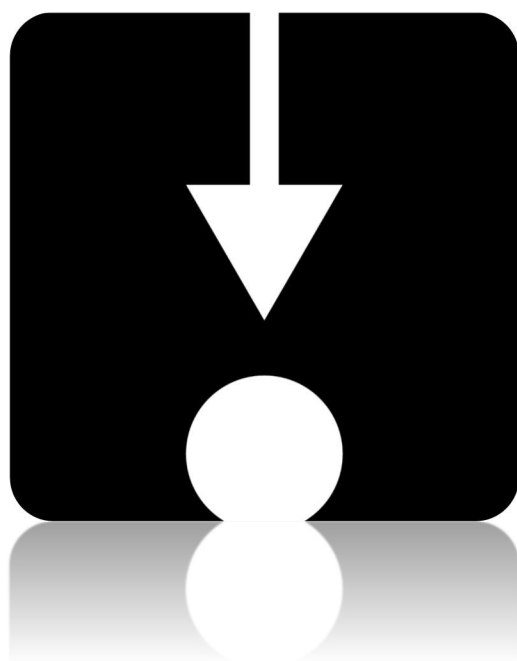


#LancsBox 5.1 マニュアル



#LancsBox を引用:

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software package]

Brezina, V., Timperley, M., & McEnery, A. (2018). #LancsBox v. 4.x. [software package]

Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139-173

.innovation in corpus linguistics

#LancsBox

@Lancaster University

目次

1	#LancsBox バージョン 5.1 をダウンロード・起動する	5	5	Whelk ツール	23
2	データの読み込み・インポート	10	5.1	Whelk タブの要覧	23
2.1	Corpora タブの概要	10	5.2	上部パネル: KWIC	23
2.2	コーパス・ワードリストを読み込む	11	5.3	下部パネル: 頻度配分	24
2.3	サポートされたファイル形式	11	6	GraphColl	25
2.4	#LancsBox のコーパス・ワードリストをダウンロードする	12	6.1	GraphColl タブの要覧	25
2.5	コーパス・ワードリストを作動する	13	6.2	コロケーショングラフの作成	26
2.6	コーパスを保存する	13	7	Words ツール	27
2.7	コーパスの前処理 (上級者向け)	14	7.1	Words ツールの要覧	27
3	キーファンクション	17	8	Ngram ツール	29
3.1	マウス・クリック	17	8.1	Ngram ツールの要覧	29
3.2	ショートカットキー	18	9	Text ツール	31
3.3	ツールとタブ	18	9.1	要覧	31
3.4	分割スクリーン	19	10	Wizard ツール	32
3.5	分析結果の保存	20	10.1	要覧	32
3.6	指定した結果のコピー/ペースト	20	10.2	研究レポート	33
4	KWIC ツール (文脈の中におけるキーワード: key word in context)	21	11	#LancsBox における検索機能	34
4.1	KWIC タブの要覧	21	12	#LancsBox と統計	41
			12.1	頻度の算出	41
			12.2	拡散の算出	41
			12.3	キーワード算出	41
			12.4	コロケーションの算出	42
			13	用語集	43
				Notes	46

#LancsBox v.5.1: ライセンス

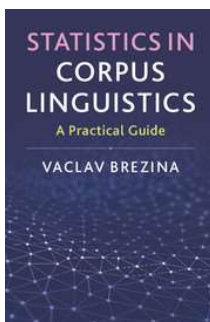
#LancsBox は BY-NC-ND クリエイティブ・コモンズ・ライセンスのもと、使用権が認可されています。#LancsBox の非営利的使用は無料です。ライセンスはこちらから閲覧可能です: <http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

#LancsBox は以下の第三者ツールやライブラリを利用しています: Apache Tika, Gluegen, Groovy, JOGL, minlog, QuestDB, RSyntaxTextArea, smallseg, TreeTagger. クレジットはこちらから参照をお願いします <http://corpora.lancs.ac.uk/lancsbox/credits.php>

#LancsBox を使用して研究を発表する際は以下のように引用をお願いします:

- ☐ Brezina, V., McEnery, T. & Wattam, S. (2015). [Collocations in context: A new perspective on collocation networks](#). *International Journal of Corpus Linguistics*, 20(2), 139-173.
- ☐ Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software package]
- ☐ Brezina, V., Timperley, M., & McEnery, A. (2018). #LancsBox v. 4.x. [software package].

