では，頻度を標準化したり，相対頻度を求めたり，ファイルごとにそのファイル中の単語頻度順で並べ替える設定ができます。図 2.57 は Brown コーパスのファイルごとの単語頻度リストを作成したものです。この他にも，n-gram でも同様のリストを作成することができますが，その際は，Word Count ツールであらかじめ同じ語数の n-gram リストを作成してから読み込んで処理してください。

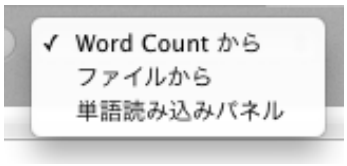


図 2.58 単語読み込み元

「TF-IDF」（図 2.59）は，コーパス全体に対してそれぞれのファイルに含まれる単語がどれくらい特徴的なのかを示す指標の一つで，CasualConc では簡単に計算することができます。「TF-IDF」でも，「単語頻度情報」と同様に，単語リストを読み込みこんでから処理しますが，その特性上，「環境設定」の「コーパス情報」にある「単語の読み込みを制限する」のチェックを外してから，Word Count で作成した単語リストの単語をすべて読み込んで実行してください。また，Word Count ツールで最低頻度が設定してある場合は，その設定が妥当かどうかを判断してリストを作成してから読み込んでください。また，単語頻度情報と同様に，n-gram でも TF-IDF リストを作成できますが，使用するコーパスが大きい場合は処理時間とメモリを大量に消費するので気をつけてください。結果の表示は，図 2.58 にあるように，ファイルごとに値の大きい順に並べ替えるか，単語ごとに値を合計してその合計値の大きい順にすべてのファイルで並べ替えるかを選べます。

ファイル	トークン	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
BROWN1_A.TXT	89010	maris (108.32)	mantle (106.79)	kennedy (59.45)	dallas (57.13)	palmer (52.73)
BROWN1_B.TXT	55224	podger (62.29)	_ (50.57)	khrushchev (42.15)	trujillo (40.62)	kennedy (38.11)
BROWN1_C.TXT	35815	comedie (43.33)	sansom (32.50)	larkin (29.79)	sloan (28.97)	cholesterol (25.75)
BROWN1_D.TXT	34691	shu (42.31)	irenaeus (40.62)	yang (37.91)	zen (37.02)	realtors (36.27)
BROWN1_E.TXT	73222	hanover (83.27)	prokofieff (70.52)	juniors (60.45)	_ (59.26)	feed (57.20)
BROWN1_F.TXT	98386	palfrey (94.78)	banion (92.07)	tsunami (59.58)	bridget (59.58)	borden (54.16)
BROWN1_G.TXT	153856	hearst (148.94)	gorton (94.78)	steele (89.37)	miriam (86.66)	woodruff (86.66)
BROWN1_H.TXT	62380	_ (200.24)	fiscal (109.86)	skywave (86.66)	allotment (74.55)	sba (73.12)
BROWN1_J.TXT	162459	f (224.93)	anode (208.52)	_ (118.51)	mg (105.61)	gyro (100.20)
BROWN1_K.TXT	59510	kate (88.66)	scotty (82.61)	alex (57.94)	ada (56.87)	andrei (54.16)
BROWN1_L.TXT	49672	alec (81.24)	shayne (78.53)	killpath (67.70)	hoag (62.29)	maude (59.58)
BROWN1_M.TXT	12320	helva (92.07)	ekstrohm (73.12)	dikkat (59.58)	mercer (45.06)	jubal (32.50)
BROWN1_N.TXT	60027	jess (127.28)	matsuo (100.20)	ramey (83.95)	brannon (81.24)	curt (78.58)
BROWN1_P.TXT	60378	cady (73.12)	linda (66.49)	bobbie (64.99)	theresa (64.99)	deegan (64.99)
BROWN1_R.TXT	18717	barco (70.41)	letch (56.87)	welch (54.16)	arlene (50.37)	moreland (46.04)

図 2.59 TF-IDF

「単語グループ頻度表」は，アドバンストコーパスモードのみで利用できる機能で，「単語頻度表」と同様の単語リストも作成できますが，大きな単語リストすべてを対象にするのではなく，特定の単語頻度を集計したり，単語をいくつかまとめたグループとして頻度集計したり，ファイルだけでなくコーパス・データベースファイルごとの頻度集計もできるようになっています。ただし，処理過程の制約により，複合語の頻度集計はできますが，厳密な意味での n-gram の頻度集計はできなくなっています。これは，「単語頻度表」では，すべての 単語や n-gram の頻度をファイルごとに集計した上で，それを基にして指定した単語もしくは n-gram の頻度を抜き出す形でリストを作成するのに対して，「単語グループ頻度表」では，指定した単語や n-gram だけを検索して頻度集計するので，頻度を数える単語やフレーズの組み合わせによっては，集計漏れが出る可能性が否定できないためです。ただ，この違いによって処理速度は大幅に向上しています。

図 2.60 単語グループ頻度表の集計単位

使い方は，まず，集計する単位を「ファイル」，「コーパス/データベース」，「混合」から選択します(図 2.60)。「ファイル」は，ファイルビューのコーパス・データベーステーブルでチェックが入っているコーパス・データベースに含まれているファイルごとに頻度集計をします。「コーパス/データベース」では，コーパス・データベースごとに頻度集計されます。「混合」を選択した場合は「設定」ボタンをクリックして現れたパネルで，それぞれのコーパス・データベースでファイルごとに集計するか，コーパス・データベースごとに集計するかを指定してから集計します。また，この選択パネルでは，結果表示のテーブルのそれぞれの集計単位にグループ名を付けることもできます。デフォルトでは，それぞれのコーパス・データベース名またはファイル名がグループ名になります。

次に，「単語頻度表」と同じく，「読み込み」ボタンを押して読み込みパネルを表示し，集計したい単語を group name->word1,word2,word3,... のように，単語グループのラベルとそれに含む単語をカンマ（,）でつないだものを -> で結んだ形式で入力していきます（図 2.61）。この時，いくつかの単語やフレーズをまとめたグループでなく一単語だけを一つのグループとして扱いたい場合は，その単語を入力して改行すれば内部的に一つのグループとして処理されます。また，単語だけのリストを読み込んだ際に，lemma や異綴り処理ができるように設定してあれば，「確認」ボタンをクリックすると現れる単語リスト確認パネル上で，lemma/異綴り処理を適用できます。図 2.62 は，Brown コーパスをそれぞれのファイルごとで，TIME という名前をつけたコーパスをコーパス全体として集計した結果です。

図 2.61 単語グループ頻度表読み込み

File/Group	TOTAL	government	citizen	president
TOTAL	231.38	94.63	15.81	120.95
BROWN1_A.TXT	291.31	95.23	13.44	182.63
BROWN1_B.TXT	268.34	120.48	32.86	115.00
BROWN1_C.TXT	42.31	19.74	5.64	16.92
BROWN1_D.TXT	31.67	17.27	5.76	8.64
BROWN1_E.TXT	46.44	15.02	9.56	21.85
BROWN1_F.TXT	57.43	20.51	12.31	24.61
BROWN1_G.TXT	95.79	34.12	19.68	41.99
BROWN1_H.TXT	337.29	237.53	15.84	83.93
BROWN1_J.TXT	72.97	52.73	9.81	10.42
BROWN1_K.TXT	27.38	6.85	1.71	18.83
BROWN1_L.TXT	6.21	.00	6.21	.00
BROWN1_M.TXT	49.74	16.58	24.87	8.29
BROWN1_N.TXT	3.42	.00	3.42	.00
BROWN1_P.TXT	34.05	17.03	.00	17.03
BROWN1_R.TXT	27.33	10.93	10.93	5.47
TIME	241.50	98.22	16.12	127.16

図 2.62 単語グループ頻度表結果

2.2.7. CasualConc Viewer

CasualConc Viewer は，CasualConc が依存する RubyCocoa のテーブル表示のバグに起因するクラッシュに対処するために制作したアプリケーションです。本稿執筆時のβ版では，この問題に対する新たな対処策を講じているため以前よりも安定はしていますが，完全に問題が解決していないかもしれないため同梱しています。また，Viewer を使うことによって，Collocation/Cooccurrence とコーパスファイル情報ツール，および，Cluster と Word Count ツールでも 2 つまたは 3 つ以上のテーブルを比較することができるようになります。

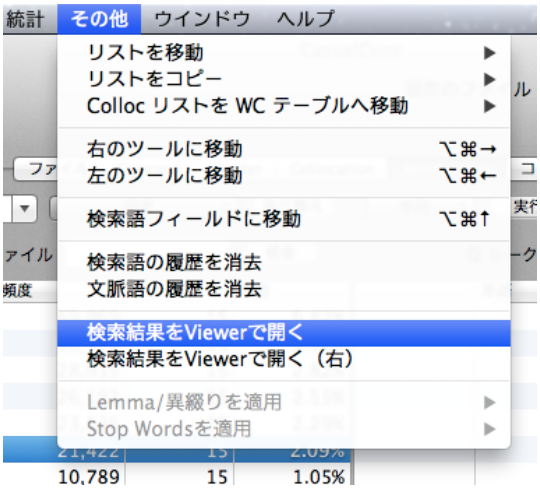


図 2.63 Viewer で開く

CasualConc Viewer は CasualConc beta のディスクイメージに入っていますので，アプリケーションフォルダに CasualConc と一緒にコピーしてください。それ以外に，特に作業は必要ありません。

CasualConc Viewer では，Concord，Cluster，Collocation，Word Count，コーパスファイル情報のいずれのツールのテーブル上の結果も開けます。まずは，これらのツールでリストを作成した後，メインメニューの「その他」から「検索結果を Viewer で開く」を選択します（図 2.63）。Cluster ツールと Word Count ツールで右テーブルの結果を開く場合は，（右）と書いてある方を選んでください。これで，CasualConc Viewer が立ち上がり，その時点で選択されているツールのテーブル上の結果が，Viewer で開きます。

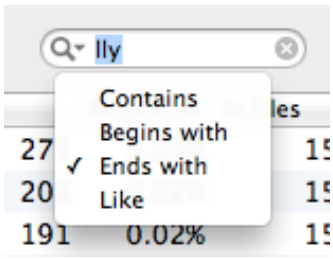


図 2.64 Viewer の検索窓

基本的には，見るだけのアプリケーションなので，結果の並べ替え程度の機能しか付いていませんが，テーブルでのデータの扱いに変更が加えてあるので，Collocation と Word Count では，検索窓で OS X 標準の機能である，指定した文字列を含むもの，その文字列で始まるもの，終

	Word Count – 1025667 tokens / 42016 types in 15 files	Freq	Propotion
1	really	275	0.03%
2	usually	206	0.02%
3	finally	191	0.02%
4	actually	166	0.02%
5	especially	162	0.02%
6	generally	132	0.01%
7	carefully	87	
8	fully	80	
9	naturally	70	
10	equally	62	
11	practically	53	
12	eventually	52	
13	gradually	50	
14	occasionally	48	
15	essentially	47	

図 2.65 Viewer での絞り込み検索

わるもの，一致するもの検索できます（図 2.64）。検索結果は，CasualConc の Word Count の単語検索のようにマッチした単語に移動して表示するのではなく，絞り込み検索になります（図 2.65）。これ以外には，テーブルの結果を保存して，後に Viewer で開くことができる程度の機能しか付いていませんが，簡単な編集などができるようにしようとは考えています。

以上で，一通り基本的な機能を紹介しました。ここから先は，論文用例検索ツールとしての使用例を，テキストデータを集めるところから簡単に示します。

3. 用例検索ツールとしての使用例

近年，研究者が望む望まないに関わらず，英語で書かれた論文を読むだけでなく，英語で論文を書く必要に迫られる状況になってきました。ただ，Biber et al. (1999) などが明らかにしたように，アカデミックな文章には日常会話やフィクションなどと明らかに異なる特徴が見られ，さらに，研究分野ごとに特徴的な表現や語彙が多く使われています。このことから，分野や研究対象に特化した論文のコーパスを作成し利用できれば，論文を書く際に大きな助けになります。

