CasualConcでのアカデミック英語分析
—単語検索からデータの視覚化まで—

今尾康裕

はじめに

CasualConc は，macOS 用のコーパス分析アプリケーションで，学習支援ツール AWSuM (Mizumoto, 2016) の開発にあたり，言語データの分析などの基礎的な部分で利用されている。アプリケーション自体は，2007 年に Mac 用の英語論文用例検索用に開発を始め，その後，様々なコーパス分析機能を追加して現在の形に至っている。開発当初は, Windows 用には有料の WordSmith Tools や MonoConc Pro, 国産の TXTANA，無料の AntConc などがあり，用例検索のみならずコーパス分析のための環境は充実していたが，Mac 用には AntConc がとりあえず動くように移植されていたという程度で，まともな環境は整っていなかった。

本稿執筆時点（2017 年初頭）では，AntConc の Mac 版も十分に使用に耐えるものとなっており, 多くの Mac ユーザーもコーパス分析には AntConc を利用しているのではなかろうか。また，古くからのコーパス分析を行っている研究者には，Mac を使っていながら, コーパス分析だけはいまだに Windows を利用している人もいるかもしれない。そこで，一つの選択肢として CasualConc の利用を提案したい。

用例検索としての CasualConc の利用法に関しては, バージョン 1.x に基づいて，細かな設定などの解説も含めて以前に書いたが (今尾, 2012)，本稿では，2015 年にリリースしたバージョン 2.0 に基づいて，アカデミック英語の分析への応用を中心に利用法の紹介をしていく。また，他のアプリケーションでも使える基本的な機能については簡単な紹介にとどめ，CasualConc のユーザーにもあまり利用されていない機能や，統計環境 R との連携で頻度集計した結果を統計処理し視覚化する機能を中心に紹介していく。

1. 基本設定

1.1 ファイルの扱い

CasualConc には，テキストデータを扱うモードとして，AntConc などのようにファイルのリストを作成して分析したり，テキストボックスにテキストをコピー&ペーストしてその分析をする「シンプル」モードと，ファイルのグループを作りコーパスとして管理しながら利用したり，テキストをパラグラフごとに分割して抽出しデータベースに登録して利用する「アドバンスト」モードがある。普段の使用では，この「アドバンスト」モードを中心に利用することになるが，「アドバンスト」モードにはさらに 2 つのモードがあり，詳細な設定をしながら分析するためには「ファイル」モード，分析の精度よりも速度が優先される用例検索には「データベース」モードを選択することが適切である。本稿での例は，すべて「アドバンスト」モードの「ファイル」モードを利用して，ファイルのグループであるコーパスを作成してテキストファイルの処理を行なっている。

1.2 分析のためのテキスト処理

CasualConc では，テキストの分析をするにあたって，単語の扱いや，読み込んだテキストを分析する前段階の処理に様々なオプションがある。細かな設定方法に関してはサイトから入手可能なマニュアルに譲るが，ここでは，どのような機能があるかを紹介しておく。

コーパス分析ツールなどで単語などの頻度を集計する際には，単にテキストに現れる文字列を集計するだけではなかなか見えてこなかったり，誤った分析結果に至ったりする可能性がある。そこで，CasualConc には，他のコンコーダンサーでもよく見られる単語のレマ処理やストップワード処理に加えて，OS の機能や外部アプリケーションを利用しての POS (Part of Speech) タグ付与, 指定文字列や複合語などの単語としての処理，異綴りの処理などの機能を備えている。

このうち，英語のレマ処理 (e_lemma.txt)，異綴り処理 (a-e spelling differences.txt)，ストップワード処理に関しては，サイトからダウンロードした際のディスクイメージにリストを含めてあるので，それらを読み込んだ上で編集して利用してもらいたい。英語以外の言語に関しては，必要に応じてリストを作成した上で読み込んで利用してもらいたい。複合語などのリストは，CasualConc 上で作成するか，各自で用意してもらう必要がある。

POS タグの処理は, あらかじめ決められた形式の POS タグが付与されたテキス

トを読み込んで処理するだけでなく，タグの付与されていないテキストに macOS の言語分析機能や TreeTagger (Schmid, 1994) を利用して POS タグを付与した上で分析できる。TreeTagger を利用するには，TreeTagger のサイトからファイルをダウンロードしてターミナルからインストールするか，CasualConc 上のインストーラを利用してインストールした上で（図 1）分析に使用する。

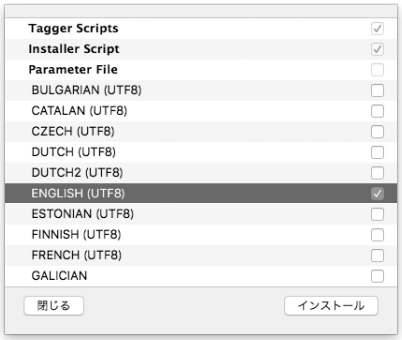


図 1 TreeTagger インストーラ

また, 使用環境に日本語形態素分析エンジン MeCab がインストールされていれば，形態素分析もしくは分かち書きした上で日本語を分析できる。MeCab がインストールされていない場合，もしくは，日本語以外で単語ごとの分析に分かち書き (tokenization) が必要な言語（中国語，タイ語など）も，macOS の標準機能を利用して分かち書きしてから処理することが可能である。

2. 基本ツール

2.1 Concord

Concord は，いわゆる KWIC 検索をするためのツールで，検索語を指定して検索すると結果がテーブルに表示される。検索にはワイルドカードを使った通常の「単語」検索以外に，正規表現での検索や POS タグでの検索モードがある。

この通常検索以外にも，一部のコーパス分析アプリケーションで可能な，指定した範囲の文脈に特定の文字列が現れる場合のみを検索することができる。使用に際しては，検索語を入力するテキストボックス左下のチェックボックスにチェックを入れ，表示されるテキストボックスに文脈に現れる文字列を入力する。通常は, 左右 5 単語以内で範囲指定するが, 環境設定で広範囲モードを選択すると，

左右 15 単語までの指定が可能になる。

図 2 の例は，北米英語の書き言葉コーパスである FROWN (Mair, 1999) コーパスを使って，文脈語として「?ly」を指定して，論文などでよく使われる伝達動詞 argue のレマ検索を行ったものである。単語検索では半角の「?」は任意の一文字以上のワイルドカードとして扱われるので, -ly で終わる単語が指定した範囲（L5–R5）に現れるものだけが結果表示され，文脈語として指定した文字列が下線で強調されている。

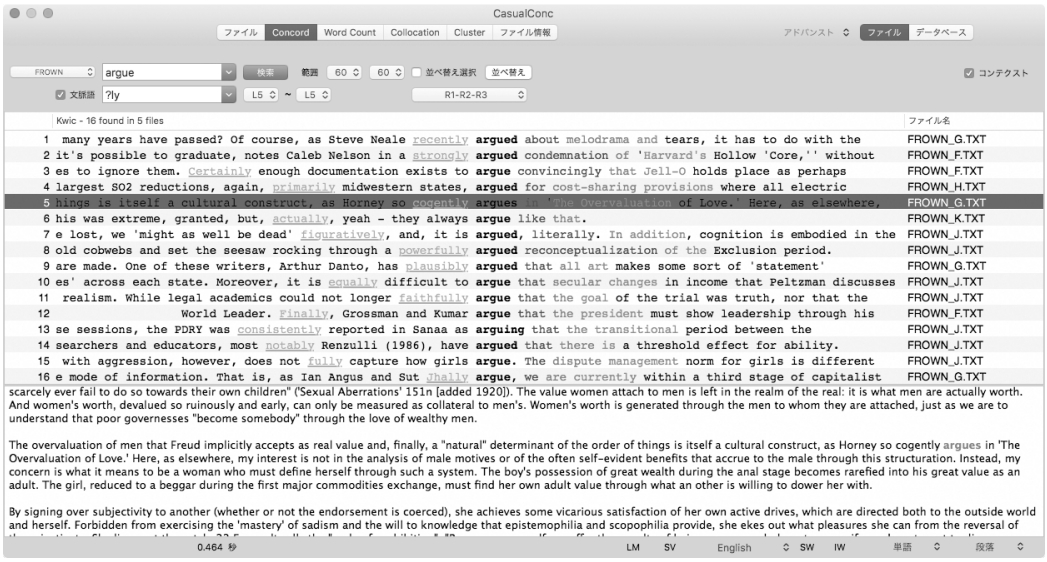


図 2 Concord での文脈語制限（argue + ?ly）

この他，Concord では，結果を別ウインドウで表示して，異なる語の検索結果や，同じ語の異なるコーパスでの検索結果を比較することが可能である。また，Concord の検索結果から，後述する Collocation の集計を行うこともできるため，Concord で検索した語を Collocation で検索し直す必要がない。

2.2 Word Count

Word Count は，単語や n-gram のリストを作成するツールである。n-gram は，n 個の連続する単語の塊で，コーパス言語学では lexical bundle (Biber et al., 1999), word cluster (Carter & McCarthy, 2006) などとも呼ばれており，語彙文法分析の重要な分析単位となっている。集計されたリストには，素頻度の他にその単語・n-gram が全単語・n-gram に占める割合，その単語・n-gram が含まれるファイル数やその割合，指定単位あたりの標準（相対）頻度（オプションで指定）などが表

示される。また，CasualConc の Word Count ツールでは，作成したリストを表示するテーブルが左右に 2 つ並んでいるため，2 つのコーパスの単語リストや同じコーパスの単語リストと n-gram などを同時に表示して比較することができる。ここで, 右テーブルに参照コーパスの単語リストを読み込むか CasualConc 上で 2 つの単語リストを作成して，左テーブルの単語リストの特徴度指数 (keyness) 統計を計算できる。特徴度指数の計算に使える統計値は，対数尤度比，カイ二乗，補完類似度 (complementary similarity measures: CMS), McNemar, %DIFF, 対数比 (log ratio) がある。その他の統計値としては，2-gram リストでコロケーション統計値を計算する機能も付いている。

Word Count では，1.2 項で紹介したように, macOS の言語分析機能（英語のみ）もしくは TreeTagger を利用して POS タグを付与しての分析もできる。また，CasualConc のレマリストを読み込んでのレマ化処理の他に，macOS や TreeTagger のレマ処理機能を使ったレマごとの頻度リストを作成することもできる。図 3 は，FROWN コーパスを用いて，TreeTagger で POS タグ付与をした単語頻度表（左）とレマ処理と POS タグ付与をした単語頻度表（右）を並べたものである。リストには，標準化した頻度も表示されているが，FROWN コーパスはほぼ 100 万語であるため，素頻度と標準化した頻度に大きな差がない。