統計処理についての参考



Brezina, V. (2018). *Statistics for corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.

コーパス言語学における統計処理についての詳細に興味のある方は、Brezina (2018)を参照してください。また、ランカスターのオンライン統計ツールもこちらから利用可能です; <http://corpora.lancs.ac.uk/stats>

役立つ文献・資料集

- Brezina, V. (2016). Collocation Networks. In Baker, P. & Egbert, J. (eds.) *Triangulating Methodological Approaches in Corpus Linguistic Research*. Routledge: London.
- Brezina, V. (2018). Statistical choices in corpus-based discourse analysis. In Taylor, Ch. & Marchi, A. (eds.) *Corpus approaches to discourse: a critical review*. Routledge: London.
- Brezina, V. & Gablasova, D. (2017). The corpus method. In: Culpeper, J, Kerswill, P., Wodak, R., McEnery, T. & Katamba, F. (eds). *English Language (2nd edition)*. Palgrave.
- Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139-173.
- Brezina, V., & Meyerhoff, M. (2014). Significant or random. *A critical review of sociolinguistic generalisations based on large corpora. International Journal of Corpus Linguistics*, 19(1), 1-28.
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus - based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, 67 (S1), 155–179.
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, 67 (S1), 130–154.

- ビデオ講座、演習、スライド等の資料は#LancsBox のウェブサイトで閲覧可能です↓

<http://corpora.lancs.ac.uk/lancsbox/materials.php>

1 #LancsBox バージョン 5.1 をダウンロード・起動する

#LancsBox は新世代のコーパス分析ツールです。バージョン 5 は最適なパフォーマンスを保証するため 64 ビット版 OS（Windows 64 ビット版, Mac, Linux）向けに構築されています。#LancsBox は 32 ビット版においても起動が可能ですが、パフォーマンスに制約がある場合があります。#LancsBox バージョン 5 はインストーラーを搭載しており、#LancsBox のインストールをより簡単に行うことができます。

❶ 選択とダウンロード: 最適な OS 版を選択し、コンピューターにインストーラーをダウンロードします。



❷ インストーラを起動する

コンピュータのセキュリティ警告に同意し、インストーラーの手順に従います。この際、必ず #LancsBox をフォルダーにインストールします。（ツールがユーザーフォルダやデスクトップなどにおいて読み書き権限を持ちます。） Windows ではプログラムファイルに #LancsBox をインストールしないよう、注意が必要です。

重要な注意点: システム権限について

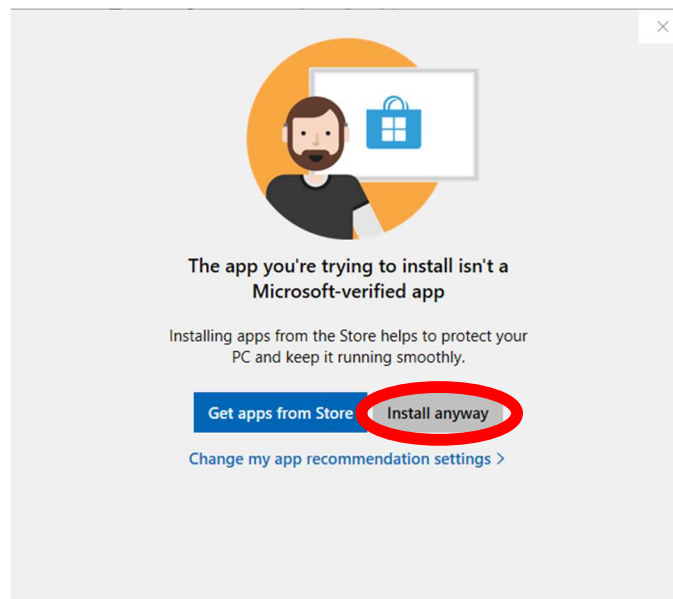
OS に対応する以下の指示に従ってください。

Windows 10

Windows 10 は 2 つのメッセージのうち、いずれかを表示します。

>新型のビルド

‘The app you are trying to install isn’t a Microsoft-verified app’ (インストールしようとしているアプリは、Microsoft Store の確認済みアプリではありません) この警告の表示後、‘Install anyway’ (インストールする) をクリックします。