そこで，このセクションでは，論文テキストの入手から始めて，テキストの整形，CasualConc の論文用例ツールとしての利用法などについて，CausalTextractor の使用法も絡めながら示していきます。

3.1 テキストデータの準備

まずは，論文テキスト集めてを CasualConc に用例テキストデータとして読み込むまでを説明します。

3.1.1. テキスト収集

以前なら紙の論文から論文の全文データベースを作成したければ，論文をコピーしてからスキャンしてコンピューターに取り込み，さらに OCR アプリケーションで文字認識させる必要がありましたが，2012 年現在，どの分野の英語学術論文も多くのものが電子化されていて，個人もしくは大学の図書館が購読していれば簡単に入手できるようになっていますので，ここでは，それを利用します。

電子化されている論文の多くは PDF フォーマットになっており，新しいものはテキストデータをそのまま PDF にしたものが多くなっているので，プロテクトがかかっていない限り簡単にテキストを抜き出すことができますが，2000 年以前の古い論文では，画像データの PDF になっているものが多くあります。このような画像 PDF のものは，以前は画像のみだったのですが，最近ではその多くが OCR 処理されてテキストが埋め込まれているので，このような PDF ファイルの場合も，比較的簡単にテキストデータを抜き出せるようになってきました。ただし，画像データにテキストが埋め込まれているものは，場合によっては認識の精度が低いものがありますので注意が必要です。また，画像データのみでテキストが埋め込まれていないものや，自分でスキャンして PDF にしたものは，市販の OCR アプリケーションを購入して文字認識させる必要があります。Mac で OCR 処理ができるものには，Abbyy Finereader Express，ReadiIris などの専用アプリケーションや Adobe Acrobat があります

が，[7] これらは Windows 用の OCR アプリケーションと違って，アプリケーション上でテキストの編集ができず，PDF にテキストデータを埋め込むか，他のファイル形式で書き出すことしかできません。

これ以外には，最近は出版社によっては，PDF だけでなく HTML で論文を読めるものがあり，Safari で HTML ソースとしてや Web アーカイブ形式で簡単に保存できます。PDF ファイルと比べると，HTML の場合は，既にテキストデータそのものを見る形で用意されているので，文字化けや誤字・脱字，レイアウトの認識ミス，ハイフンで改行分割された単語の処理，ページ番号，ヘッダー・フッターの処理など手間のかかる行程を考える必要が少なくなります。このような理由で，HTML で入手可能な論文の場合は，PDF だけでなく HTML ページも Web アーカイブファイルとして保存しておくとよいでしょう。ただし，闇雲に何でもかんでも入手するとなると，いろいろと問題が出てきますので，自分が読む目的で PDF をダウンロードする際についでに入手するようにしてください。

テキストデータの埋め込まれた PDF ファイルや，HTML，Web アーカイブファイルからテキストを抜き出すには，Preview や Safari で一つひとつファイルを開いて，テキストエディットなどのテキストエディタのファイルにテキストをコピー & ペーストして保存する事もできますが，Google で検索して一括処理できる好みのフリーウェアを探すことも簡単にできますし，Automator の使い方がわかる方はワークフローを作って一括処理するのもいいでしょう。しかし，ここでは，まさにその目的で開発した CasualTextractor を使ってテキストを抜き出すことにします。

3.1.2. CasualTextractor

CasualTextractor は，PDF や HTML ファイルからテキストを抜き出して整形処理し，プレインテキストファイルとして保存する目的で制作したアプリケーションです。[8] このアプリケーションには，PDF ファイルからテキストを抜き出して整形して保存する「PDF モード」，ウェブページもしくは，保存した HTML または Web アーカイブファイルからテキストを抜き出して保存する「Web モード」，そして，複数の PDF，HTML，Web Archive，Microsoft Word，リッチテキストファイルなどからテキストを一括して抜き出す「Batch モード」の 3 つのモードがあります。ここでは，PDF ファイルからテキストを抜き出して整形する方法を少し詳しく説明し，後の 2 つを簡単に説明します。

3.1.3. PDF ファイルの処理

CasualTextractor を初めて立ち上げると PDF モードになっているので（図 3.1），メインメニューの「File」から「Open...」を選んで開きたい PDF ファイルを選択するか，左側の PDF ビューに PDF ファイルをドラッグ & ドロップして，PDF ビューに PDF ファイルを読み込みます。PDF ビューに PDF ファイルが表示されている状態で左下の「Extract Text」ボタンを押すと，テキストデータが埋め込まれていれば，それが抜き出されて，右側のテキストビューに表示されます。

図 3.1 CasualTextractor PDF モード

これをこのまま，メインメニューの「File」から「Save Text」を選んで保存してもいいのですが，時間があるならば，不必要なテキストを削除して整形してから保存した方が用例データとしての有用性が高まります。

まずは，用例検索用データとしてあまり有用でない表の数値や Reference List などを削除します。全文データベースとして利用する場合は，残しておいてもいいでしょう。この作業は，左側の PDF ビューでスクロールしながら表や図を探すか，左上の検索ボックスに「Table」などと入力して検索して探します。表などが見つかったら，そのタイトルまたは表の一部を PDF 上で選択して「Go to Selected」ボタンをクリックし，テキストビューに抜き出されたの表の部分に移動して削除します。もしうまくいかなければ，表のタイトルの一部などを選んで繰り返してください。これを繰り返して，表や Reference List などのテキストを削除していきます。

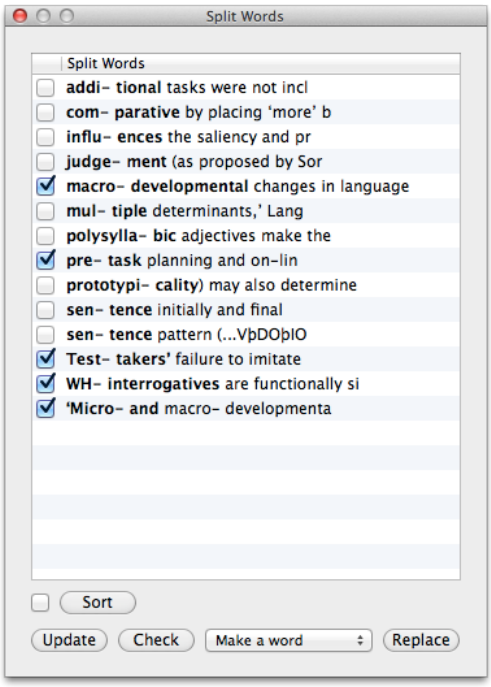


図 3.2 Split Words パネル

次に，PDF ファイルから抜き出したテキストによく見られる，ハイフンで改行されて分割された単語の処理をします。ウインドウ右下の「Split Words」をクリックして，Split Words パネルを表示すると，ハイフンの後に半角スペースが続く文字列がパネル上のテーブルに表示されます（図 3.2）。ここで，下のポップアップメニューで「Make a word」が選ばれている事を確認して「Check」ボタンをク

リックすると，ハイフンとそれに続く半角スペースを削除したときにできる文字列が OS X のスペル辞書および本文中の別の箇所に存在しないものにチェックが入ります。ここでは，チェックの入っていないものがハイフンを取り除く処理の対象になるということです。この時，テーブル上の文字列はデフォルトでは先頭の文字順で並べ替えられていますが，単語数が多い場合は，「Sort」ボタンの左のチェックボックスにチェックを入れてボタンをクリックして，分割された後半部分の先頭の文字で並べ替えて確認する事もできます。リストの文字列を見て，処理しても問題がないことを確認したら，「Replace」ボタンをクリックして，チェックの入っていない文字列からハイフンとそれに続く半角スペースを削除します。次に，ポップアップメニューで「Remove space」を選んで「Check」ボタンをクリックし，ハイフンでつながれた連語としての可能性をチェックします。この場合も，チェックの入っていないものが連語としておかしくない可能性のあるにものになります。確認が終わったら同様に「Replace」ボタンをクリックしてハイフン後の半角スペースを削除します。

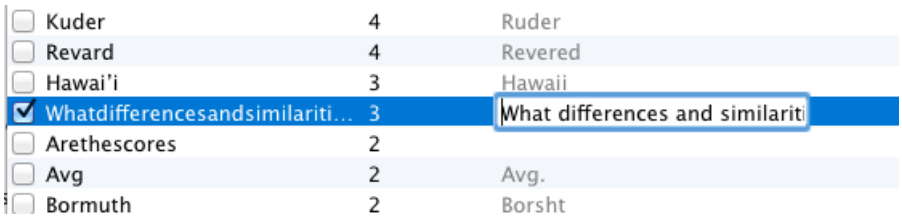


図 3.3 Misspell Word List パネル

ハイフン処理が終わったら，「Misspells」ボタンをクリックして，OS X の辞書に載っていない単語を探すことで，OCR エラーやテキストを抜き出した際に問題がある部分を探します。デフォルトでは，アメリカ英語とイギリス英語でチェックするようになっていますが，どちらか一方もしくは他の地域の英語もしくは英語以外のいくつ化の言語に設定することもできます。これらは，Misspel パネル下部か環境設定で変更してください。ミススペルは最近のテキストから作った PDF ではあまり問題にはならないかもしれないのですが，画像 PDF を OCR 処理した古いものや自分でスキャンして OCR 処理したものでは多く見つかる場合があります。また，よくある OCR の認識ミスだけでなく，PDF に変換した際にレイアウトの問題で隣り合った単語間のスペースが挿入されていないものが多く見られます。ミススペルの頻度が少ない場合は，直したい単語を選んで「Go to Selected」ボタンをクリックしてその単語の位置に移動し，修正してもいいですが，もし頻度が多い場合は，左側のチェックボックスにチェックを入れてから，右側の列のに正しいスペルを入力し，右下の「Replace」ボタンをクリックして，チェックの入った項目をすべて置換します（図 3.3）。標準では，OS X のサジェスチョン機能を使って正しい可能性のある単語が配置されているので，その中から選ぶこともできます。

次に，「Format」ボタンをクリックして，テキストの整形および特定文字列の置換をします。置換される文字列は，「Preferences」の Format Text にあるので，置換したくないもののチェックを外してください（図 3.4）。

整形が終わったら，ページヘッダーなどの処理をします。PDF ビューでページヘッダーがあるところまでスクロールして選択し，「Go to Selected」ボタンをクリックしてその部分まで移動します。そこで，ヘッダーのテキスト部分を改行まで含めて選択して，右クリックもしくは 2 本指クリックでコンテ