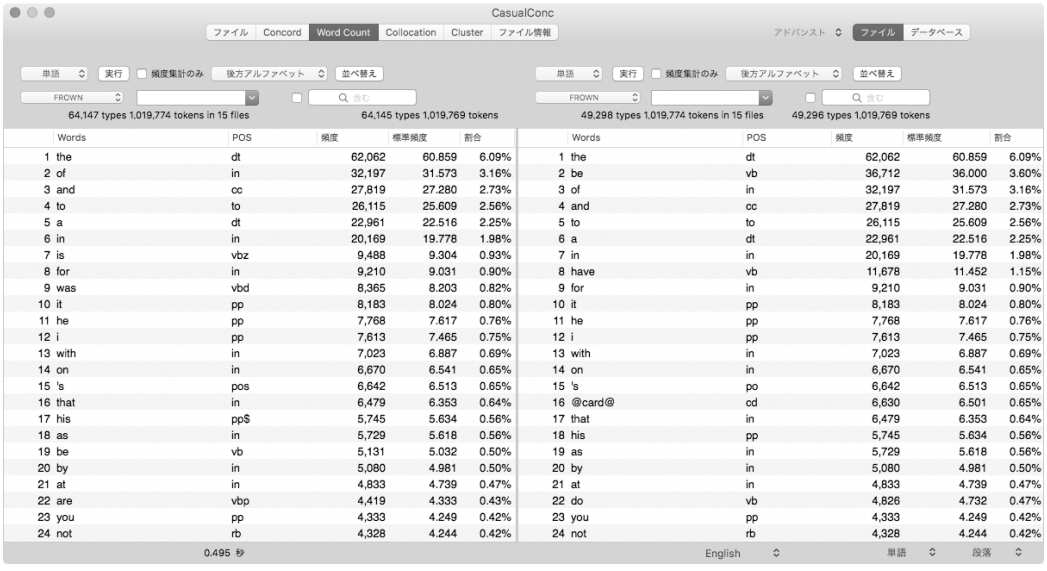
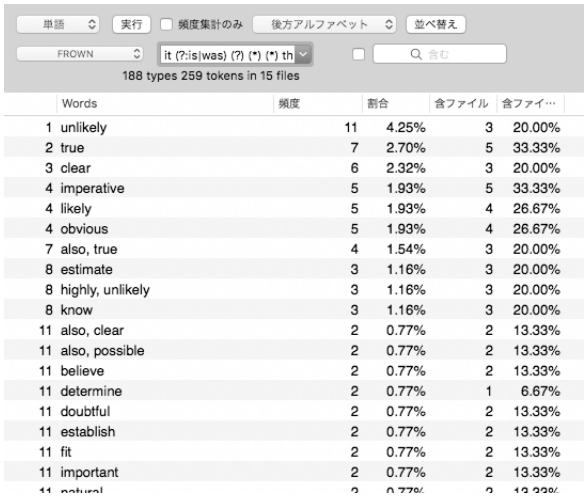


図 3 POS 処理した単語（左）およびレマ化した単語リスト（右）

環境設定で Word Count の「アドバンスト」モードに入ると，指定した文字列を検索してリストを作成することや，n-gram の一つの単語をギャップとして，ギャ

ップ n-gram，いわゆる phrase frame のリストを作成することもできる。図 4 は，FROWN コーパスを用いて，指定文字列検索で it (?:is|was) (?) (*) (*) that という文字列を検索した結果である。この指定文字列検索では，()を検索文字列で使用すると，その中の単語だけでリストが作成される。(?:)を使うと，その中の単語はリストに含まれず，|で OR 検索ができるので，一つ目の単語が it，二つ目の単語が is もしくは was，その次が(?)で，?は一文字以上の任意の文字列にマッチしてリストに含まれ，(*)の*はゼロもしくは一文字以上の任意の文字列なので，最後の that までの間に 1–3 単語が含まれる it is ... that もしくは it was ... that の...の部分のリストができる。この結果のリストの上位 5 つは，unlikely，true，clear，imperative，likely となっており， 7 番目には also true，8 番目には highly unlikely いう 2 単語が入っている。このように，ワイルドカード文字を利用することで，phrase frame で使用される単語・n-gram のリストを作成でき，学術論文とその他のジャンルでギャップ位置での出現頻度を比較して，特定の表現で使われる単語の分析が可能となる。



	Words	頻度	割合	含ファイル	含ファイ…
1	unlikely	11	4.25%	3	20.00%
2	true	7	2.70%	5	33.33%
3	clear	6	2.32%	3	20.00%
4	imperative	5	1.93%	5	33.33%
4	likely	5	1.93%	4	26.67%
4	obvious	5	1.93%	4	26.67%
7	also, true	4	1.54%	3	20.00%
8	estimate	3	1.16%	3	20.00%
8	highly, unlikely	3	1.16%	3	20.00%
8	know	3	1.16%	3	20.00%
11	also, clear	2	0.77%	2	13.33%
11	also, possible	2	0.77%	2	13.33%
11	believe	2	0.77%	2	13.33%
11	determine	2	0.77%	1	6.67%
11	doubtful	2	0.77%	2	13.33%
11	establish	2	0.77%	2	13.33%
11	fit	2	0.77%	2	13.33%
11	important	2	0.77%	2	13.33%
11	natural	2	0.77%	2	13.33%

図 4 指定文字列検索

Word Count の「アドバンスト」モードで n-gram を選択すると，2-gram では，指定した範囲もしくは位置の単語との 2-gram，3-gram 以上では，「ギャップ」オプションが現れて，phrase frame のリストが作成できる。ギャップ n-grams のリスト作成では，n-gram のすべて位置の単語をギャップの対象にする「フル」と中間位置の単語のみをギャップの対象にする「p フレーム」オプションが選べる。さらに，詳細オプションにチェックを入れると，ギャップの位置に現れる単語も集

計でき（図 5），ギャップ位置に現れる単語は別のパネルでリストとして表示することもできるが，大量のメモリと処理時間が必要となるため，ギャップ n-gram のリスト作成後に指定文字列検索でギャップ位置の単語のリストを作成するのが現実的かもしれない。また，2017 年 1 月現在のバージョン 2.0.6 では，ギャップの位置の単語は 1 単語のみであるが，将来的には，複数語のギャップも集計できるようにしたい。現状では，これらの phrase frame のギャップ位置に複数のワイルドカード文字を使って指定文字列検索することで，ギャップ位置に現れる単語の詳細な分析は可能である。

	4-grams	頻度	割合	含ファイル	含ファイ…	
1	the * of the	2,305	0.06%	15	100.00%	end (61), rest (51), middle (27),…
2	in the * of	969	0.03%	15	100.00%	middle (40), case (34), context…
3	the * of a	572	0.02%	15	100.00%	form (13), end (12), size (9), exi..
4	to the * of	476	0.01%	15	100.00%	point (11), end (7), developmen…
5	of the * of	475	0.01%	15	100.00%	department (6), division (6), uni…
6	of the * and	420	0.01%	15	100.00%	car (5), day (4), kind (4), world (..
7	the * and the	396	0.01%	15	100.00%	poet (9), press (5), man (4), ma…
8	on the * of	387	0.01%	15	100.00%	basis (21), part (15), use (10), b..
9	at the * of	362		14	93.33%	end (46), university (30), begin…
10	one of the *	350		15	100.00%	most (37), few (17), first (14), b…
11	and the * of	335		15	100.00%	rest (8), number (6), developme..
12	* one of the	325		14	93.33%	was (34), is (27), as (15), of (13..
13	for the * of	302		15	100.00%	sake (18), rest (15), purpose (1…
14	the * in the	295		15	100.00%	man (6), machine (5), one (5), c..
15	a * of the	292		15	100.00%	member (22), part (15), copy (9…
16	* the united states	283		12	80.00%	in (92), of (36), to (17), between..
17	the united states *	279		12	80.00%	and (33), in (18), as (11), has (8..
18	with the * of	255		15	100.00%	help (9), exception (8), rest (7),…

図 5 ギャップ 4-gram リスト

表 1

FROWN コーパスの高頻度 Phrase Frames（上位 10）

Phrase Frames	頻度
the * of the	2305
in the * of	969
the * of a	572
to the * of	476
of the * of	475
of the * and	420
the * and the	396
on the * of	387
at the * of	362
one of the *	350

表 1 は，図 5 で作成した FROWN コーパスの頻度上位 10 位までのギャップ 4-gram のリストである。このリストのギャップ 4-gram は，ほとんどが前置詞と冠詞などの機能語の組み合わせとなっており，再頻出の the * of the のギャップの位置に現れる頻度上位 5 単語は，end (61)，rest (51)，middle (27)，beginning (25)，center (24)となっている。ここでは，ジャンルごとの比較はしていないが，ギャップ n-gram 自体だけでなく，ギャップの位置に現れる単語などをジャンルごとに比較することで，学術論文と小説や報道などの文章との違いを検証でき，EAP などの教材作成などにも応用可能である。また，ここで作成した頻度リストなどをもとに，R と連携した頻度情報の視覚化も CasualConc 上で可能である。

2.3 Collocation

Collocation は，検索語の左右の指定範囲内の位置に現れる単語の頻度を集計した表を作成することができる。AntConc の Collocation 機能は，左右と合計の頻度のみが表示されるだけであるが，CasualConc の Collocation ツールでは，WordSmith Tools と同様に，それぞれの位置での頻度も集計される。左右の範囲は，L10 から R10 までの指定が可能である。

Collocation では，同じコーパスで単語リストを作成することにより，簡単な視覚化ができるようになっている。図 6 は，応用言語学の学術論文コーパスを用いて，argue/argued/argues の L5–R5 の範囲でコロケーションを集計して視覚化したものである。ここでは，文字の大きさを右上の統計値で，色のバランスを指定した統計値の組み合わせで表示できる。