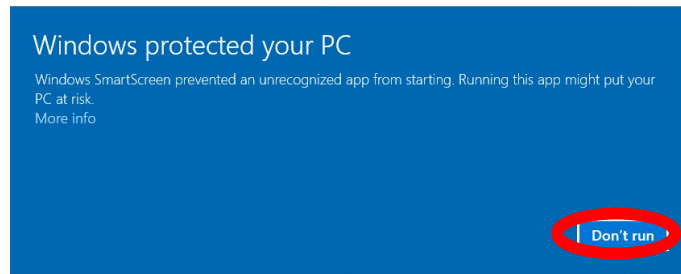


>旧型のビルド

①

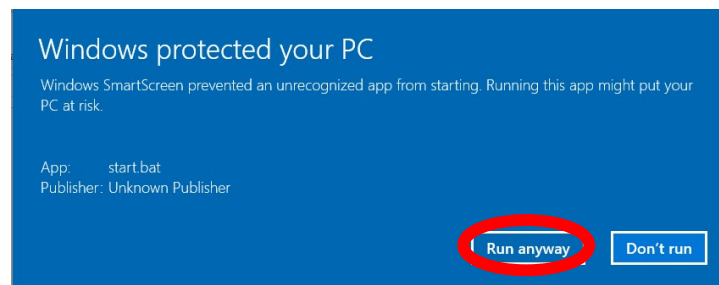
‘Windows protected your PC’ (Windows によって PC が保護されました) :

この警告が表示されたら、‘More info(詳細情報)’をクリックします。



②

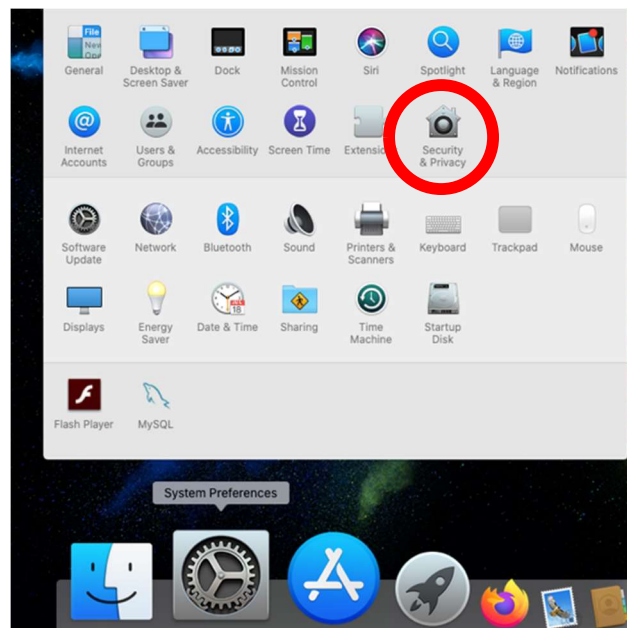
‘Run anyway’ (実行) をクリックします。



OS X

①

ドックの'System Preferences'（システム環境設定）を開き、'Security & Privacy'（セキュリティとプライバシー）をクリックします。

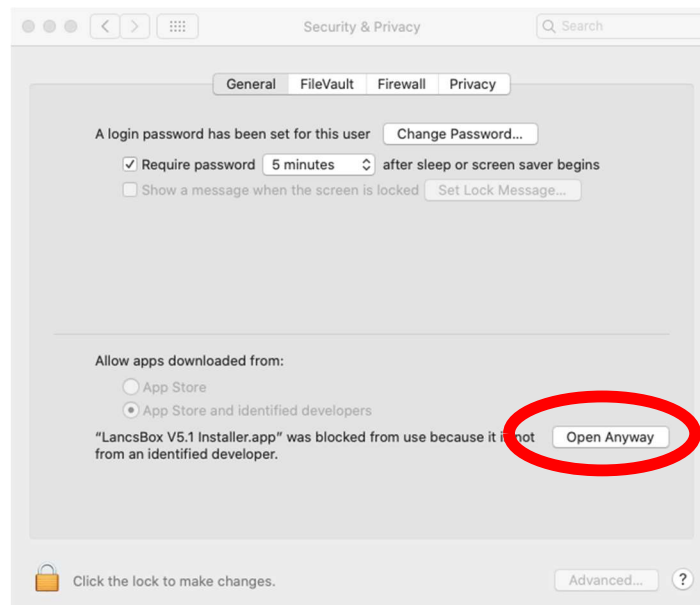


②

‘LancsBox V5.1 Installer was blocked because it is not from an identified developer’