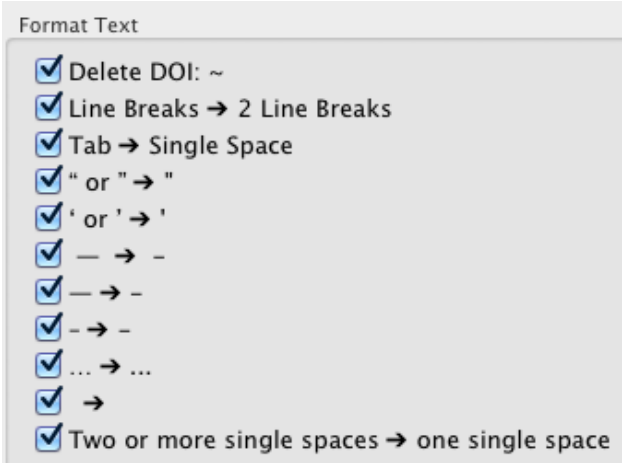


図 3.4 Format Text 設定

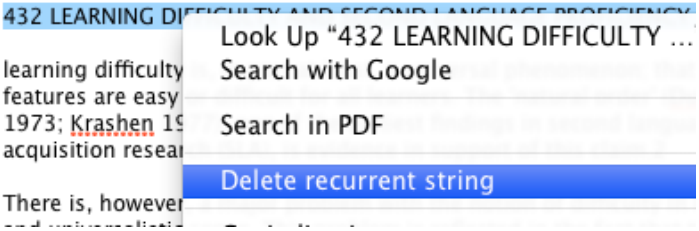


図 3.5 Delete recurrent string

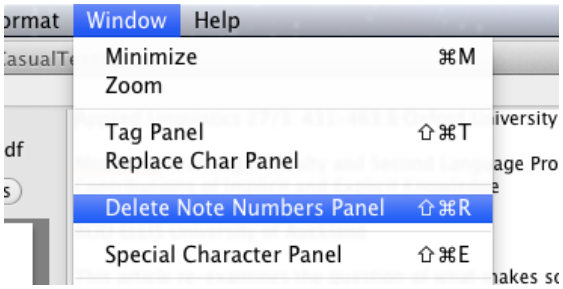


図 3.6 Delete Note Numbers Panel を開く

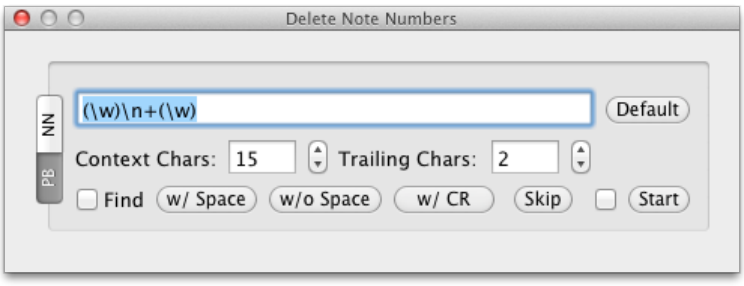


図 3.7 Delete Note Numbers Panel (Page Break)

クストメニューを表示し「Delete recurrent string」を選択します（図 3.5）。これで，選んだ部分と同じ文字列の部分がすべて削除されます。選んだ部分に数字含まれていればその数字の位置に別の数字であっても数字が使われている文字列も削除されます。つまり，ヘッダーの部分は本や論文のタイトルまたは著者とページ番号が記されていることが多いので，ページ番号が違っても，それをなんとか自動で認識させようとしています。これをページの切れ目のところで繰り返してページヘッダーを削除していきますが，うまく行かない場合は自力で見つけていくことになります。

最後に，細かいことになりますが，PDF から抜き出したテキストでページヘッダーなどを削除した後には，パラグラフの途中で改行されてしまっている部分が残っているので，その処理をします。CasualTectractor には，正規表現が使える検索・置換パネルがありますが，この処理だけのためのパネルが用意されています。メインメニューの「ウインドウ」から「Delete Note Number Panel」を選んで開き（図 3.6），「PB」タブを選択します（図 3.7）。ここには，あらかじめパラグラフが途中で改行されてしまっている際に当てはまるであろう正規表現が入力されていますが，これよりも高度な正規表現に置き換えてもらっても結構です。ただ，デフォルトの正規表現のように，最初と最後は半角の括弧 () でくくり，その部分を参照できるようにしてください。「Default」ボタンをクリックすれば，いつでもデフォルトの正規表現が呼び出せます。

これで，「Start」をクリックして検索を開始し，該当箇所が見つかるまで「Skip」をクリックしながら移動します。見つかったら「w/ Space」をクリックして単語間の改行を半角スペースに置き換えます。単語の切れ目で分割されているものが見つかった場合は，「w/o Space」をクリックして置き換え

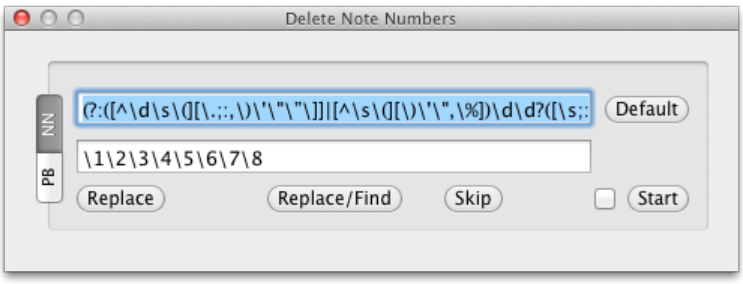


図 3.8 Delete Note Numbers Panel (Note Numbers)

ます。「NN」タブの方はおまけのような機能ですが，本文中にある脚注を参照する上つき文字の番号を探して削除するためのものです（図 3.8）。このため，Delete Note Numbers パネルという名前になっています。使い方は，改ページ部分を削除する機能とほぼ同じです。

あとは，お好みでテキストを整形して，メインメニューの「File」から「Save Text」を選んで保存してください。これで，PDF ファイルの処理は終わりにして，次に HTML/Web アーカイブファイルの処理を少し説明します。

3.1.4. HTML/Web アーカイブファイルの処理

CasualTextractor では，PDF ファイルの他に，HTML や Web アーカイブファイルを開いてテキストを抜き出し，整形して保存することができます。まずは，ウインドウ下部のタブで「Web」を選んで Web モードに切り替えてください（図 3.9）。そして，メインメニューの「File」にある「Open...」を選び，開きたい HTML ファイル，もしくは Web アーカイブファイルを選んで開きます。Web モードには，簡単なウェブブラウジング機能もあって，Safari のブックマークを読み込みます。また、プロキシなどは Safari と同じものが適用されるので，ウインドウ左上のテキストボックスに出版社や論文のサイトアドレスを入力して直接アクセスすることもできます。

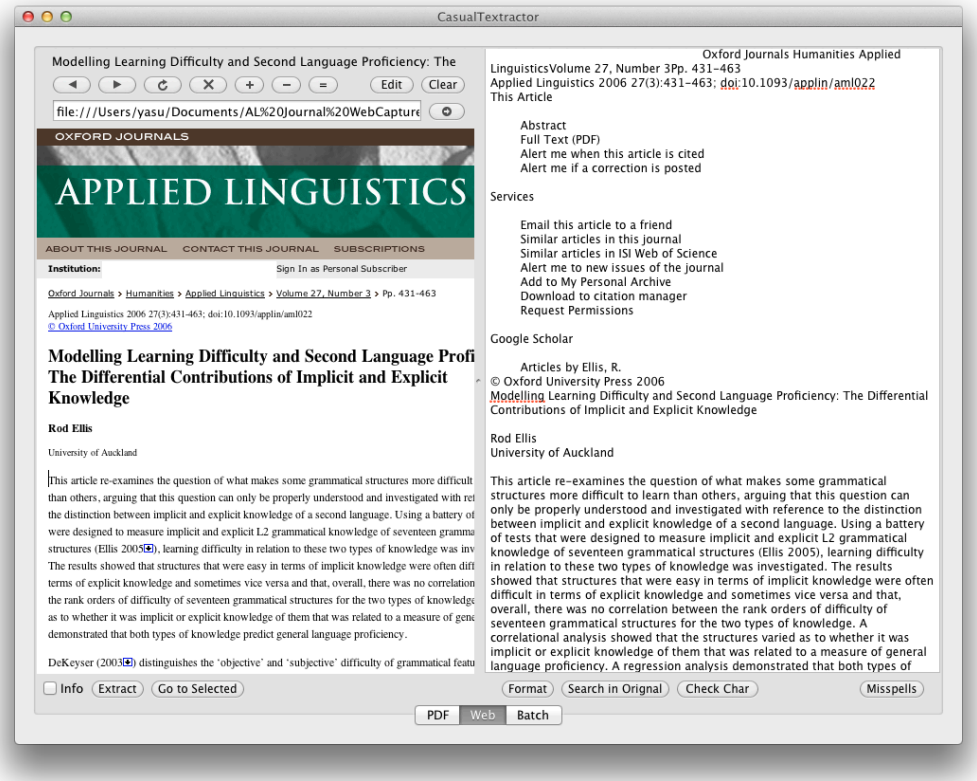


図 3.9 CasualTextractor Web モード

ウェブページからテキストを抜き出すのは，PDF ファイルからに比べて問題が少ないので，機能は少なくなっています。また，出版社によってフォーマットが異なるので，多少の工夫が必要になります。基本的には，ページを表示させてウインドウ左下の「Extract」ボタンをクリックし，テキストを抜き出します。抜き出したら，左側の Web ページを参照しながら不必要な部分を削除していきます。場合によっては，特殊な文字やリンクが本文中に挿入されていますので，対処が必要となる場合があります。HTML が扱える場合は， メインメニューの「Window」から「Web Source Panel」を選んでソースコードを正規表現パネルなどを使って直接編集することもできます。ソースコードを編集した場合は，ソースコードパネル右下の「Apply」をクリックして，Web ビューに反映させてからテキストを抜き出してください。また，Web モードでも Note Numbers Panel は利用できるので，脚注番号などが本文に入っている場合は活用してください。ここでも，編集が終わったら，メインメニューの「File」から「Save Text」を選んで保存します。

3.1.5. バッチ処理

これまでの 2 つのモードは，抜き出したテキストを細かく整形したい場合に使うことを想定していましたが，テキスト化の作業に時間が取れない場合やそれほどの正確さを求めない場合は，バッチモードで，指定したファイルから一括してテキストを抜き出してプレインテキストファイルとして保存することができます。

まずは，ウインドウ下のタブで「Batch」を選んでください（図 3.10）。次に， メインメニューの「File」から「Open...」を選んでテキストを抜き出したいファイルを開きます。ファイルを追加したい

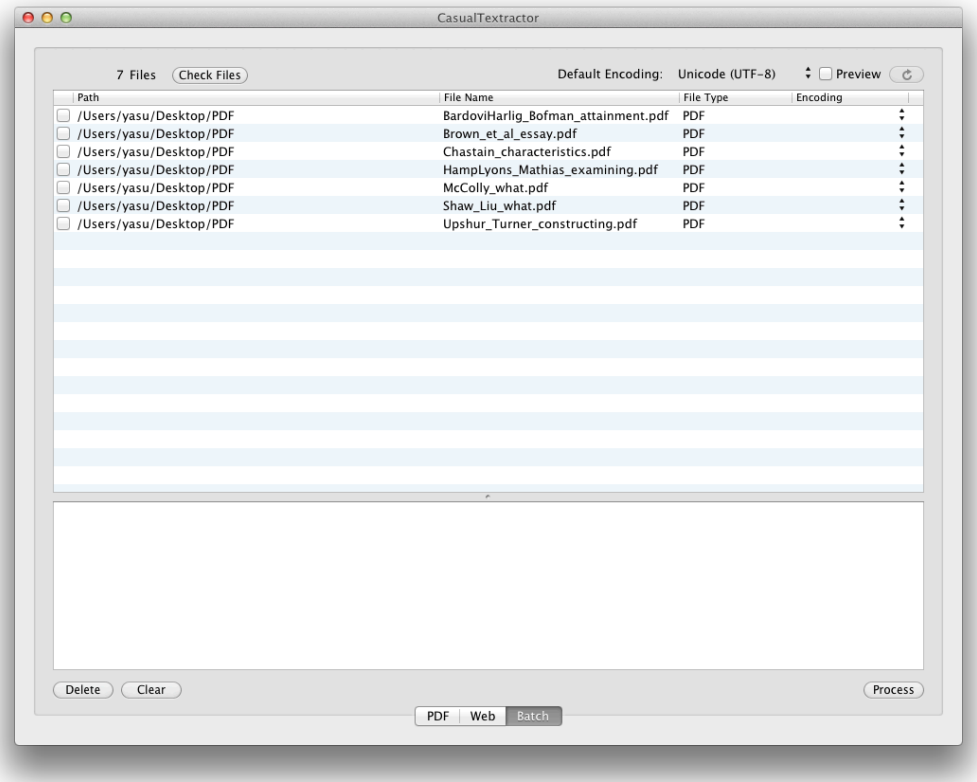


図 3.10 CasualTextractor Batch モード

場合はこの作業を繰り返すか，もしくは、テーブルに直接ファイルをドラッグ ＆ ドロップすることもできます。

テーブルにファイルを追加したら，「Preferences」にある「Batch」の設定をします（図 3.11）。「Process split words」は，PDF モードの Split Words 機能を簡易的に適用するものです。「Create Misspells List」は，OS X のスペルチェッカーでアメリカ綴りとイギリス綴りでチェックして，単語の登録がない文字列を抜き出してリストを作りテキストファイルとして保存するので，後でチェックすることもできます。このファイルは，テキストファイルを保存したフォルダまたは処理後のテキストファイルを元のファイルと同じフォルダに保存する設定をした場合はデスクトップに，MISSPELL_LIST.txt というファイル名で保存されます。「Apply Formatting」にチェックを入れると，環境設定ウインドウの左側にある文字の置換など（図 3.4）を適用します。「Save to Original Folder」にチェックが入れると，元のファイルが入っているフォルダにテキストファイルが保存されますが，そうでない場合は，変換後のファイルを保存するフォルダを処理実行時に指定します。