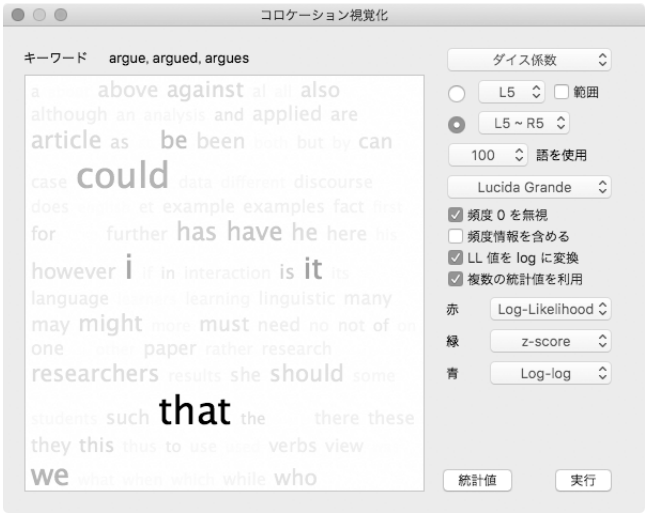


図 6 Collocation の視覚化

この他にも, 環境設定で Experimental Features を使えるようにすることで, Word Chain の機能を使うことができる。このツールでは，検索語（キーワード）の左右独立してではあるが，指定した単語ごとにその左右の次の位置に現れる単語の頻度リストが表示される。例えば，応用言語学の学術論文コーパスで argue/argued/argues で検索した場合には，図 7 のような表示になる。ここではキーワードは複数表示されているが，Collocation の集計時に区別していないため，どの語を選んでも結果に変化はない。まずは，左（L1）で, have (76) を選択すると，その左は have argued の一つ左に現れる単語のリストが表示される。上位 3 つは i (29)，we (12)，#（8: 数字）の順となっており，i (29) を選択すると，さらにその左に，i have argued の左に現れる単語が頻度順に並んでいる（as [5]，and [2]，for [2]）。キーワードの右も頻度表の単語をクリックすることで，その右に現れる単語のリストが表示されていく。これで，共起語がただ単にそれぞれの位置に現れるだけでなく，どの単語と結びついて現れるかも確認できる。

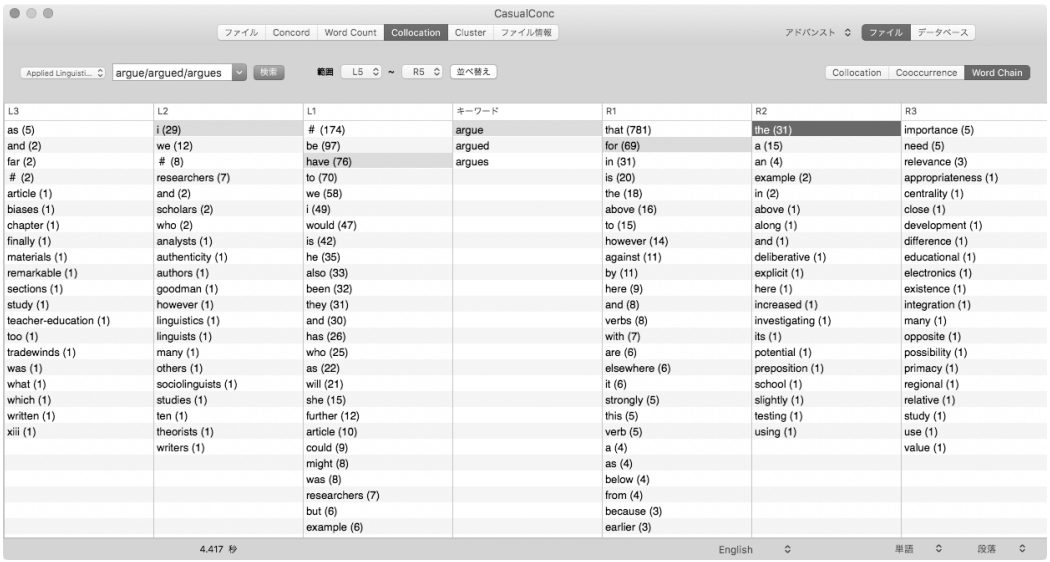


図 7 Collocation の Word Chain

2.4 Cluster

CasualConc の Cluster は，指定した検索語を含む n-gram のリストを作成するツールである。2 から 9-gram までを選択してリストを作成する。また，検索語で始まる n-gram，検索語で終わる n-gram のリストも作成できる。

2.5 ファイル情報

ファイル情報は,それぞれのファイルに含まれる単語数や Type/Token 比などの基本情報を集計するだけでなく，ファイルごとやコーパスごとの単語・n-gram を集計して頻度表を作成することや，TF-IDF の集計表，指定文字列（キーワードグループ）の頻度表，指定した単語同士の共起頻度表などが作成できる。指定文字列は，単に単語や正規表現でマッチする文字列の集計だけでなく，任意の単語や正規表現をグループとしてまとめて集計することもできる。TF-IDF は，それぞれのファイルもしくはコーパスに特徴的な単語・n-gram を抽出するのに有効である。

図 8 は，FROWN コーパスのサブコーパスごとの法助動詞の頻度を集計した結果である。それぞれの頻度は，数え上げる文字列を指定することで否定形の頻度も含み，100 万語ごとの相対頻度となっている。この表から，分野ごとに法助動詞の使用頻度に違いがあることがわかるが，タグが付与されていないテキストであるため,それぞれの頻度は正確なものではないことは理解しておく必要がある。正確な頻度を集計するには，POS タグがつけられているコーパスをそのタグを認識させた上で集計する必要があるだろう。

Group	Total	can	could	may	might	should	must	will	would
TOTAL	11,272	2,189.00	1,475.00	940.00	640.00	792.00	671.00	1,998.00	2,567.00
FROWN_A	982	151.63	133.66	56.16	57.28	58.41	34.82	330.22	280.80
FROWN_B	910	297.26	128.45	89.91	49.54	207.35	88.08	462.40	346.80
FROWN_C	277	175.12	96.31	75.88	70.05	49.62	20.43	175.12	145.93
FROWN_D	426	285.15	104.75	122.21	81.47	66.92	107.66	238.59	232.77
FROWN_E	711	302.33	86.58	71.46	41.23	75.58	50.85	199.27	149.79
FROWN_F	1,147	306.66	120.81	146.62	67.11	98.09	72.28	173.46	199.28
FROWN_G	1,407	130.38	132.36	88.68	72.80	53.61	56.92	93.98	302.44
FROWN_H	720	156.19	65.76	177.56	19.73	110.16	154.55	333.76	166.06
FROWN_J	1,734	245.52	94.24	167.40	48.98	92.38	78.74	180.42	167.40
FROWN_K	658	214.16	251.85	30.84	104.51	49.69	56.54	138.78	280.98
FROWN_L	602	257.23	273.82	22.82	116.17	64.31	47.71	120.32	346.43
FROWN_M	158	157.93	365.72	16.62	58.18	49.87	41.56	157.93	465.46
FROWN_N	606	183.31	229.57	23.98	66.82	46.26	32.55	167.89	287.82
FROWN_P	706	207.45	264.03	34.29	58.29	65.15	66.86	145.73	368.61
FROWN_R	228	182.90	188.44	11.08	94.22	49.88	83.13	110.85	543.15

図 8 FROWN コーパスの分野ごとの法助動詞の頻度

この法助動詞の例と同様に，POS タグなどがついていなければ正確な頻度の集計はできないが，簡易的な集計として，Hyland (2005) のメタ談話標識などの使用頻度を集計してジャンルごとの使用頻度の比較なども可能である。その際は，大文字小文字の区別や，コンマが後に続く場合などのみ集計するなど，メタ談話標識としての使用以外の使用例を排除する工夫が必要である。すでにメタ談話標識のタグ付けがされている場合は，そのタグを数えることでより正確な頻度集計が

可能となる。

2.6 語彙プロファイラ

語彙プロファイラは，基本ツールではないが，実験的なツールとしてバージョン 2.0 で追加した機能である。語彙プロファイラツールは，あらかじめ用意した複数の基準となる単語リストなどを読み込み，Word Count の単語リストや，ウインドウ下部のテキストエリアにペーストしたテキストの単語のうち，どれほどの割合が基準となる単語リストに含まれているかを集計する。CasualConc のディスクイメージには，General Service List (GSL) と Coxhead (2000) の Academic Word List (AWL) が読み込めるフォーマットで用意してあるので，英語の分析では読み込んで利用してもらいたい。

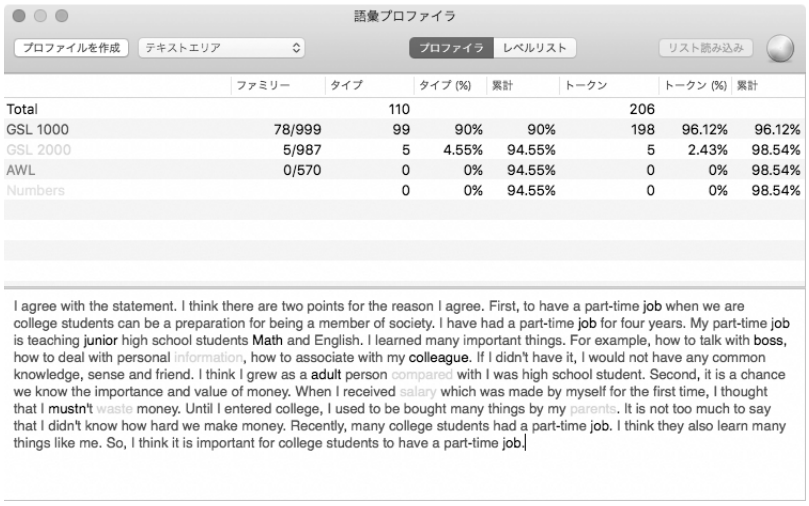


図 9 学習者エッセイの語彙プロファイル（Japanese-B1-1）

図 9 の例では，GSL と AWL を読み込んだ上で，ICNALE (Ishikawa, 2011) の B1-1 レベルと分類された日本人学習者のエッセイを一つ選んでテキストエリアにペーストし，語彙プロファイラを実行した結果である。総語数 206 語のうち，96.12%（198 語）が GSL の 1000 語レベルの単語で，2.43%（5 語）が GSL の 2000 語レベルの単語であり，AWL に含まれる単語は使われていなかった。また，ここに示した図では確認できないが，この分析を行うと，テキストエリアのテキストがあらかじめ単語リストごとに指定した色で色付けされる。図 10 は，同じトピックで書かれた B2 レベルと分類された日本人学習者のエッセイのプロファイルである。総語数が多い（295 語）だけでなく，GSL1000 語レベルが 90.17%（266 語），

GSL2000 語レベルが 6.44%（19 語），AWL が 1.36%（4 語）と，あらかじめ分類されたレベルの差が書かれたエッセイの語彙使用にも表れているのがわかる。