（開発元”LancsBox V5.1 Installer app”のシステムソフトウェアの読み込みがブロックされました）

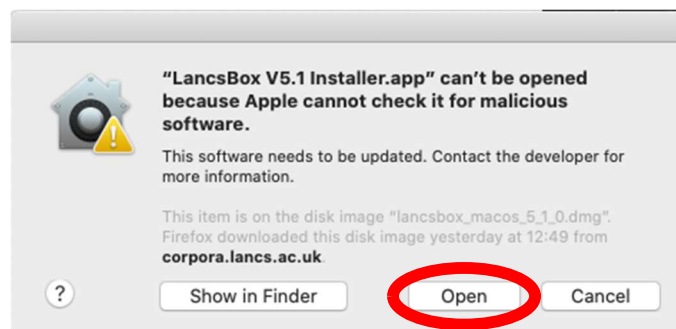
というメッセージの隣の‘Open Anyway’（許可）をクリックします。



③

‘LancsBox V5.1 Installer.app’ can’t be opened because Apple cannot check it for malicious software’（LancsBox V5.1 Installer.app”は開発元が未確認のため開けません）

というメッセージが新しいウィンドウに表示されたら、“open（開く）をクリックします。



2 データの読み込み・インポート

データは#LancsBoxの‘Corpora’タブで読み込み、インポートが可能です。このタブは#LancsBoxを起動する際に、自動で開かれるものです。#LancsBoxは異なる形式でコーパス（.txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx, .zip等）とワードリスト（.csv）を作動します。

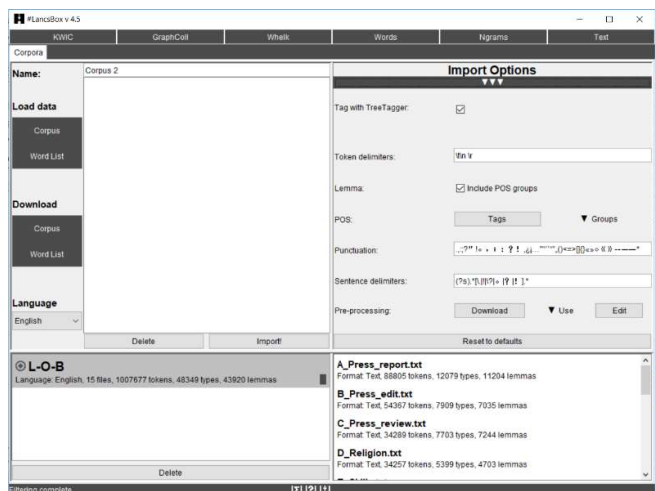
コーパスとワードリストの読み込みには2つのオプション（①自身のデータを読み込む ②#LancsBoxに配信されているコーパスやワードリストをダウンロード）があります。

2.1 Corpora タブの概要

画面上部: コーパス、ワードリストをインポート

ここでは、

- 読み込むコーパスやワードリストを選択できます。
- #LancsBoxに配信されたコーパスやワードリストをダウンロードできます。
- 言語を選択できます。
- POSタグを見直すことができます。
- 句読点や区切り文字を見直すことができます。
- カスタマイズ可能なスクリプトを通じて前処理をセットアップできます。
- 基本的な分類（トークン、レマ、POSグループ、句読点）を定義できます。



画面下部: コーポラやワードリストを作動する

ここでは、

- インポートしたコーパスやワードリストを起動・削除できます。
- コーパスとテキストサイズ（トークン、タイプ、レマ）を見直すことができます。
- テキストをプレビューできます。
- POSタグ等で処理されたコーパスを保存できます。