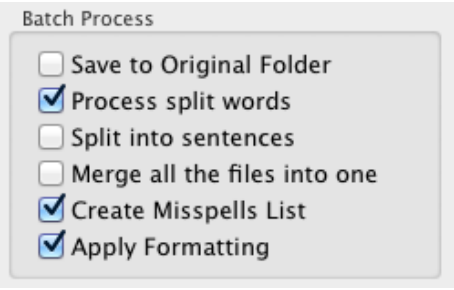


図 3.11 Batch 環境設定

最後に「Process」ボタンをクリックして保存します。保存時には，「File Type」でプレインテキスト（.txt）かリッチテキスト（.rtf），「Encoding」で文字コード，「Line Break」で改行コードを選択します（図 3.12）。デフォルトでは元の改行コードをそのままにする「As is」になっていますが，Mac でテキストファイルを管理するのであれば，LF を選ぶと問題が少なく，Windows でも使いたい場合は，CR + LF を選ぶとよいでしょう。

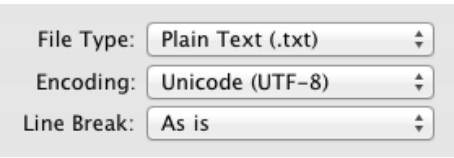


図 3.12 Batch 保存オプション

これで，プレインテキストファイルに変換できたので，CasualConc にテキストファイルを読み込む段階へと移ります。

3.1.6. CasualConcへの読み込み

用例検索ツールとして CasualConc を使う場合は，高速検索が可能で，複数のデータベースファイルを管理できるアドバンストデータベースモードが最適です。ただ，ワイルドカード文字を多用したり，正規表現を使って検索する場合はファイルモードの方が速い場合もありますので，目的によって両モードを使い分けてください。

このモードに入るには，前述の通り（2.1.2.，2.1.3.），メインウインドウ右上のモード切り替えでデータベースモードを選び，「環境設定」の「一般」にある「コーパスモード」で「アドバンスト」を選択します。設定したら，ファイルビューに切り替えて，一つのデータベースファイルにまとめたいテキストファイルを右上のテーブルに追加し，左上のテーブル右下にある「新規データベース」をクリックしてデータベースファイルを作ります。既に使用しているデータベースファイルがある場合は，左上のテーブルに直接追加します。

論文データベースを作る場合は，テキストファイルの数が少なければ，すべてのファイルをまとめて一つのデータベースファイルにした方が便利ですが，ファイル数が多い場合は，すべてのファイルを含む大きなデータベースファイルとともに，専門誌ごとや，研究手法，研究分野などに分けたデータベースファイルを作っておくと，切り替えて検索できたり，横断的に検索できたりして便利です。ただし，多数のデータベースファイルを横断的に検索すると，処理速度が遅くなるので，頻繁に組み合わせて使うデータベースファイルに含まれるテキストファイルは一つにまとめて，新しいデータベースファイルを作っておくことをお薦めします。

前述の通り，データベースモードでは，データベースファイルを作成する際に 2.1.6 にある文字置換処理が適用されて，検索時には文字置換設定は無視されます。前述（3.1.3.）のように CasualTextractor でテキストを抜き出す際に文字置換を行っていれば，同じ文字の置換を設定する必要はありませんが，それ以外の文字（箇条書きの点など）を処理する必要がある場合は，「環境設定」の「文字の置換処理」にチェックを入れてリストにない文字を追加し，この段階で処理してください。どの文字を追加すればよいのかよくわからない場合は，データベースファイルを作る前に，ファイルモードで追加したいファイルの単語リストを作って，単語ではない全角の記号などを見つけてください。例えば，2.2.5. にある Word Count ツールの「指定文字列検索モード」で検索語を正規表現にして [^A-Za-z\W\d] （すべて半角文字）を検索すると（図 3.13），半角アルファベット，半角数字，半角記号，空白文字を除いた文字のリストができるので，全角の記号など，文字として認識させたくないものがあるかどうかを簡単に確認できます。ただし，この正規表現では，半角のアクセント付き文字などもマッチしてしまうので，使用する言語によっては工夫が必要となります。

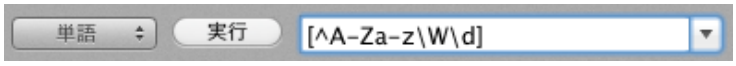


図 3.13 全角記号の確認例

また，例えば，<info> ~ </info> タグでファイルの情報を管理したり，<text> ~ </text> タグなどで本文を区分するなど，コンテクストタグが付けてあるファイルを扱う場合，ファイルモードでは，検索するたびに，削除したり指定したタグの部分だけ読み込んだりする処理をしますが，データベースモード

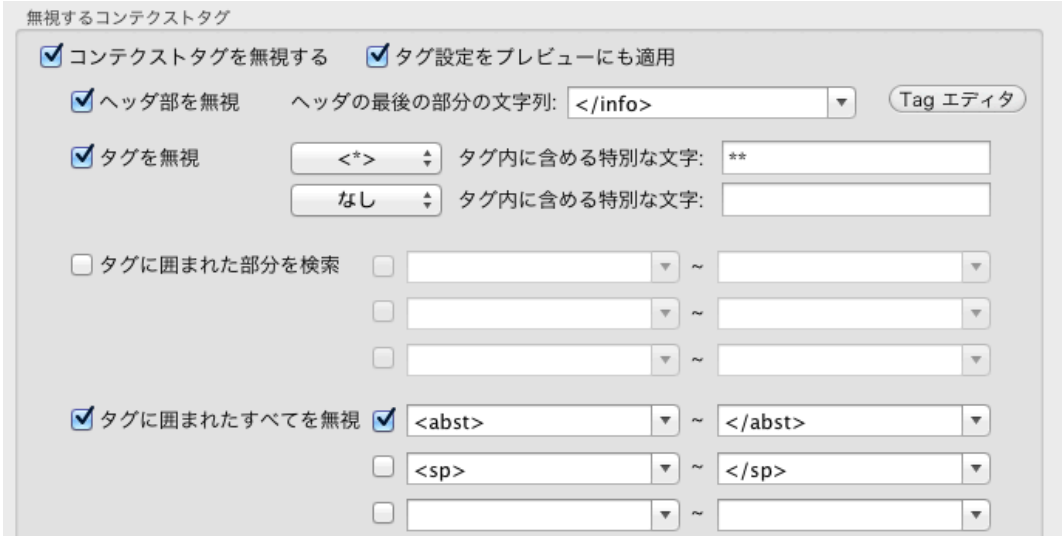


図 3.14 コンテクストタグの設定

では，データベースファイル作成時にそれらの処理がされるので，本文のみをデータベースに含めたい場合は，あらかじめコンテクストタグの設定をして必要のない部分を削除してからデータベースを作ってください。コンテクストタグの設定は「環境設定」の「タグ」でできます（図 3.14）。この機能を有効にするのチェックボックスのラベルは「コンテクストタグを無視する」になっていますが，「タグに囲まれた部分を検索」にチェックを入れて，そこで指定したタグに囲まれた部分のテキストだけ使って検索したり，データベースファイルに含めたりできます。ここでの実際の処理は，タグではなく文字列で見ていますので，タグとして指定した文字列に囲まれたテキストをその指定した文字列ごと無視するか（「タグに囲まれたすべてを無視」），指定した文字列を除いてその文字列に囲まれた部分のテキストだけを処理（『タグに囲まれた部分を検索」）するようになっています。「タグを無視」の所は，右側のテキストボックスに何も指定しないと，<text> </text> など正規表現の \w にマッチする，記号と空白記号を除いた「文字」だけを含むタグだけを無視しますが，それ以外の文字も含めたい場合は，右のテキストボックスに無視するタグに含まれる文字を入力してください。** を入力すると，<nn lemma=“see”> のように改行記号を除く記号や空白記号を含むすべての文字で構成されるタグを無視することができます。

一番下の「指定した文字列を無視」（図 3.15）は，正規表現で指定した文字列を無視する設定ができます。追加する際は，「追加」ボタンをクリックしてテーブル上に挿入された行を選び，左側のテキストボックスに正規表現を入力します。右側のテーブルの一番左のチェックボックスにチェックを入れると，その行の正規表現にマッチした文字列が削除されます。つまり，チェックが入っていない列は飛ばされます。二番目にチェックを入れると大文字小文字の区別をし（C），三番目では複数行に渡ってマッチするかどうかの指定をします（M）。図 3.15 の正規表現は論文によく見られる (2000) などの引用の際の年号を削除することを試みるものです。

図 3.15 指定した文字列を無視

これらの設定が終わったら，メインウインドウのファイルビューでデータベースファイルを作成してください。ここから先は，論文を書く際に各ツールを利用して単語などの使い方を探る例を示していきます。用例検索の際には，lemma 処理や，異綴り処理などがされると便利な場合が多いので，2.1.1.4. を参考に，lemma リストファイルと異綴りリストファイルを読み込んで準備しておきます。それに加え，Stop Words も有用なので，2.2.1.5. を参考に，Stop Word リストを入手して読み込んでおきます。

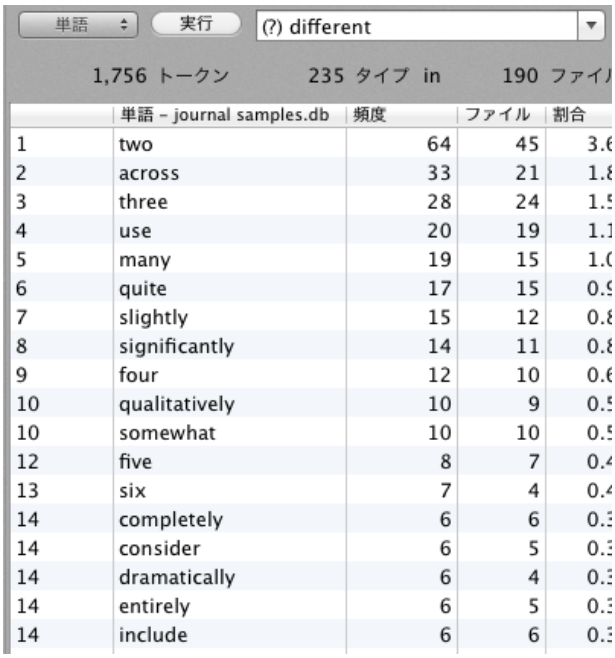
前述のように，個人的な用例コーパスを作る際は，自分が専門とする分野の論文を集めた方が使い勝手がよくなります。また，十分な用例を見つけるには，ある程度の論文の量を集めることが必要になってきます。ここに示す例は，作者個人の専門である，言語テスト・評価法，英語教育分野での論文を集めて作ったサンプルデータベースファイルを使うため，用例に偏りがあることを最初に断っておきます。論文数は，学位論文や何本も論文を書かれた方なら十分にあり得る 200 本ほど含んでおり，前述の CasualTextractor で PDF などからテキストを抜き出して多少の整形をしてあります。