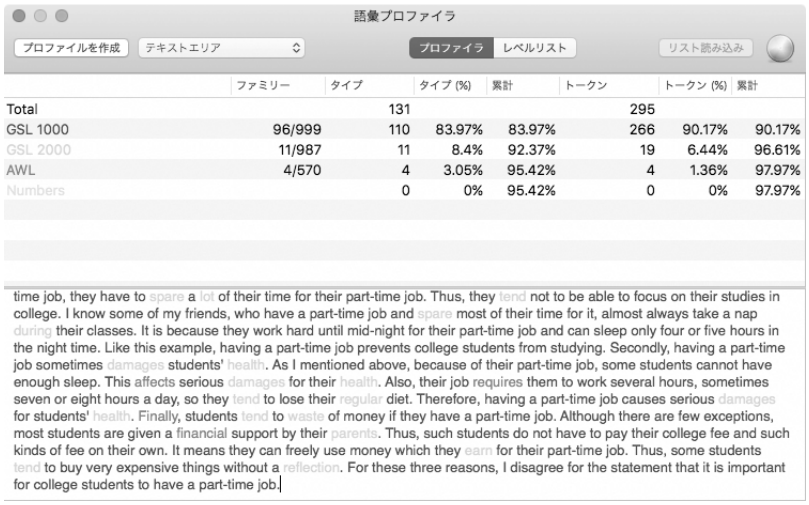


図 10 学習者エッセイの語彙プロファイル（Japanese-B2）

ここまで見てきたように，CasualConc には，単に KWIC 検索をしたり，単語・n-gram リストを作成したりする以上の機能が備わっており，アカデミック英語の表面的な分析だけでなく，一歩踏み込んだ分析が可能となっている。次の項では，頻度集計した結果を統計アプリケーション R と連携して視覚化する機能を紹介する。

3. R との連携機能を使った頻度情報の扱い

CasualConc のバージョン 2.0 以降では，統計環境 R を利用して頻度情報の統計処理や視覚化を行うことができる。この機能を利用するためには，別途 R をインストールした後に CasualConc を起動し，環境設定で利用可能にする必要がある。初めて R との連携機能を利用する場合や，CasualConc のアップデートによって新たな R のライブラリを利用することになる場合は，利用開始時に必要なライブラリを自動でダウンロードしてインストールするため，インターネットにアクセスできる必要がある。2017 年 1 月現在では，最初の利用の際に 18 のライブラリをインストールするが，すでに R を利用している場合はこの限りではない。

現在のバージョンでは，Word Count やファイル情報で作成した頻度リスト・頻

度表を利用しての統計処理や，多変量解析を用いて視覚化する機能を多数備えているが，スペースの都合上，本稿では一部の紹介にとどめる。

3.1 ファイルごとの頻度情報を使った特徴語抽出

CasualConc の Word Count も含めて，多くのコンコーダンサーでは，2 つのコーパスを比較して一方のコーパスでより多く出現する単語・n-gram，いわゆる特徴語を統計処理で識別する機能を持っている。その際に使われるのは，コーパス全体の頻度リストとカイ二乗検定（χ^2）や対数尤度比（Log-likelihood）などの統計値が一般的である。しかしながら，この方法はコーパス全体の頻度を利用するため，全体の頻度リストしかない場合でも特徴語の抽出が可能ではあるが，コーパスの一部のファイルにしか出現しない単語などでも頻度が高ければそのコーパス全体の特徴語として扱われてしまうことや，頻度が低くても一方にしか現れなければ統計値が高くなってしまうこと，一方を規模の大きな参照コーパスとする前提があることなどが問題になる。これらの問題に対処するため，コーパスに含まれるファイルごとの頻度情報を利用する特徴語抽出の方法として，CausalConc のバージョン 2.0 以降には，マン・ホイットニー (Mann-Whitney) 検定とランダムフォレスト (Random Forest) を利用した，2 つもしくはそれ以上のコーパスからの特徴語抽出機能が試験的に追加してある。

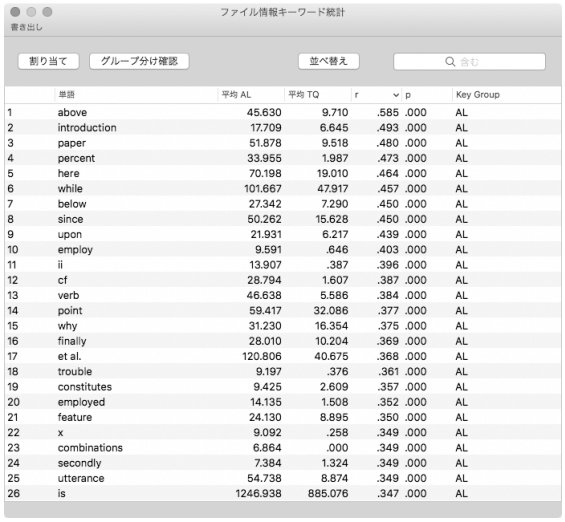
3.1.1 マン・ホイットニー検定を利用した特徴語抽出

マン・ホイットニー検定は，ノンパラメトリック，つまり，母集団に対して正規分布を仮定しない検定で，パラメトリックでの対応のない平均値の差の検定である *t* 検定に相当する (竹内・水本, 2014)。CasualConc では，ファイル情報の単語頻度機能でファイルごとの単語もしくは n-gram 頻度を集計し, 頻度集計されたすべての単語・n-gram ごとに検定をかける。頻度集計は素頻度ではファイルごとの総悟数に影響されるため，一定語数ごとに標準化した相対頻度で集計する必要がある。当然のことながら，マン・ホイットニー検定は，2 集団の比較を行う検定であるため，扱うことのできるコーパスの数は 2 つである。

まずは, ファイルビューで比較したいコーパスを 2 つ選び, ファイル情報の「単語頻度」でグループ分けを「ファイル」に設定して頻度表を作成する。この際，レマ，英米などの綴りの違い，ストップワードなどは，分析の目的に従ってあらかじめ設定しておく必要がある。頻度表作成後，メニューの統計からファイル情報の「順位／平均比較」を選び，ファイルを 2 つのグループに分けて実行する。

オプションで，*t* 検定（Welch の検定）も選べるが，通常は Mann-Whitney-U を選択したままにしておく。

結果テーブルには，それぞれの平均値と効果量 (*r*)，*p*，および平均値の大小に基づくグループへの割り当て (Key Group)，つまり，どちらのグループの特徴語になるのかが表示される（図 11）。



	単語	平均 AL	平均 TQ	r	p	Key Group
1	above	45.630	9.710	.585	.000	AL
2	introduction	17.709	6.645	.493	.000	AL
3	paper	51.878	9.518	.480	.000	AL
4	percent	33.955	1.987	.473	.000	AL
5	here	70.198	19.010	.464	.000	AL
6	while	101.667	47.917	.457	.000	AL
7	below	27.342	7.290	.450	.000	AL
8	since	50.262	15.628	.450	.000	AL
9	upon	21.931	6.217	.439	.000	AL
10	employ	9.591	.646	.403	.000	AL
11	ii	13.907	.387	.396	.000	AL
12	cf	28.794	1.607	.387	.000	AL
13	verb	46.638	5.586	.384	.000	AL
14	point	59.417	32.086	.377	.000	AL
15	why	31.230	16.354	.375	.000	AL
16	finally	28.010	10.204	.369	.000	AL
17	et al.	120.806	40.675	.368	.000	AL
18	trouble	9.197	.376	.361	.000	AL
19	constitutes	9.425	2.609	.357	.000	AL
20	employed	14.135	1.508	.352	.000	AL
21	feature	24.130	8.895	.350	.000	AL
22	x	9.092	.258	.349	.000	AL
23	combinations	6.864	.000	.349	.000	AL
24	secondly	7.384	1.324	.349	.000	AL
25	utterance	54.738	8.874	.349	.000	AL
26	is	1246.938	885.076	.347	.000	AL

図 11 マン・ホイットニー検定の結果

表 2

抽出された特徴語（上位 10 単語）

MWU		Log-Likelihood	
Applied Linguistics	*TESOL Quarterly*	*Applied Linguistics*	*TESOL Quarterly*
above	p	complexity	assessment
introduction	students	metaphor	students
paper	teachers	movies	p
percent	tesol	is	teachers
here	assessment	planning	english
while	university	caf	pragmatics
below	develop	grammar	pt-e
since	majority	et al.	programs
upon	skills	te	revision
employ	education	skehan	revisions

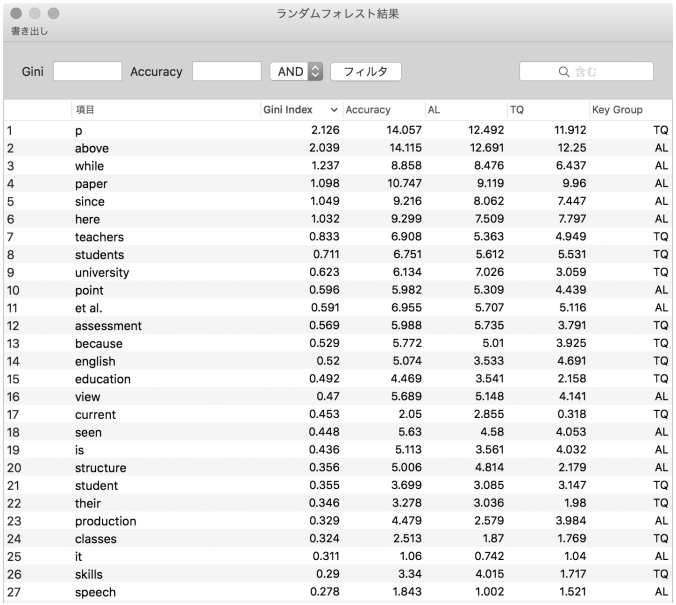
ここでは,応用言語学の専門学術誌である *Applied Linguistics* と *TESOL Quarterly* の 2008–2009 年に出版された号に載った論文を使って特徴語抽出を試みた。表 2 は，マン・ホイットニー検定でのそれぞれのコーパスに特徴的だと識別された上位 10 語と，Word Count の対数尤度比で特徴語と識別された上位 10 語を比較したものである。両方の抽出方法で共通する語もあるが，単語リスト全体の頻度を利用して対数尤度比で抽出したリストには，一部の論文のみに出現する単語も含まれる一方，マン・ホイットニー検定ではそのような傾向はあまり見られない。マン・ホイットニー検定の結果を見ていくと，*Applied Linguistics* の論文に特徴的な語には，while，since といった従属接続詞や，above，below といった前置詞，introduction など，文体や論文のフォーマットに関連するようなものが現れているが，*TESOL Quarterly* の論文に特徴的な語には，students，teachers，assessment，university，education，skills などの言語教育・習得に直接関連する内容語が多く見られる。これは，*Applied Linguistics* は，応用言語学の幅広い分野の研究論文が掲載されているが，*TESOL Quarterly* には，英語教育という応用言語学の中でも狭い分野の論文が掲載されているため，扱う内容が限定されて内容語が特徴語として現れていることが考えられる。ただし，その要因を確認するためには，実際にこれらの語がどのように使われているかを量的質的の両面からさらに詳しく検証していく必要がある。また、当然のことではあるが，マン・ホイットニー検定も万能ではないため，対数尤度比や，CasualConc に備わっているその他の検定などの結果とも比較した上で，特徴語の抽出を行う必要がある。

3.1.2 ランダムフォレストを利用した特徴語抽出

ランダムフォレストは，Breiman (2001) が提案した集団学習を使ったデータの分類手法であり，規模の大きいデータからのデータマイニングに適しているとされる (金, 2009)。CasualConc では，分類自体の結果を利用するのではなく，あらかじめ指定したグループにファイルが分類される際に寄与する単語・n-gram などの項目を,その寄与率に基づいて特徴語として取り出す方法として利用している。ランダムフォレストを利用した特徴語抽出機能には，マン・ホイットニー検定のように 2 つのコーパスに制限されることはないが，あまり多くのコーパスを利用すると必要なデータの精度や結果の解釈が難しくなるため，3 つ 4 つ程度が適当ではないかと考えられる。また，計算に必要なリソースなどの制限から，分析にかける項目数は 1,000 程度までが適当であるようだ。それより多くの項目を利用

したい場合は，指定語数ごとに分割して処理するオプションを利用するか，あらかじめ語数を絞ってから分析をしてもらいたい。また，利用する項目からランダムにサンプルを取り出して分類するという性質から，項目の寄与率が分析ごとに異なることもあるため，オプションの設定を変更するなり，分析を繰り返すなどして確認する必要がある。

使い方は，マン・ホイットニー検定と同様に，ファイル情報の単語頻度でファイルごとの単語頻度表を作り，メニューから呼び出す。その後の手順はマン・ホイットニー検定と多少異なるが，グループの振り分けをしたのちにオプションを設定して実行する。結果には，ジニ係数と正解率の値およびそれぞれのグラフ，グループ割り当てなどが表示される（図 12）。



	項目	Gini Index	Accuracy	AL	TQ	Key Group
1	p	2.126	14.057	12.492	11.912	TQ
2	above	2.039	14.115	12.691	12.25	AL
3	while	1.237	8.858	8.476	6.437	AL
4	paper	1.098	10.747	9.119	9.96	AL
5	since	1.049	9.216	8.062	7.447	AL
6	here	1.032	9.299	7.509	7.797	AL
7	teachers	0.833	6.908	5.363	4.949	TQ
8	students	0.711	6.751	5.612	5.531	TQ
9	university	0.623	6.134	7.026	3.059	TQ
10	point	0.596	5.982	5.309	4.439	AL
11	et al.	0.591	6.955	5.707	5.116	AL
12	assessment	0.569	5.988	5.735	3.791	TQ
13	because	0.529	5.772	5.01	3.925	TQ
14	english	0.52	5.074	3.533	4.691	TQ
15	education	0.492	4.469	3.541	2.158	TQ
16	view	0.47	5.689	5.148	4.141	AL
17	current	0.453	2.05	2.855	0.318	TQ
18	seen	0.448	5.63	4.58	4.053	AL
19	is	0.436	5.113	3.561	4.032	AL
20	structure	0.356	5.006	4.814	2.179	AL
21	student	0.355	3.699	3.085	3.147	TQ
22	their	0.346	3.278	3.036	1.98	TQ
23	production	0.329	4.479	2.579	3.984	AL
24	classes	0.324	2.513	1.87	1.769	TQ
25	it	0.311	1.06	0.742	1.04	AL
26	skills	0.29	3.34	4.015	1.717	TQ
27	speech	0.278	1.843	1.002	1.521	AL

図 12 ランダムフォレストの結果

表 3 には，マン・ホイットニー検定での例と同じコーパスで，頻度上位 500 語で分析した結果を示す。全体的な傾向としては，マン・ホイットニー検定での結果と同じであるが，いくつか異なる単語も含まれている。これらをさらに検証していくことで，特に内用語よりも機能的な語が特徴語として選別された要因などを探っていくと新たな知見が得られるかもしれない。

表 3

ランダムフォレストで抽出された特徴語（上位 10 語）

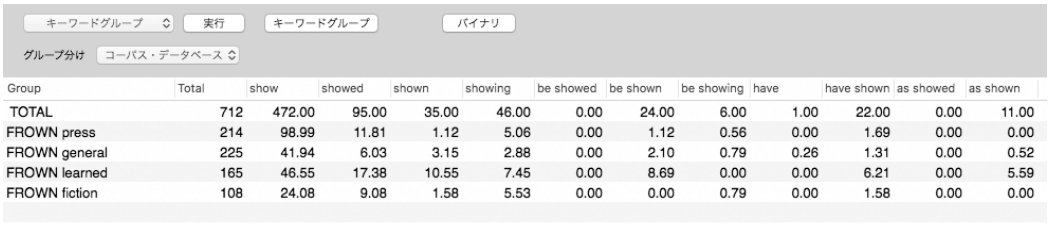
Random Forest	
Applied Linguistics	*TESOL Quarterly*
above	p
while	teachers
paper	students
since	university
here	assessment
point	because
et al.	english
view	education
seen	current
is	student

3.2 頻度データの視覚化

CasualConc の各ツール，特に Word Count とファイル情報で得られた頻度データを利用して，R で視覚化する機能のうち，コーパス分析でよく利用されるものに限って簡単に紹介する。

3.2.1 単純な視覚化

CasualConc では，頻度データを簡単に棒グラフや折れ線グラフにすることができる。以下の例では，ファイル情報のキーワードグループ機能を利用して，FROWN コーパスのサブコーパスごとの動詞 SHOW の頻度集計をした。このサブコーパスの分類は厳密ではなく, Press (A-C), General (D-H), Learned (Academic: J), Fiction (K-R) としてある。showed，shown，showing については，be，have，as との共起の場合（受け身，進行形など）の頻度を，集計のためのリストを指定することで除いてあるが，これらとの間に副詞などが存在する場合は含まれるため，あくまでも簡易的な集計となっている。また，be showed と as showed は存在しなかったため結果に含めていない。頻度は，10 万語ごとに標準化された頻度となっている（図 13）。



Group	Total	show	showed	shown	showing	be showed	be shown	be showing	have	have shown	as showed	as shown
TOTAL	712	472.00	95.00	35.00	46.00	0.00	24.00	6.00	1.00	22.00	0.00	11.00
FROWN press	214	98.99	11.81	1.12	5.06	0.00	1.12	0.56	0.00	1.69	0.00	0.00
FROWN general	225	41.94	6.03	3.15	2.88	0.00	2.10	0.79	0.26	1.31	0.00	0.52
FROWN learned	165	46.55	17.38	10.55	7.45	0.00	8.69	0.00	0.00	6.21	0.00	5.59
FROWN fiction	108	24.08	9.08	1.58	5.53	0.00	0.00	0.79	0.00	1.58	0.00	0.00

図 13 ファイル情報での頻度集計

頻度集計の結果は表 4 に示す。show/shows は現在形での使用，showed，shown は，過去形及び分詞としての使用，showing も分詞としての使用，be shown は受動態，be showing は進行形，have showed と have shown は完了形，as shown は，分詞ではあるが，共起する頻度が高い組み合わせとして集計に加えてある。集計結果を見ると，Press では現在形での使用が多いが，Learned（つまり Academic）では受け身の分詞や受動態，完了形での使用頻度が高く，現在形を除いた形で概ね使用頻度が高いことがわかる。

表 4

FROWN コーパスでの動詞 SHOW の使用頻度

	show/ shows	showed	shown	showing	be shown	be showing	have showed	have shown	as shown
Press	98.99	11.81	1.12	5.06	1.12	0.56	-	1.69	-
General	41.94	6.03	3.15	2.88	2.10	0.79	0.26	1.31	0.52
Learned	46.55	17.38	10.55	7.45	8.69	-	-	6.21	5.59
Fiction	24.08	9.08	1.58	5.53	-	0.79	-	1.58	-

この集計結果は，表の数値だけでもある程度傾向はつかみやすいが，グラフにすることでその傾向がより明確になる。CausalConc では，ファイル情報の結果を読み込んで簡単に折れ線グラフや棒グラフを描くことができる。図 14 は、表 4 のデータをそのまま折れ線グラフにしたものである。ただし，今回はあくまでも例として示すために折れ線グラフを描いたが，連続性のないデータに折れ線グラフを使用することの是非については考える必要がある。

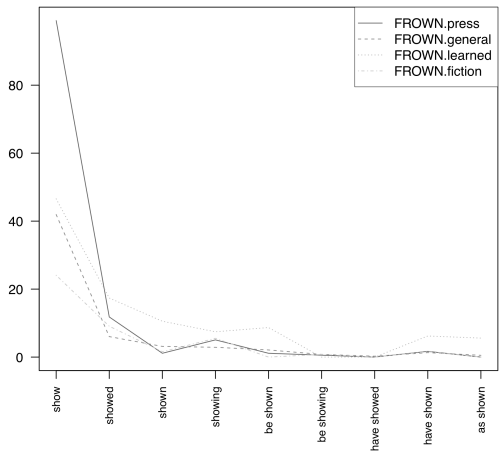


図 14 頻度集計結果の折れ線グラフ

図 15 は同じデータを棒グラフにしたもので，連続性のないデータでは，こちらの方が好ましい。また，サブコーパス間の差異も明確になっている。CasualConc では，この他にパイチャートやレーダーチャートも描けるが，実際に利用する際は，データの性質や提示する目的などを考慮した上でグラフを選択することが重要である。