3.2. 検索例 1

まず始めは，「大きく違う」感じを表現がしたい時，different という形容詞を修飾するにはどの単語を使ったらよいか知りたいとします。このような場合，いきなり 2.2.2. の Concord で KWIC 検索をかけて L1 の単語で並べ替えてもいいのですが，結果の量が多くなって情報量が多くなり，判断がしづらくなります。また，2.2.4.1 にある Collocation で different を検索するのもいいかもしれませんが，その場合は，範囲に制限をかけて，左側だけに絞ったり，L3 - L1 に絞ったりした方がいいでしょう。ただ，different のような頻出する単語の場合，直前に使われる単語だけ見れば十分情報が得られる可能性が高いので，2.2.5. で説明した Word Count ツールの機能である「指定文字列検索モード」を使うことにします。

図 3.16 指定文字列検索モードの設定

まずは，「環境設定」の「その他」で，「指定文字列検索モード」にチェックを入れます（図 3.16）。ここでは，半角の丸括弧 ( ) にワイルドカード文字を入れて，そこに現れる単語だけのリストを使いたいので，「文字列全体を検索」のチェックは外したままにしておきます。さらに，この場合には，機能語などの頻出語はあまり有用ではないので，2.1.1.5. にある Stop Words 処理をするように設定します。



	単語 - journal samples.db	頻度	ファイル	割合
1	two	64	45	3.6
2	across	33	21	1.8
3	three	28	24	1.5
4	use	20	19	1.1
5	many	19	15	1.0
6	quite	17	15	0.9
7	slightly	15	12	0.8
8	significantly	14	11	0.8
9	four	12	10	0.6
10	qualitatively	10	9	0.5
10	somewhat	10	10	0.5
12	five	8	7	0.4
13	six	7	4	0.4
14	completely	6	6	0.3
14	consider	6	5	0.3
14	dramatically	6	4	0.3
14	entirely	6	5	0.3
14	include	6	6	0.3

図 3.17 指定文字列検索モードでの検索結果

これで，ワイルドカードモードに設定して Word Count ツールに移動し，「実行」ボタンの右に現れたテキストボックスに，(?) different と入力し検索します。このように一文字以上の文字列のワイルドカード文字である ? を括弧に入れると，その括弧の部分の文字列（単語）だけのリストが作れます（図 3.17）。また，-ly で終わる単語だけに絞って検索する場合には，(?ly) different と ? の後に続けて ly を入力することで，ly で終わる単語だけを検索することもできます。

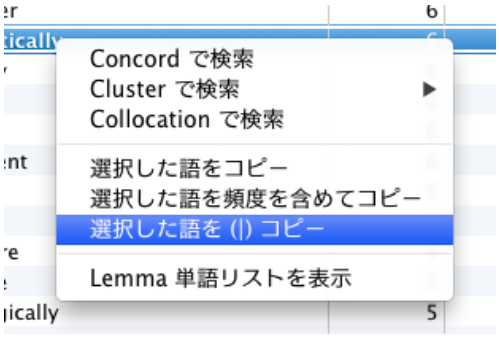


図 3.18 選択した語を (|) コピー

そこで，図 3.17 のリストから「大きく」という感じの意味を表す単語を探すと，quite，significantly，dramatically などが候補になりそうです。ここで，この 3 つの単語をテーブル上で選び，右クリックでコンテクストメニューを表示させて，「選択した語を (|) コピー」を選択します（図 3.18）。これで，この 3 の単語が | で区切られて括弧に入った形でコピーされるので，Concord に移動して，検索テキストボック

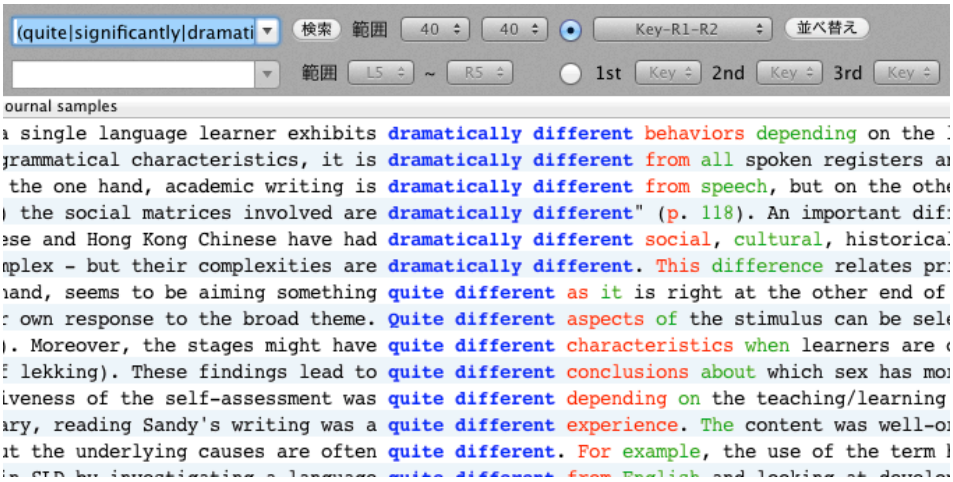


図 3.19 (quite|significantly|dramatically) different の Concord での検索結果

スにペーストし，different を付けて検索します。このとき，検索語は (quite|significantly|dramatically) different になっているはずです。また，検索時か検索後にプリセットの並べ替えで Key-R1-R2 のような Key が最初になっているものを選択するか，任意の並べ替え順で一番目に Key を選んで検索または並べ替えをすると，検索された語で同じものが並んで表示されるので，確認しやすくなります（図 3.19）。

この結果の行をクリックしてもう少し広い文脈を下部のコンテクストビューで確認すると，この 3 つの単語がどのように使われているかの詳細がわかります。この場合は，significantly が使われている場合は，数値，特に，統計の値が使われていることが多いことがわかります。dramatically と quite では，この結果での情報からははっきりとはわかりませんが，それぞれの単語の意味や頻度などから，quite の方が多少一般的に different を修飾するのに使われる副詞ではないかと推測できます。このような情報から，「大きく違う」の意味では quite が多少無難で一般的であり，statistically を使う時は統計の値などを示した時の方がよいのではないか，ということになります。実際には，これでもサンプルが小さく絶対的なことは言えないので，不安な場合は Google Scholar などで二重引用符つきで “quite different” や “significantly different” で検索して確認するのもいいかもしれません。

3.3. 検索例 2

次に，言語学・応用言語学や言語教育でよく出てくる competence と ability がどう違うのかを探ってみることにします。両語とも日本語では「能力」とされることが多いですが，同じような文脈で同じように使われているのだろうかという疑問が生じます。そこで，同じ論文コーパスを使って，検索例 1 と同様に Word Count の「指定文字列検索モード」から始めます。今回は 2 つを比べたいので，Word Count の左右のテーブルを使い，片方で ability を，もう一方で competence を検索します。また，複数形も検索結果に含めたいので，ワイルドカードモードで次のような文字列を検索します。

Ability: (?) (?:ability|abilities)

Competence: (?) competence*

ここでは，直前の単語だけでリストにするために，両方ともに (?) で，ワイルドカード文字を使ってマッチした単語を抜き出します。そこの続く部分で ability の方は，単数系と複数形を (?:) の括弧でくくり | で区切っています。他のツールでワイルドカード検索をする場合と違うのは，括弧内に ?: があることですが，これは，Word Count の「指定文字列検索モード」で括弧だけを使うと，その単語がリストに抽出されてしまうので，それを避けるために付けています。この処理は正規表現で後方参照しない処理と同じです。また，competence の最後に使っている * は他のツールのワイルドカード文字と同じで，0 文字以上の連続する文字列に一致します。ここでは，competences 以外に competence に文字が続く単語が考えられないので使用します。他の可能性が考えられる場合は，ability と同じように処理します。この 2 つの検索語で Stop words 処理をせずに検索した結果が図 3.20 です。

556 トークン 119 タイプ in 177 ファイル

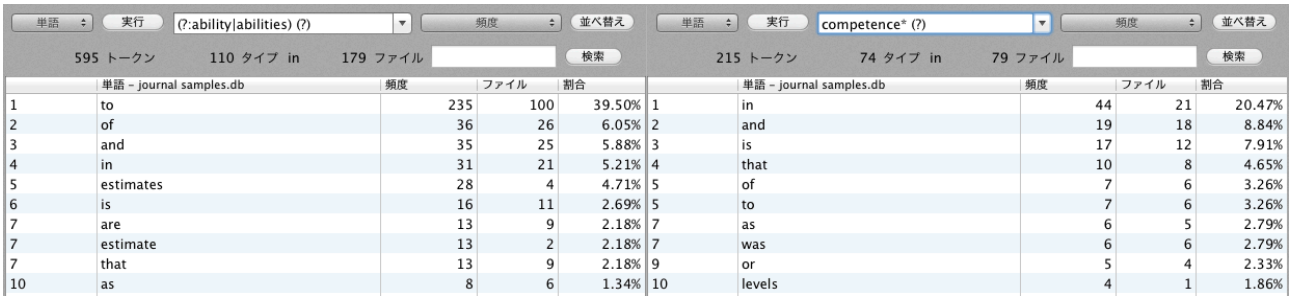
	単語 - journal samples.db	頻度	ファイル	割合
1	the	107	60	19.24%
2	their	49	36	8.81%
3	writing	45	12	8.09%
4	language	40	26	7.19%
5	musical	36	1	6.47%
6	of	18	13	3.24%
7	and	12	12	2.16%
7	person	12	1	2.16%
7	verbal	12	4	2.16%
10	an	10	8	1.80%

343 トークン 88 タイプ in 79 ファイル

	単語 - journal samples.db	頻度	ファイル	割合
1	communicative	42	23	12.24%
2	linguistic	27	16	7.87%
3	language	20	12	5.83%
4	rhetorical	19	1	5.54%
5	of	18	8	5.25%
6	english	16	4	4.66%
6	humor	16	1	4.66%
8	pragmatic	14	6	4.08%
9	l2	13	7	3.79%
10	sociolinguistic	11	6	3.21%

図 3.20 ability と competence を直前で修飾する語の検索結果

このコーパス全体としては，ability の方が competence よりも，1.5 倍ほど多く現れているのがわかります。そして，一目見てわかるのが，ability には the が冠詞として使われる例が二割近くあるのに対し，competence は直接冠詞が付いている場合が少ないということです。実際に，the * * * * * competence* と修飾語の前についている場合を検索しても ability ほどではありませんでした。そこで，両方の単語で直後に使われている単語のリストを作って比べてみます。



595 トークン 110 タイプ in 179 ファイル

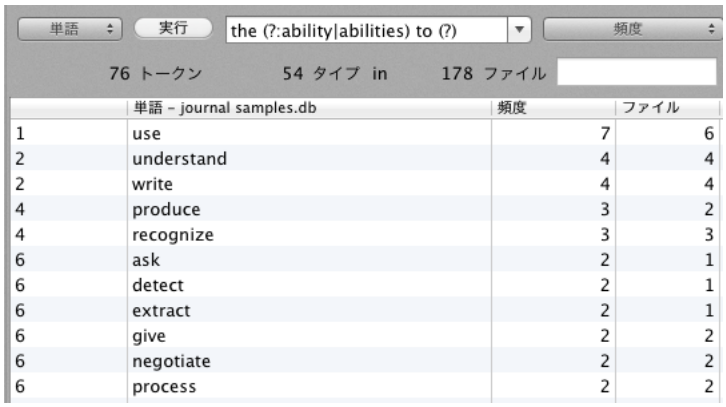
	単語 - journal samples.db	頻度	ファイル	割合
1	to	235	100	39.50%
2	of	36	26	6.05%
3	and	35	25	5.88%
4	in	31	21	5.21%
5	estimates	28	4	4.71%
6	is	16	11	2.69%
7	are	13	9	2.18%
7	estimate	13	2	2.18%
7	that	13	9	2.18%
10	as	8	6	1.34%