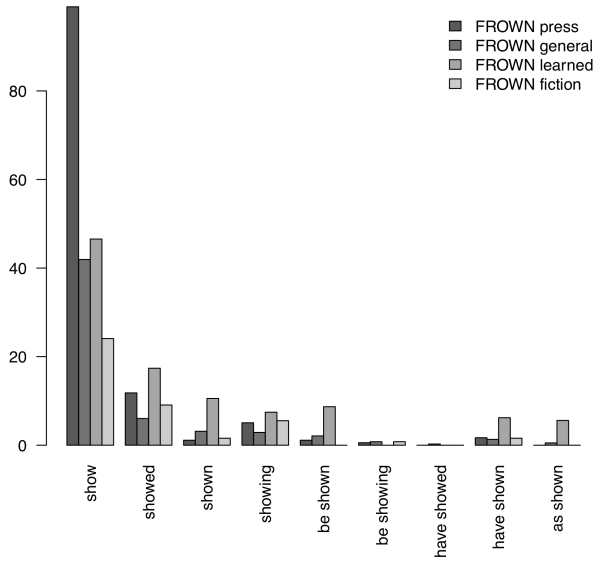


図 15 頻度集計結果の棒グラフ

3.3 テキストデータの特徴の可視化

ここまでの単純な頻度データのプロットによる可視化だけでなく、R のライブラリを活用した高度な可視化も可能となっている。

3.3.1 Word Cloud を使った特徴語の可視化

ウェブ上でも見ることが多い，手軽に可視化ができて視覚に訴える手法の一つに Word Cloud がある。Word Cloud では，頻度情報を元に文字の大きさや色などを割り当てられ，テキストの特徴が一目でわかる図が描ける（図 16）。研究の分析対象として用いるのは難しいかもしれないが，データの特徴を大まかに把握する際や，他の手法やキーワードでの分析をプレゼンテーションなどで提示する前に，データの性質などを説明する補助としては有用であろう。CasualConc では，Word Count やファイル情報ツールでの頻度データを扱うことができる。

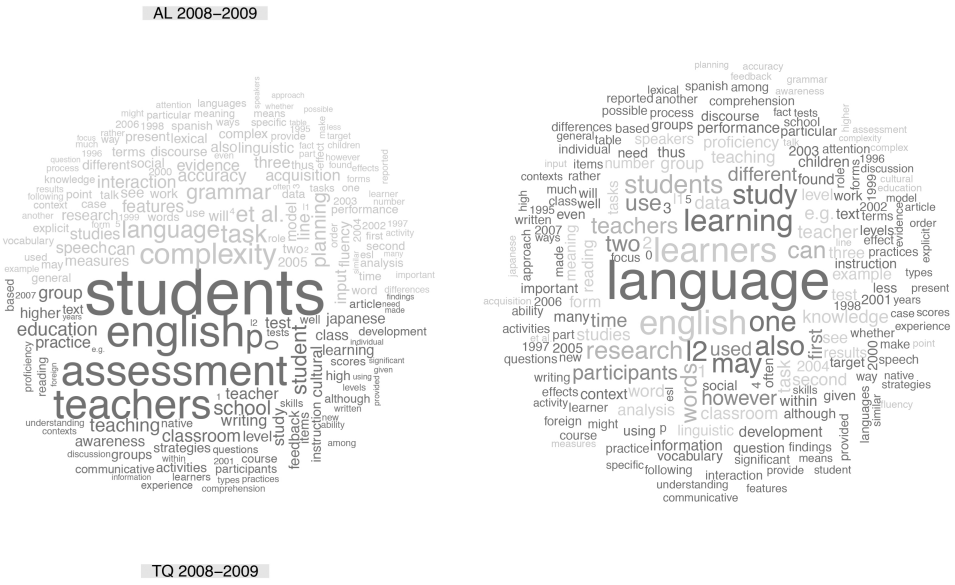


図 16 Word Cloud（左: *Applied Linguistics* と *TESOL Quarterly*; 右: *TESOL Quarterly*）

図 16 は, 上でも扱った応用言語学系の論文誌の 2 誌のコーパスを可視化したものである。いずれも，ファイル情報でコーパスを単位として，キーワード抽出で扱った *Applied Linguistics* と *TESOL Quarterly* の 2 つのコーパスの単語リストをストップワードを除いて作成し，左側は 2 つのコーパスの単語リストを比較して，右側は 2 つの単語リストに共通する単語で Word Cloud を描いたものである。2 つのコーパスを比較したもの（左）を見ると，キーワード抽出の結果を反映するよ

うに，*Applied Linguistics* では特徴語があまり偏っていないため，特に大きな文字で示される単語は見られないが，*TESOL Quarterly* では特定の単語が大きく表示され，強い特徴を示す単語となっていることがわかる。2 つのコーパスに共通する単語の Word Cloud は，どちらにも共通する単語として当然のことながら language が最も特徴的な単語として示されており，learners や English などの研究対象としてよく扱われている単語や，research，study，participants などの研究報告の文章でよく使われる単語が特徴的な語として現れていることがわかる。

3.3.2 ネットワーク分析を用いた可視化

ネットワーク分析とは，グラフ理論に基づく要素の関連性を視覚化する手法で(金, 2009)，テキストデータでは，語と語の関係性を視覚化するのに用いられる。CasualConc では，Word Count の 2-gram リストから語と語の関係性を示すネットワークを，4-gram リストから 2-gram 間の関係性を示すネットワークを，ファイル情報のコロケーション頻度表から単語間のコロケーション情報に基づくネットワークを描くことができる。

図 17 は，上記の 2 つの応用言語学論文コーパスを合わせてストップワードを除いた 2-gram のリストを作りネットワーク図を描いたものである。左下の language を中心に，English，L2，learning，education などの関連する多くの語が繋がり，そのほかでは，statistically-significant-differences や current/present-study といった研究論文で多く見られる語のつながりが現れている。

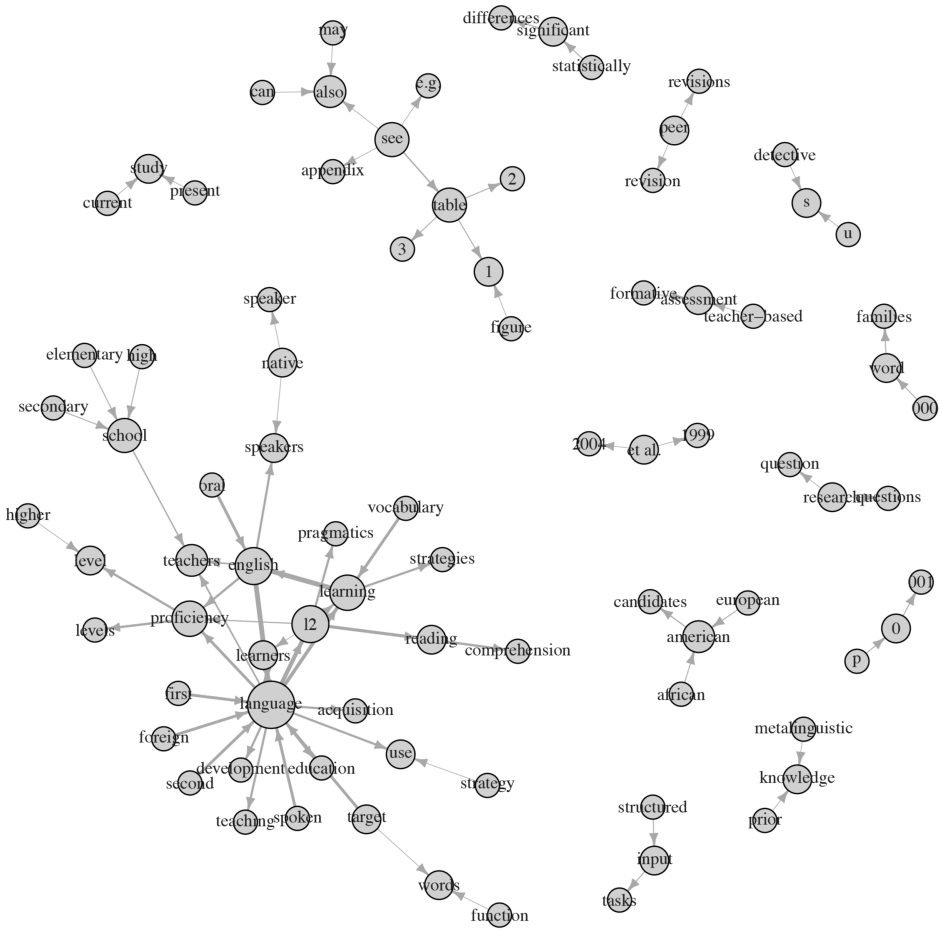


図 17 応用言語学論文コーパスの 2-gram による単語のネットワーク図

図 18 は，同じコーパスでストップワードを排除しないで 4-gram のリストを作り，2-gram のネットワークを描いたものである。左端の of the を中心とした集合には，多くの「冠詞＋名詞」の 2-gram から線がつながっており，名詞が of を使って後置修飾されるというアカデミックな書き言葉の特徴を示している。それ以外には，at the，in the，at the などの「前置詞＋冠詞」の 2-gram から名詞を含む 2-gram へ線がつながっており，その多くは「名詞＋of」になっていることから，ジャンルが異なるコーパスではあるが，表 1 に見られる phrase frame は学術論文でも頻出することがわかる。

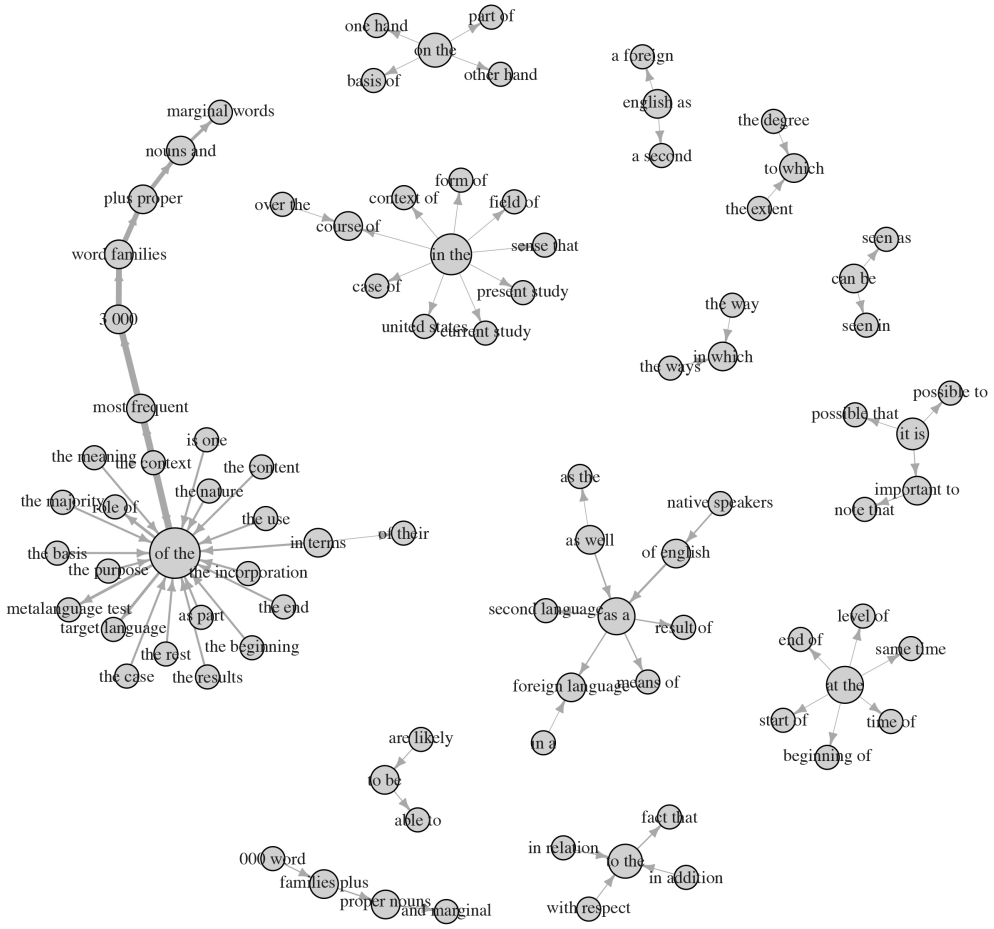


図 18 応用言語学論文コーパスの 4-gram による 2-gram のネットワーク図

これらの例が示すように，Word Cloud やネットワーク図のみで学術的に議論することは難しいかもしれないが，他の分析の結果を視覚的に示す目的や，研究初期段階でテキストデータの大まかな傾向などを把握するためには，大いに利用価値があるのではなかろうか。

3.4 多変量解析を利用したデータの可視化

CasualConc には，ここまでの頻度データをそのまま可視化する手法以外にも，多変量解析を利用して単語やコーパスの関係性を描画するオプションが多く備わっている。利用できる統計解析は，クラスター分析，コレスポンデンス分析，主成分分析，探索的因子分析，多次元尺度構成法，ヒートマップがあるが，ここでは，クラスター分析，コレスポンデンス分析，主成分分析について簡単に説明する。

3.4.1 クラスター分析

CasualConc で利用できるクラスター分析は，一般に階層的クラスター分析と呼ばれる手法で，頻度表のデータ間の距離を元に情報を集約し，類似性の高いものを集めて，デンドログラム（樹形図）を描く (石川・前田・山崎, 2010)。

クラスター分析は多変量の分析であるため，デンドログラムを描くにはファイル情報で頻度表を作成する必要がある。頻度表は，単語頻度作成したファイルやコーパスごとの頻度表だけでなく，キーワードグループで特定の文字列の頻度を集計したものも利用できる。まずはファイル情報で頻度表を作成し，クラスター分析ツールに読み込み描画する。図 19 は，FROWN とそれに対応するイギリス英語のコーパスである FLOB，応用言語学の論文コーパス，Open American National Corpus (OANC) の口語サブコーパス，British National Corpus (BNC) の一部で構成された sampler の口語サブコーパス，北米大学の大学で使われる口語英語を集めた Michigan Corpus of Academic Spoken English (MICASE)，およびアーサー・コナン・ドイルの作品 Sherlock Holmes のコーパスを用いて，一般に入手可能な COCA のレマ頻度リストの上位 100 まで（http://www.wordfrequency.info/free.asp より入手可能）を使って，ファイル情報のキーワードグループで集計した頻度表のデータで描いたデンドログラムである（FROWN と FLOB はそれぞれサブコーパスごとに集計してある）。

ここでは，大きく 2 つのクラスターに分かれているが，上のクラスターはその下に 2 つのクラスターがあると見ることができ，上のクラスターは，MICASE，BNC，ANC の口語コーパスのクラスター，下のクラスターは Sherlock Holmes を含む FROWN と FLOB の Fiction のサブコーパスのクラスターになっており, 登場人物の台詞として口語が含まれる小説と口語のコーパスが近い関係にあることがわかる。そして，下の大きなクラスターもその下に 2 つのクラスターがあると見ることができ，上のクラスターは FROWN と FLOB の報道文書や一般書などがあり，下のクラスターは，FROWN と FLOB の政府文書，学術文章のサブコーパスと応用言語学の論文コーパスのクラスターとそれ以外の FROWN，FLOB のサブコーパスのクラスターに分かれている。そして，FROWN と FLOB に関しては，おおよそ同じサブコーパスとして分類されているテキストは近い位置に分類されていることもわかる。

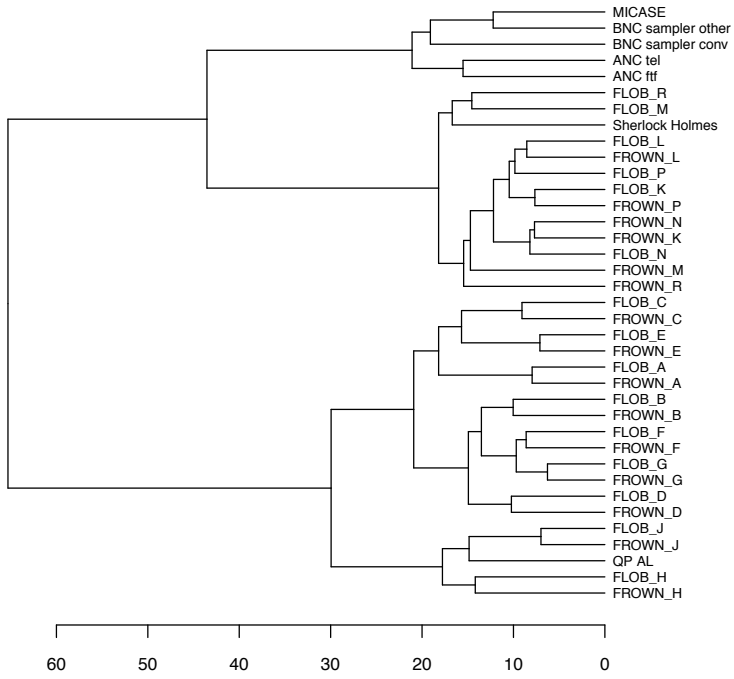


図 19 デンドログラム

ここでわかることは,それぞれのコーパスの特徴語を選別して分類しなくても，高頻度で現れる語でも分類ができるということである。つまり，言語としての英語で高頻度で使われる単語には，高頻度であるがゆえ人間の目には止まりにくいが，ジャンルもしくはレジスターを区別する多くの情報が含まれているということである。当然ではあるが，高頻度で使われる語の中にそれぞれのコーパスやジャンルに特徴的なものも含まれているため，それらが大きく影響している可能性は否定できないので，研究として利用する際は，さらに詳しく分析する必要がある。また，ここではスペースの都合上詳しくは扱えないが，2-gram で同じ分析を行ったところ，同様の結果が得られた。

3.4.2 コレスポンデンス分析

コレスポンデンス分析は，頻度表の行と列の対応関係をその相関が最大となるように並べ換えることで，行列のデータを集約する手法である (石川 et al., 2010)。CasualConc では, ファイル情報の頻度表から項目 (列) とファイル・コーパス (行) の関係を視覚化する機能となっている。そのため，基本的には，ファイル情報の単語頻度もしくはキーワードグループで作成した頻度表のデータを利用する。

ここでの例では，小山・水本 (2010) にならって，クラスター分析の例で用いた話し言葉・書き言葉コーパスの 4-gram の頻度集計を行い，それに基づいてコレスポンデンス分析を試みた。小山・水本では，分析に使用したコーパスから 4-gram を抽出したが，ここでは使用するコーパスが多いこと，4-gram は単語に比べて頻度が低いため含まれるコーパスの影響が出やすいことから，クラスター分析の例と同様に，一般に入手可能な COCA の 4-gram リストの上位 100 までを利用して（http://www.ngrams.info/download_coca.asp より入手可能），ファイル情報のキーワードグループで FROWN と FLOB についてはそのサブコーパスファイルごと，その他はそれぞれのコーパスごとの頻度を集計した。その後，コレスポンデンス分析ツールにデータを送り分析を行った。

図 20, 21 はコレスポンデンス分析の結果の第 1 次元と第 2 次元の値をプロットしたものである。通常は，行項目と列項目を同一平面上にプロットしてその関係性を見るが，視認性をよくするため，ここでは別々にプロットしている。

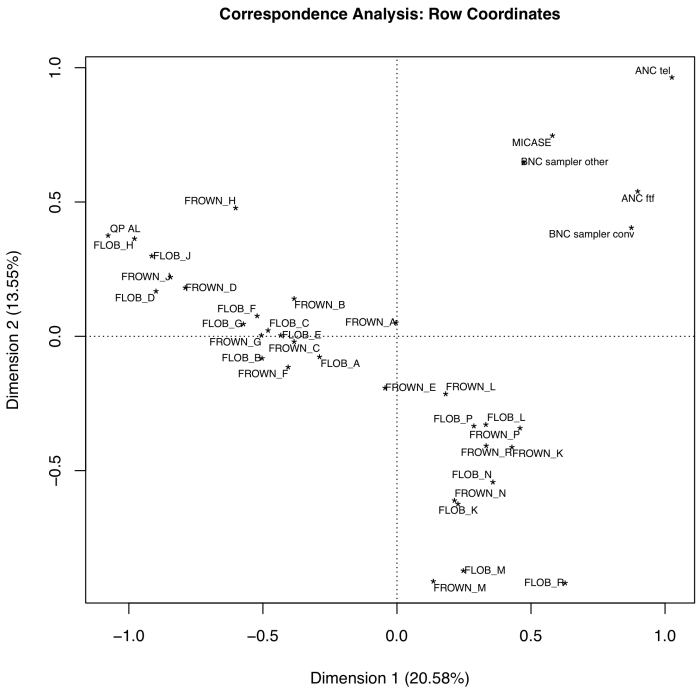


図 20 コレスポンデンス分析（行項目）

行項目のプロットである図 20 を見ると，左端中央やや上に応用言語学の論文コーパスや，クラスター分析で同じクラスターに分類された，FROWN や FLOB の

J，H，D などのフォーマルな書き言葉のサブコーパスが位置し，そこからやや斜め左下に向かって Press や General なサブコーパス，さらに右上に話し言葉を含む Fiction のサブコーパスが位置する。つまり，左上から右下にかけて，フォーマルからインフォーマルへと並んでいる。右上の端の方には，話し言葉のコーパスが位置しており，書き言葉とは異なる軸にある情報を含むとも取れるが，左から右へ第 1 次元で見れば，フォーマルからインフォーマルへと並んでいる。ただし，このようなプロットによく見られる馬蹄形 (horseshoe) に並んでいる (Greenacre, 2007) とも解釈できるため注意が必要である。