215 トークン 74 タイプ in 79 ファイル

	単語 - journal samples.db	頻度	ファイル	割合
1	in	44	21	20.47%
2	and	19	18	8.84%
3	is	17	12	7.91%
4	that	10	8	4.65%
5	of	7	6	3.26%
5	to	7	6	3.26%
7	as	6	5	2.79%
7	was	6	6	2.79%
9	or	5	4	2.33%
10	levels	4	1	1.86%

図 3.21 ability と competence の直後に使われる語の検索結果

その結果が図 3.21 です。この比較からわかることは，ability の場合はかなりの割合で to が直後に続いており，competence の場合は in が続いている割合が高いということです。さらに，ability to の後に続く単語も調べてみると，動詞がほとんどであることがわかります（図 3.22）。このことから，大まかではありますが，ability の場合は，the ability to do という形で「〜する能力」を表すことが多く，competence は competence in で「〜の能力」になることがよくあるようだということが見て取れます。



	単語 – journal samples.db	頻度	ファイル
1	use	7	6
2	understand	4	4
2	write	4	4
4	produce	3	2
4	recognize	3	3
6	ask	2	1
6	detect	2	1
6	extract	2	1
6	give	2	2
6	negotiate	2	2
6	process	2	2

図 3.22 the ability to の検索結果

さて，ここでまた，この 2 つの単語の直前に使われる単語を見てみることにします。最初の検索では，stop word 処理をしなかったので，次は stop word 処理をしたリストを見てみます（図 3.23）。ability を修飾する単語は，writing, verbal, listening など，実際の言語運用に関する単語が並んでいますが，competence の方は，ability の方に見ら

	単語 – journal samples.db	頻度	ファイル	割合
1	writing	45	12	8.09%
2	language	40	26	7.19%
3	musical	36	1	6.47%
4	person	12	1	2.16%
4	verbal	12	4	2.16%
6	rasch	10	2	1.80%
7	listening	9	3	1.62%
7	reading	9	6	1.62%
9	spelling	7	1	1.26%
10	examinee	6	2	1.08%

	単語 – journal samples.db	頻度	ファイル	割合
1	communicative	42	23	12.24%
2	linguistic	27	16	7.87%
3	language	20	12	5.83%
4	rhetorical	19	1	5.54%
5	english	16	4	4.66%
5	humor	16	1	4.66%
7	pragmatic	14	6	4.08%
8	l2	13	7	3.79%
9	sociolinguistic	11	6	3.21%
10	oral	9	2	2.62%

図 3.23 ability と competence を直前で修飾する語の検索結果（Stop words 処理適用）

れない communicative, linguistic など，抽象的な単語が並んでいます。前述の in の後に続く単語を Concord でもう少し広い文脈を含めて見ても，同じような傾向が見られます。language など，両方で使われている単語もあり，使用例が重なる部分もありますが，全体的な傾向としては，ability は動作を伴う言語運用に関係する「能力」を ~ ability もしくは (the) ability to (do) という形で表す傾向が強く，competence の方は，直接確認できない言語に関する抽象的な「能力」を ~ competence もしくは competence in ~ の形で表す傾向が強いことがわかります。ability は一般的な用語で，competence はどちらかというと学術的な用語であることを考慮すると，ある程度当然の区別ではありますが，実際の使用例からもそのことがよくわかり，論文を書く際には，その辺りの違いを意識して使い分けることが重要になります。

3.4. 検索例 3

検索例の最後に，lemma 処理と異綴り処理をして Concord で検索した場合に，どのような検索結果になるかを示してみます。lemma リストファイル，異綴りリストファイル共に，各自で用意したファイルを読み込めるようにしてありますが，ここでは CasualConc のディスクイメージに入っている「e-lemma」ファイルと「a-e spelling differences」ファイルを使っています（手順については 2.2.1.4. を参照してください）。ただし，ここでは検索語を処理したいので，「検索語も Lemma 処理する」と「検索語も異綴り処理する」にチェックを入れます。

準備ができたら，前述の論文コーパスで Concord で realise を検索してみます。結果は，図 3.24 のように，イギリス綴りである realise だけでなく，アメリカ綴りである realize と共に，lemma に含まれる活用形が両方の綴りで検索されています。あくまでも，lemma リストと異綴りリストに単語があるという前提ですが，すべての活用形と異綴りを手入力する場合に比べて，手間や検索漏れを減らすことができます。ただし，CasualConc の lemma 処理では品詞情報が扱えないため，名詞と動詞が同じ綴りの場合など，見出し語の綴りが同じ場合は，元々の lemma リストに別々に登録してあったとしても，ここでの処理では区別されません。実際にどのような単語が検索されるのかをチェックしたい場合は，メインメニューの「ウインドウ」から「Lemma チェッカーパネル」を選んで，Lemma チェッカーパネルを開き，そこでどのような単語が lemma と異綴り処理された場合に含まれるのかをチェックできます。また，複合語や成句など，検索語に 2 つ以上の単語が含まれる場合は，すべての単語に lemma および異綴り処理が適用されます。また，Concord 以外に，Cluster と Collocation でも同様の検索語処理がなされます。

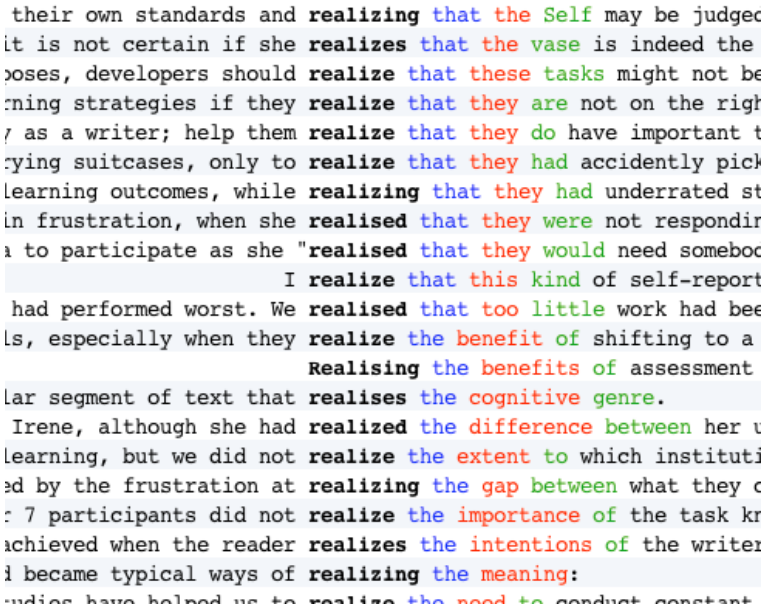


図 3.24 realise の検索結果（Lemma，異綴り処理）

Lemma チェッカー（図 3.25）では，一番上のテキストボックスに単語を入力して「Check」ボタンをクリックすると，「環境設定」の「Lemma」のところでチェックが入っている項目を反映した結果が返ってきます。ここでの例では，lemma と異綴り処理の両方が適用される設定になっていて，イギリス綴りの原型である realise で検索すると，アメリカ綴りである realize が lemma として登録されており，含まれる単語にはその両方の綴りの活用形が表示されています。元々の e-lemma リストファイルには，realize とその活用形しかありませんが，異綴りリストファイルの方にそれぞれの活用形のイギリス綴りが登録されているため，結果としてすべてを網羅したリストが返ってきています。ここで，

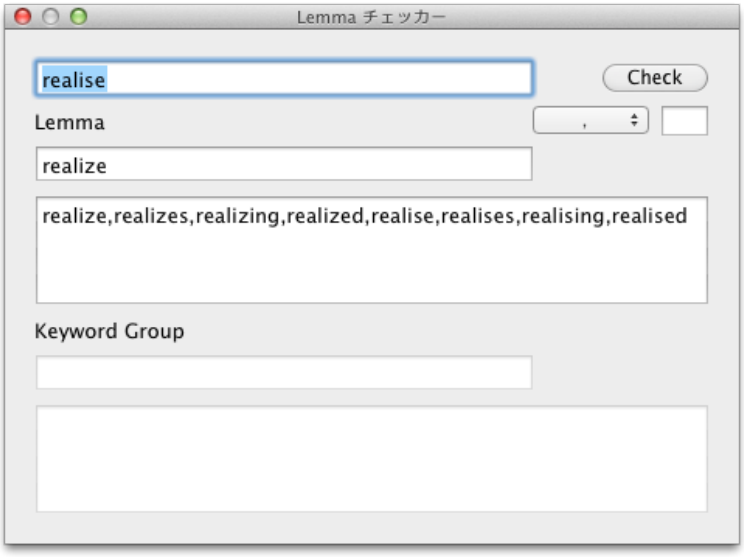


図 3.25 Lemma チェッカーパネル

キーワードグループにも登録してあれば，その結果も返ってきます。また，lemma に含まれる単語のリストは，標準ではカンマ（,）で区切られて表示されますが，スラッシュ（/）や縦棒（|），もしくは指定した文字列に変更することもでき，CasualConc 以外のアプリケーションやスクリプト，文書などで利用したい場合にも結果をコピー & ペーストして使うことができます。

これ以外にも，様々な検索ができますのでいろいろと工夫してください。例えば，左一つ目の単語だけでなくもう少し広く見たい場合は，Collocation で範囲を左だけにして検索したり，ある単語を含む連語にはどんなものがあるか見たいときには，Cluster で検索したりしてください。ワイルドカード文字を使って検索すれば，かなり自由度の高い検索ができるはずです。

3.5. 単語頻度を使った予備研究

ここまでは，用例検索ツールとしての CasualConc の応用方法を見てきましたが，最後に少しだけ言語研究の予備研究としての活用法に触れてみます。ここであえて「予備研究」としているのは，アカデミックな研究をされる場合は，CasualConc の結果だけに頼らずに他の方法でも検証していただきたいという思いを込めています。これは，開発段階で問題が見つかったり，問題の報告があれば修正するので，出てくる結果は基本的に大きな問題はないと考えていますが，厳密な結果を求められる場合は，各種の設定や速度を優先させるための処理方法などのため，利用者が望むような厳密性に答えられない可能性があるためです。本稿は，基本的な設定など，内部でどのような処理がなされているかも多少明らかにすることで，CasualConc の処理の特性なども理解してもらいたいという希望も込めています。

さて，ここでは，2.2.6 のコーパスファイル情報ツールにある「単語グループ頻度表」機能を使って Brown Corpus (Francis & Kucêra, 1964) と MICASE (Simpson et al., 2002) の分析を試みます。Brown Corpus はアメリカのブラウン大学で作られた世界初の電子コーパスで，アメリカ英語の書き言葉が約百万語集められています。最初の版ができたのが 1964 年と少し時代は感じますが，15 の項目に分けてバランスの考慮されたコーパスになっています。MICASE はアメリカのミシガン大学で作られたアカデミックな話し言葉のコーパスで，ミシガン大学でのさまざまな状況における話し言葉を集めたコーパスです。

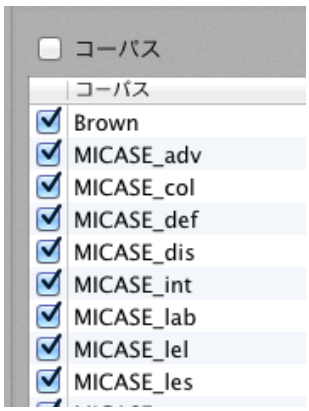


図 3.26 使用コーパス

ここでは，高速検索よりも，数え漏らしの少ないことが優先されるので，集計のたびにファイルからテキストを読み込む「アドバンストファイルモード」を使うことにします。Brown Corpus は，項目ごとに一つのファイルとなっているので，すべてまとめてコーパスを作成し，MICASE はそれぞれの項目に複数のファイルがあるので，ファイル名を参考にして項目ごとにコーパスを作って，作成したコーパスすべてにチェックを入れておきます（図 3.26）。