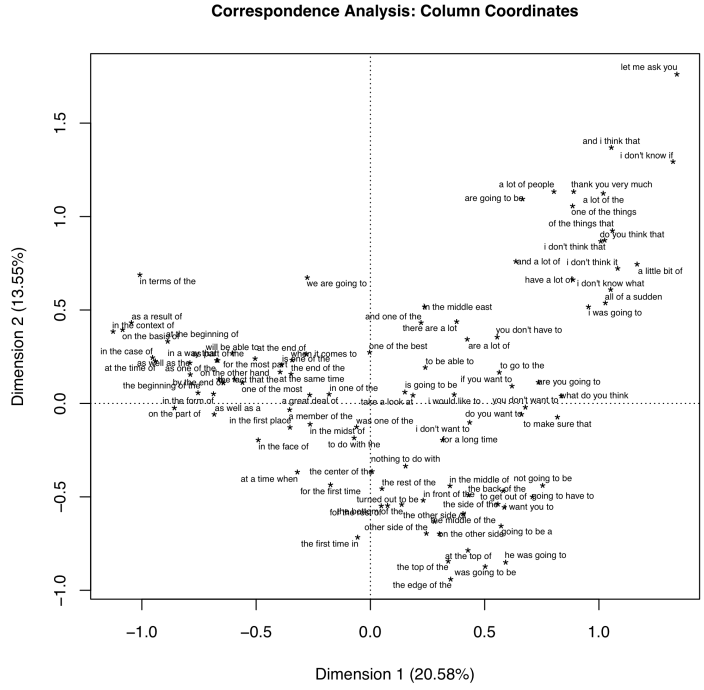


図 21 コレスポンデンス分析（列項目）

列項目（図 21）を見ると，左中央あたりには，前置詞句の名詞に of や that が続く形の 4-gram が多く見られ，中央右下の Fiction に対応する位置に行くに従って，be going to を含むものや動詞で始まる 4-gram が見られるようになる。右上の話し言葉コーパスに対応する位置には, 口語特有の I think や don't, a lot of や a little bit などの程度を表す表現などが見られる。このように，コーパスと単語・n-gram の対応が見られるのがこの分析手法の特徴である。ここから，さらに KWIC などで文脈などを精査して分析を進めていくことになる。

4.3.3 主成分分析

主成分分析は，多量のデータを圧縮し，相関のない主成分に情報を集約していく手法である (石川 et al., 2010)。コーパス分析では，主に第 1 主成分と第 2 主成分の負荷量をプロットして，項目の位置付けにより関係性を視覚化する。CasualConc では，コレスポンデンス分析と同様に，ファイル情報で頻度表を作成し，そのデータを利用して分析したものを描画する。

ここでは，ランダムフォレストで特徴語抽出を行い，その特徴語の頻度表で主成分分析を行うという応用的な例を示す。ここでのコーパスは，ランダムフォレストの例で使用した *Applied Linguistics* と *TESOL Quarterly* の論文コーパスを使用する。まずは，ファイル情報の単語頻度で，ストップワードを削除したファイルごとの単語頻度表を 100 万語あたりの相対頻度で作成し，これを使って，ランダムフォレストで 2 つのコーパスの分類に寄与する単語を抽出する。Gini 係数減少度の上位から，それ自体の頻度が意味を特に持たない単語（リストの項目記号や複数著者の引用の et al.など）を取り除いた 100 語を特徴語として取り出し，ファイル情報の頻度表にこの 100 語でフィルターをかけて特徴語の頻度表を作成する。特徴語を使って，ファイルごとのキーワードグループの頻度表を作成しても同じ結果になるはずである。

この特徴語の頻度表を主成分分析にかけた結果が図 22 である。第 1 主成分と第 2 主成分の因子負荷量をプロットしたもので，円の大きさはその単語の頻度を表している。このプロットの左側が *TESOL Quarterly*，右側が *Applied Linguistics* に特徴的な単語となっており，中央から離れるほど，それぞれにより特徴的な語であると判断できる。*TESOL Quarterly* に特徴的なものには，students，curriculum，teaching，teachers，education などの教育に関わる単語が多い。*Applied Linguistics* に特に特徴的な単語というのは *TESOL Quarterly* ほどはっきりしていないが, verb, utterance，structure，acquisition などが並んでいる。このあたりの単語は，ランダムフォレストの結果の Gini 係数の減少度大きさや精度の減少の大きさと必ずしも一致しておらず, 特徴語の分析には多角的な評価が必要なことを示唆している。

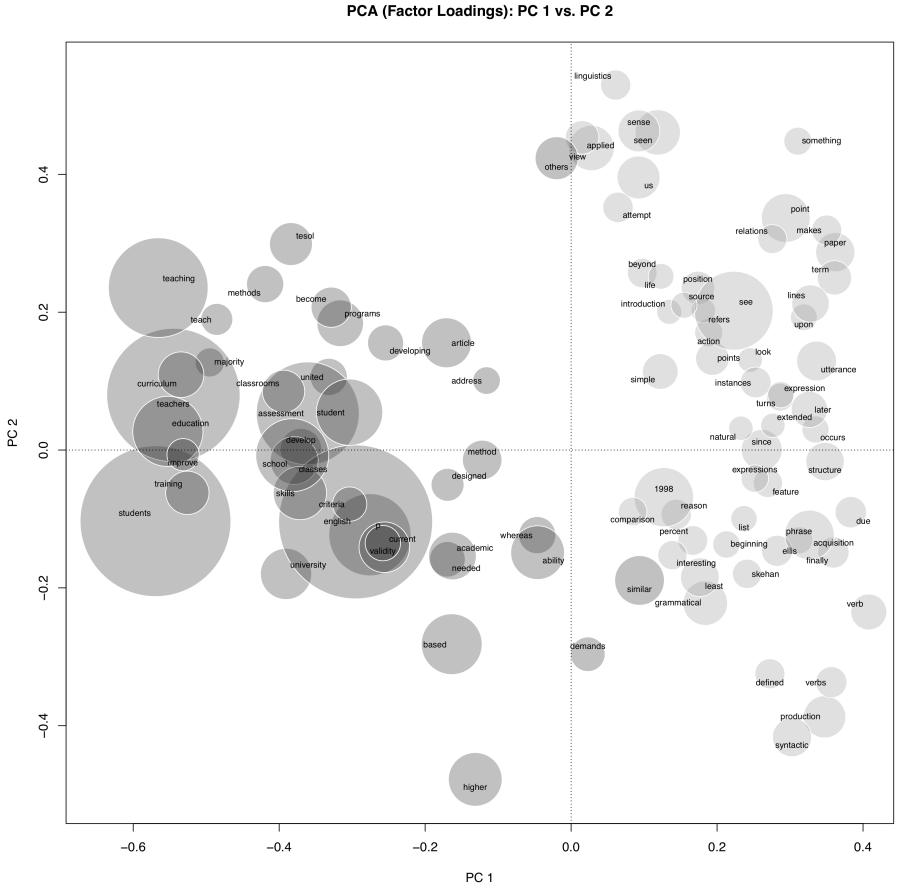


図 22 主成分分析（因子負荷量）

これら以外の多変量解析なども含めて，とりあえずついているから使ってみるというのも一つの考えではあるが，何か興味深い現象などが見つかった際には，それぞれの分析手法について理解を深めた上で研究に使うことが望ましい。

おわりに

CasualConc 2.0 は，これまでのテキスト検索や頻度集計が中心のコンコーダンサーの枠にとどまらない，テキスト分析の機能を備えたアプリケーションとなっている。しかしながら，コーパス分析において大切なのは，あくまでもテキストであり，アプリケーションはその分析の補助をする道具に過ぎない。

CasualConc によって，これまでは高価な統計分析アプリケーションや，無料であっても利用に高度なコンピュータスキルが必要なアプリケーションを使う必要

があったために手を出せなかった分析手法に手が届くようになるかもしれない。手軽に使えることによって，これまでは思いもしなかったような発見がある可能性が高くなる一方，手法の理解をしないまま間違った使用や解釈が行われる可能性も高くなる。

これらのことを踏まえて，新たな手法に手探りでも取り組み，興味深い現象が見えたところで手法についての理解を深め，さらには，そこでテキストに立ち返り，その現象についての分析を進めていただきたい。CasualConc がそのプロセスの一部になり，言語現象の解明に役立つことがあれば幸いである。

引用文献

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

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide; spoken and written English grammar and usage*. Cambridge, UK: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Greenacre, M. (2007). *Correspondence analysis in practice* (2nd). Boca Raton, FL: Chapman & Hall/CRC.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London, UK: Continuum.

今尾康裕 (2012) 「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—」『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』121–178.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In G. Weir, S. Ishikawa, & K. Poonpon (Eds.), *Corpora and language technologies in teaching, learning and research* (pp. 3–11). Glasgow, UK: University of Strathclyde Publishing.

石川慎一郎・前田 忠彦・山崎 誠 (2010) 『言語研究のための統計入門』くろしお出版

金明哲 (2009) 『テキストデータの統計科学入門』岩波書店

小山由紀江・水本 篤 (2010) 「単語連鎖にみる科学技術分野と他の分野の英語表現比較 」『統計数理研究所協同研究リポート』239, 1–11.

Mair, C. (1999). *The Freiburg-Brown corpus ('FROWN')*. Freiburg: Department of English. Albert-Ludwigs-Universität Freiburg.

Mizumoto, A. (2016-). *AWSuM*. Available online at: http://langtest.jp/awsum

Schmid, H. (1994). Probabilistic part-of speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing* (p. 154). Manchester, UK.

竹内理・水本篤 (2014) 『外国語教育研究ハンドブック（改訂版）』松柏社