コーパスの準備ができたら，「コーパスファイル情報」ツールに移動して，ポップアップメニューから「単語グループ頻度表」を選んで切り替え

図 3.27 単語グループ頻度表

図 3.28 グループ分け設定

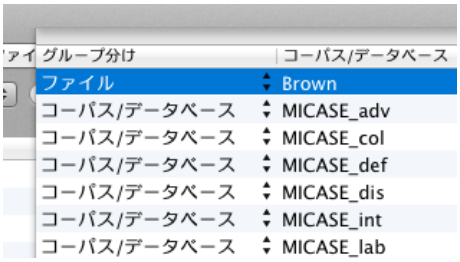


図 3.29 グループ分け指定

ます（図 3.27）。この機能は，シンプルモードでは使えなくなっていますので，必ずアドバンストコーパスモードに切り替えてください。次に，集計のグループ分けを指定しますが，Brown Corpus はファイルごと，MICASE はコーパスごとに集計したいので，「混合」を選び（図 3.28），「設定」ボタンをクリックして現れるパネルで，Brown Corpus だけを「ファイル」に変更します（図 3.29）。

このパネルでは，集計結果テーブルの左端の列に表示される各行のラベル（グループ名）も指定できますが，何も指定しなければ，グループ分けにファイルが選択してあればファイル名が，コーパス/データベースが指定してあれば，コーパス/データベース名が割り当てられます。ラベルを変更したい場合は，グループ分けにファイルを選んだコーパス/データベース一つを選択して，左下の「グループラベル」ボタンをクリックします。そこで，もう一つパネルが現れるので，一行につき一つのラベルで含まれるファイルに付けたいラベルを入力して，「閉じる」をクリックします。コーパス/データベースを選んだものは，複数を選択して「グループラベル」ボタンをクリックし，同様にラベルを入力していきます。どちらの場合も現れたパネルの左下にラベルが必要なファイル数，もしくはコーパス/データベース数が表示されるので参考にしてください。修正が必要な場合は，一つであれば修正したい行の「ラベル」列のセルをダブルクリックして修正します。このラベルを変更する機能は，テーブルの結果を書き出したり，コピー & ペーストして Excel などで編集する場合は特に必要ないかもしれませんが，統計処理をさせるアプリケーションに直接読み込ませたりする際に，そこでの修正が容易でない可能性があるので追加した機能です。すべての入力が終わったら，「閉じる」ボタンをクリックしてパネルを閉じてください。

次に，頻度を数える単語を読み込むのですが，まずは，分析に使う単語リストを用意します。ここでは，文部科学省中学校学習指導要領の外国語のウェブページ[9]にある別表 1 の英単語を使ってみることにします（サイトのアドレスは巻末注にあります）。ここでは，ウェブページなので、直接ブラウザーで開いて単語をコピー & ペーストすればいいのですが，多少の加工が必要なので，CasualTextractor を使います。CasualTextractor を開いたらウインドウ下のタブで「Web」をクリックして Web モードに切り替えて，左上のテキストボックスにウェブアドレスを入力して右のボタンをクリックするか Enter キーを押し，文部科学省の該当ページを開きます。そして，ページを下の方の別表 1 まで移動して単語の部分を選択し，左下の「Extract」ボタンをクリックして単語を抜き出します（図 3.30）。単語以外の部分も抜き出してしまった場合は削除してから，メインメニューの「Edit」から「Regular Expression Search」を選んで，Regex Search パネルを表示させます（図 3.31）。これで，上のボックスに半角で \s+ 下のボックスに \n と入力し（\ は Option キーと ¥ キーを押して入力します），

図 3.30 CasualTextractor Web モード - 文部科学省中学校学習指導要領 外国語

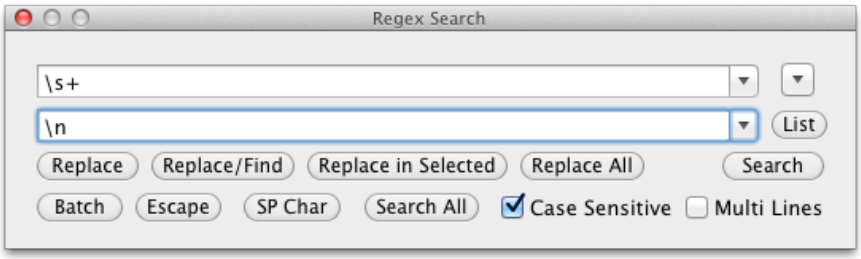


図 3.31 正規表現検索パネル

図 3.32 読み込みボタン

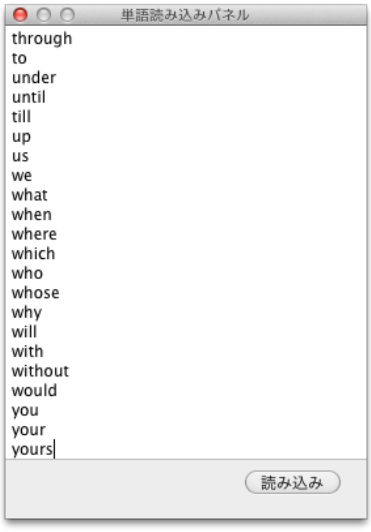


図 3.33 単語読み込みパネル

「Replace All」をクリックします。これで一行一単語になったはずですが，till にだけ括弧がついているので消しておきます。また，CasualTextractor を使いたくない場合は，これと同じことが正規表現が使えるテキストエディターでもできます。

使いたい単語のリストができたら，CasualConc に戻って，「読み込み」ボタン（図 3.32）をクリックして読み込みパネルを開き，CasualTextractor で作った単語リストをコピー & ペーストして，パネル上の「読み込み」ボタンをクリックします（図 3.33）。これで単語リストが読み込まれたので，「確認」ボタンをクリックして確認用の単語リストパネルを開きます。さっと確認して，空白行が見つかった場合は，選択してからパネル右下の「削除」ボタンをクリックして削除します。このリストは，派生系まで含んで網羅的なので，このまま使うのも一つの手ですが，CasualConc に lemma リストが読み込ん

図 3.34 単語リスト確認

である場合は（2.2.1.4），単語リストを lemma 化してみてもいいでしょう（図 3.34）。ただし，このリストの場合には，活用形も含めて網羅的になっているので，単語として見れば重複するものや，もともとは前置詞であった単語が動詞として使われている場合などは考慮されていないであろうと想定されることから，単語リストパネル上での編集が必要になります。ヘッダーのところをクリックすると，アルファベット順で並べ替えられますし，単語のセルをダブルクリックすることで編集もできますので，試してください。リストの編集が終わったら，パネルを閉じます。

これで，すぐに頻度表を作成してもいいのですが，もう少し設定をつめてみます。初期設定の状態では，頻度表は実際にファイルもしくはコーパス内に現れた頻度の集計になり，各ファイルやコーパスなどの総頻度が同じであればグループ間の比較に使えるのですが，ほとんどの場合はそのようなことはないので，何らかの標準化が必要になります。CasualConc では，何語ごとに現れるかを指定して，標準化した頻度を集計することができます。また，コピーする際に，合計頻度の列や行が不要な場合は除く設定にできます。これらは，「環境設定」の「ファイル情報」にある「単語グループ頻度表」で設定します（図 3.35）。標準化は，まずチェックを入れてから，何語ごとの頻度にするかを指定します。総頻度行，列のコピーは，両方ともコピーする，どちらか一方だけコピーする，両方ともコピーしない設定にできます。

図 3.35 単語グループ頻度表環境設定

File/Group	TOTAL	a	about	across	after	all	among	and	another	anyone
TOTAL	49107.65	2510.19	361.73	18.87	73.07	332.72	18.66	2768.02	73.36	13.59
B_Press_report	40396.40	2815.59	164.70	17.93	169.18	217.36	43.70	2454.82	52.66	12.32
B_Press_edit	44557.42	2513.65	115.00	16.43	69.37	292.07	29.21	2482.61	69.37	16.43
B_Press_review	42442.53	3130.73	211.54	5.64	101.54	420.25	45.13	3283.03	45.13	28.20
B_religion	47139.94	2377.87	184.24	11.52	63.33	371.36	71.97	2775.14	54.70	5.76
B_skill/hobby	42783.97	2897.07	139.32	25.95	101.08	303.23	21.85	2989.95	75.12	10.93
B_popular_lore	45290.00	2947.21	170.23	27.69	124.08	257.39	52.30	2915.42	101.52	15.38
B_biography/memoir	46912.83	2675.48	168.61	24.93	104.97	316.23	43.96	2935.29	72.17	16.40
B_Misc_gov/report	42488.64	1946.13	121.93	11.08	68.09	272.36	31.67	3097.34	44.34	3.17
B_academic	43662.01	2614.05	133.06	12.26	78.49	240.99	36.18	2629.38	57.64	7.97
B_Fict_general	48437.50	2570.51	292.65	56.48	114.66	364.53	25.67	3032.58	85.57	17.11
B_Fict_myst/detect	49185.35	2778.30	279.49	74.53	118.01	341.59	18.63	2656.15	91.09	22.77
B_Fict_sci	47521.14	2230.14	149.23	16.58	116.07	547.17	91.20	2437.41	58.03	16.58
B_Fict_adv/west	47955.84	2758.60	211.94	78.62	114.51	333.29	11.96	2922.68	56.40	17.09
B_Fict_roman/love	49944.66	2652.77	280.94	34.05	119.19	415.45	20.43	3265.74	49.38	18.73
B_humor	48226.49	3322.95	196.75	16.40	158.50	349.78	38.26	2891.18	98.38	38.26
M_advise_session	51935.09	2343.28	510.65	6.72	52.07	372.91	3.36	2988.31	68.87	5.04
M_colloquia	51478.02	2693.14	484.68	18.07	68.53	375.66	17.44	3559.70	85.97	7.48
M_dis_defence	52516.22	2163.79	553.69	8.79	40.43	349.79	14.06	2596.19	52.73	12.30
M_discuss_sec	52134.30	2318.36	571.76	2.61	49.60	288.49	6.53	2665.59	92.68	22.19
M_interview	49684.35	2577.78	435.89	7.52	75.15	300.62	22.55	4005.71	82.67	15.03
M_lab_section	50342.74	2530.39	279.82	5.36	49.54	409.68	.00	1891.77	107.11	25.44
M_lrg_lecture	51018.86	2526.20	527.32	14.03	65.11	378.77	14.03	2979.06	103.95	11.51
M_sml_lecture	51170.28	2541.96	457.58	15.30	60.60	321.99	6.73	2751.00	78.36	16.83
M_meetings	51182.23	2467.76	312.66	6.98	37.69	351.74	2.79	2471.94	62.81	27.92
M_office_hours	52006.44	1933.97	478.65	9.27	31.18	318.54	4.21	2580.31	56.46	4.21
M_seminars	51844.50	2402.28	578.54	7.66	47.89	323.75	10.86	2516.59	58.11	7.66

図 3.36 単語リストの頻度集計結果

最終的にすべての設定が終わったら「実行」ボタンをクリックして，頻度表を作成します（図 3.36）。この例では，Brown，MICASE それぞれの下位分類ごとにラベルを設定してあります。これで，この結果をメインメニューの「ファイル」から「テーブルの結果を書き出す」を選んで CSV もしくは tab-delimited テキストファイルとして保存するか，テーブル上で右クリックしてコンテクストメニューから，「すべてをコピーする」を選んでテーブルの結果をコピーし（図 3.37），R などの統計処理アプリケーションなどに読み込んで処理します。テーブル上で選択されている行があれば，その部分だけコピーすることもできます。コピーする際は，ヘッダー部分のラベル表示も一緒にコピーされるので注意してください。

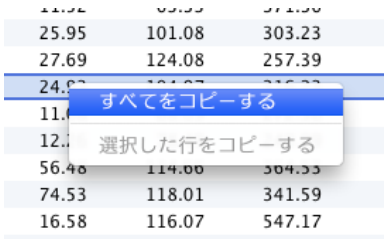


図 3.37 結果のコピー

では，このテーブルの頻度表でどのようなことができるのかを少しだけ紹介して終わりにします。今回の例のように，口語と文章語という特性の違う 2 つのコーパスを用いて中学校学習指導要領で提示されている基本的な単語の頻度を集計してみました。このような基本的な単語は高頻度で使われていて，特に口語で多く使われることがわかっています (Biber et al., 1999 など)。ただ，MICASE はアカデミックな環境での口語なので，一般的な会話の英語とはまた少し違う可能性があり，文章語とは明確に分類できるのかどうかが気になります。そこで，R を使って，クラスター分析とコレスポンデンス分析のグラフを書いてみます。ただ，それぞれの分析についての詳しいことはここでは触れません。

図 3.38 は，ここで作成した頻度集計表をデータとしてウォード法で分析したクラスター分析の樹形図です。まず，最初の段階で 2 つに分かれていますが，Brown コーパスと MICASE できれいに分かれ

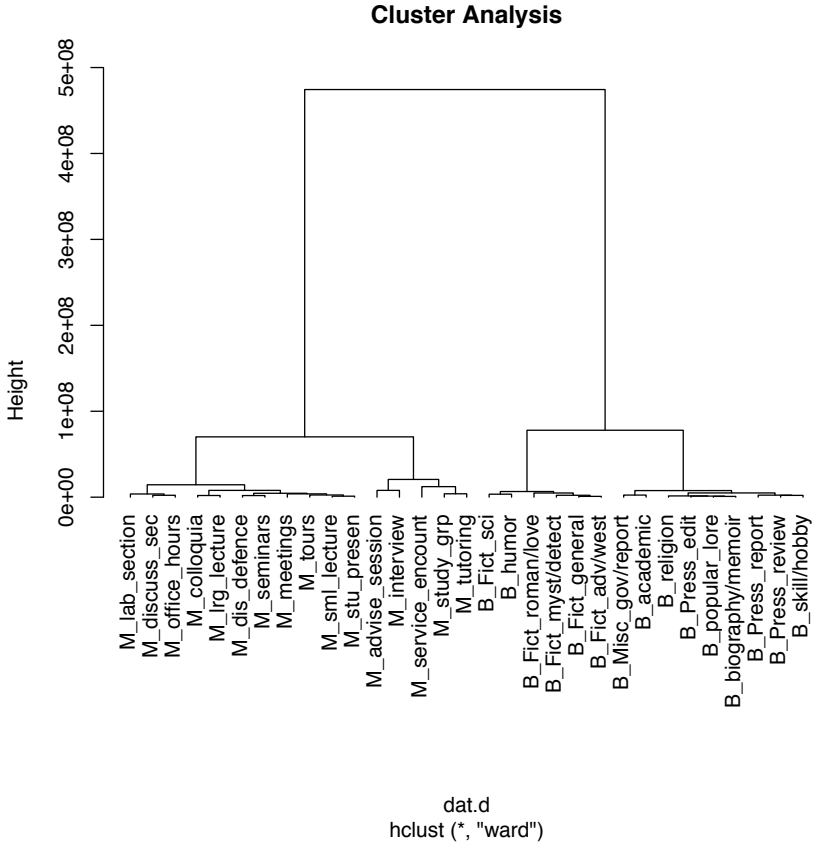


図 3.38 クラスター分析の樹形図

ています。このことから，アカデミックな口語であっても口語は口語であり，文章語とは高頻度語を用いて分類可能であるようだということが推測できます。さらに，Brown コーパスの方を見ていくと，人物描写や会話文が含まれることが多いであろうフィクションと，それ以外のもう少し堅い文章が分類されています。MICASE の方を見ると，レクチャーやセミナーなどアカデミックで双方向性の低いものと，スタディーグループや，ミーティングなど双方向性の高いものが分類されています。このように，基本的な語彙だけをとっても，言語使用域に関して多くのことが見えてきます。

図 3.39 のコレスポンデンス分析の散布図は少し見づらいですが，大きく 3 つ程に分かれているように見えます。一つは MICASE の群，一つは Brown コーパスのフィクションなどの群，もう一つは，Brown コーパスのフィクション以外の固めの文章の群となっています。それに対応する単語が赤い文字でプロットされているのですが，MICASE の近くには，I や you などの，一人称，二人称の代名詞や，this や that などの指示代名詞，Brown コーパスのフィクションの辺りには，he や she などの三人称単数の人称代名詞が多く見られます。Brown コーパスのその他の群のあたりには，フォーマルな場面でしかあまり見られない shall や may などの助動詞や，among や during などの前置詞が見られます。ただ，Brown コーパスは年代が古いので，新しいアメリカ文章語のコーパスでは，shall は同じように見られるかどうかはわかりません。

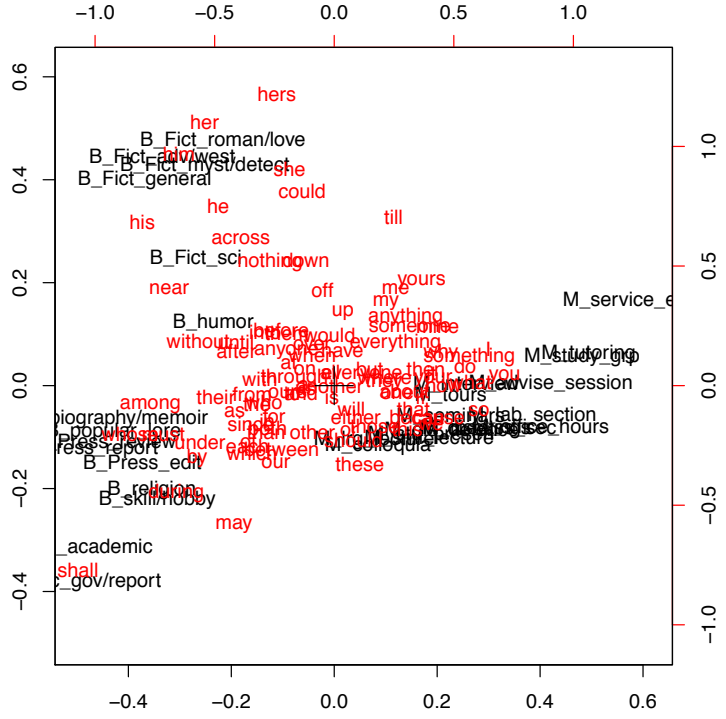


図 3.39 コレスポンデンス分析の散布図

以上，さっと応用的な分析を見てみましたが，この 2 つの分析法から絶対的な何かがわかるかどうかは何とも言えません。ただ，大まかな傾向をつかんだり，さらに詳しく見る価値のある何かが見つかるかもしれません。そのような意味では，簡単に必要なデータが整うということに多少の価値があるのではないかと考えますし，そのような発見に CasualConc が少しでも役立てたらと願っています。

4. 今後の課題

長々と描いてきて，ようやく最後になりますが，現在の課題と今後の予定について簡単に触れておきます。冒頭で書いたように，現行バージョン（1.9.2）の CasualConc は，RubyCocoa という Ruby と Objective-C をつなぐブリッジアプリケーションで書かれています。本稿執筆時点（2012 年 3 月末）では，RubyCocoa 自体に重大なバグがありますが，この問題を回避すべく，バージョン 1.9.2 では標準的ではない方法を試していて，多少の処理速度とメモリ消費を犠牲にしながらも，少なくともクラッシュの頻度は下がっているようです。しかし，この問題が回避できても，RubyCocoa の将来性に不安があるのも確かです。RubyCocoa は現在 Mac OS X Lion に標準でインストールされている Ruby 1.8.7 に依存していますが，Ruby には，これより新しいバージョン 1.9.x がリリースされており，将来的には Mac OS X に標準でインストールされる Ruby も 1.9.x へ移行することが考えられます。そのため，遅かれ早かれ，CasualConc およびその他の RubyCocoa で書いているアプリケーションも別の方法を探らなくてはならなくなっています。

現在，Ruby で Mac OS X アプリケーションを開発する環境には，RubyCocoa の他に，MacRuby というものがあります。これは，RubyCocoa の後継といえるもので，ブリッジの役割しか果たしていない RubyCocoa とは異なり，Ruby 1.9.x で直接 Mac OS X のコアテクノロジーにアクセスするもので，RubyCocoa が抱えている問題を解決するスクリプト言語となっています。ただし，MacRuby にも問題はあって，執筆時点で未だに開発途中であるため多くの問題が解決されておらず，さらに，その特性上，一部の文字列処理の速度が RubyCocoa と比べてかなり遅くなっています。特に，処理速度の低下は，一番利用されるであろう KWIC 検索のときに顕著に現れます。これからもコンピューターの処理能力は上がるでしょうから，将来的には処理速度はあまり大きな問題ではなくなるかもしれませんが，現時点で置き換えるのには速度差が大きすぎます。

また，後継という位置づけではあるものの，移行するには，ほぼ一から書き直す必要があります。現状では，将来を見越して移植は始めているものの，基本的な機能がなんとか動く程度で，現行バージョンとは大きな隔たりがあります。そろそろ，現行バージョンはバグ修正を中心にして，MacRuby 版の開発を進めようと考えていますが，これを機会に増えすぎた機能の整理をし，重要度の高い機能から移植していくつもりです。また，CasualConc 以外のアプリケーションも，順次 MacRuby に移植していきます。そこで，ユーザーの皆様で，よく使われる機能やアプリケーションがあり，早期の移植を望む場合は，連絡いただけると助かります。また，使っている方の声を聞ける（文字を読むだけですが）のは，とても励みになりますので，よろしくお願いします。

と，ここまで書いた後に，Ruby から R を使う方法を見つけて，実験的に CasualConc に組み込むことを試してみました。まずは，コーパスファイル情報ツールの「単語グループ頻度表」機能でできる頻度集計表を使うことを考えたのですが，実は，最後の項で使っているクラスター分析の樹形図とコレスポンデンス分析の散布図は，CasualConc から PDF を書き出したものです。現時点ではオプションなどは一切受け付けず，ただ，頻度表から決めうちのグラフを作るだけですが，これから R の使い方や，コーパスの分析手法などが理解でき次第，もう少しまともな機能にしていきたいとは考えています。その辺りに詳しい方は，情報をいただけると助かります。

いろいろと問題もある CasualConc ですが，使っていただいている方の研究や論文執筆，教材準備などのお力に少しでもなれたらと願っています。

注

[1] http://sites.google.com/site/casualconcj/

[2] ワイルドカード文字はすべて半角記号を使ってください。

[3] https://github.com/nltk/nltk_data/tree/master/packages/corpora にある stopwords.zip をダウンロードして使うことができます。

[4] 最新の Ruby 1.9.x では文字ごとの処理になっていますが，RubyCocoa が動く 1.8.7 ではバイトごとになっています。

[5] http://homepage.mac.com/bncweb/manual/bncwebman-collocation.htm

[6] 要望があれば，もう少し柔軟にリストの書式に対応できるように改良するかもしれません。

[7] この他にも精度はわかりませんが Mac App Store などに OCR アプリケーションがあるかもしれません。

[8] http://sites.google.com/site/casualconcj/yutiriti-puroguramu/casualtextractor から入手できます。

[9] http://www.mext.go.jp/b_menu/shuppan/sonota/990301/03122602/010.htm

参考文献

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, UK: Longman.

Francis, W. N. and Kucera, H. (1964). *A standard sample of present-day English for use with digital computers*. Report to the US Office of Education on Co-operative Research Project no.E -007. Providence, RI: Brown University.

Simpson, R. C., S. L. Briggs, J. Ovens, and J. M. Swales. (2002) *The Michigan corpus of academic spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.