2.2 コーパス・ワードリストを読み込む

#LancsBox では簡単に自身のコーパスやワードリストを作動することができます。これらのコーパスは自分のコンピュータに保存されたものに限らず、あるいはコンピュータからアクセス可能な場所（メモリースティック, 共有ドライブ, ドロップボックス, クラウド等）についても同様です。

1. Corpora タブで、コーパス・ワードリストのどちらを読み込むかに応じて、‘Load data’の下にある ‘Corpus’、もしくは ‘Word List’ を左クリックします。
2. コーパスやワードリストが保存されている場所（フォルダー）を検索できるウィンドウが開かれます。
3. 特定のファイルを選択できます。Ctrl キー+左クリックで複数のファイルを選択できます。Ctrl キー+ A で全てのファイルを選択できます。
4. ‘Open(開く)’を左クリックしてファイルを読み込みます。
5. コーパス・ワードリストの言語を選択します。#LancsBox は複数言語での自動レマ化、POS タグ付けをサポートしています。これは Tree Tagger を用いた機能です。言語がリストにない場合 ‘Other’を選択します。この場合、自動レマ化や POS タグ付けは作動しません。
6. [任意: 3 つの三角形 (▲▲▲) のバーを左クリックすることでインポート・オプションを見直し・変更できます。多くの場合、デフォルトでの使用が可能です。
7. ‘Import!’を左クリックにて、コーパスを#LancsBox にインポートします。#LancsBox は通常、自動的に POS タグをコーパスに追加します。

2.3 サポートされたファイル形式

#LancsBox は異なるファイル形式のコーパスファイルをサポートしています。(.txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx, .zip などその他多数) #LancsBox は自動的にコーパスファイルで利用可能なテキストを抽出し処理します。ワードリストについては、#LancsBox はコンマ区切りのファイル形式(.csv)をサポートしています。

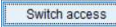
1. コーパスのファイル形式: .txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx, .zip
2. ワードリストのファイル形式: csv (以下参照)

```
Corpus: BNC| Language: English| 4055 files| 96996843 tokens| 662414 types| 716618 lemmas|
"Type", "Frequency: 01 - Freq", "Dispersion: 01_CV"
"the", "6054524.000000", "0.286889"
"of", "3049295.000000", "0.400166"
"and", "2622080.000000", "0.263099"
"to", "2599355.000000", "0.223254"
"a", "2168976.000000", "0.221813"
"in", "1945319.000000", "0.333547"
```

2.4 #LancsBox のコーパス・ワードリストをダウンロードする

#LancsBox では特定のライセンスの元に無料配信された既存のコーパスを作動することが出来ます。コーパス共有について2つのモード（① オープンアクセス ② 制限されたアクセス）において可能です。開発チームはこのリストにコーパスを定期的に追加しています。

1. Corpora タブで、'Download'の下にある'Corpus'または'Word List'を左クリックします。
2. #LancsBox に配信されているコーポラやワードリストを選択できるウィンドウが開きます。
コーパスを左クリックすると、そのコーパスやワードリストについての追加情報(言語、日付、テキストタイプ、ライセンス等)が表示されます。
3. コーパスのライセンスを見直し・同意します。
4. 'Download'を左クリックして選択したコーパスやワードリストを読み込みます。
5. 'Import!'を左クリックしてコーパスを#LancsBox にインポートします。デフォルトで#LancsBox は自動的に POS タグをコーパスに追加します。

▶ **メモ:** オープンアクセスと制限アクセスを切り替えるためには、画面左下の'Switch access'ボタン()を使います。制限アクセスのコーポラは暗号化されて配信されており、いくつかの表示・使用制限があります。例えば、テキストツールでの表示、コンピュータにてローカル保存することはできません。

2.5 コーパス・ワードリストを作動する

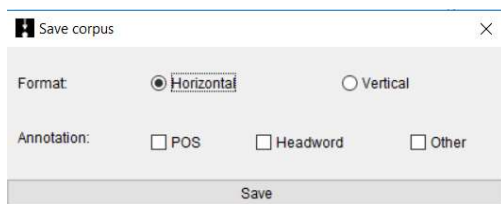
#LancsBox にインポートされたすべてのコーパス・ワードリストは'Corpora'タブの画面下部に表示されます。ここではコーパスのプレビュー、レビュー、また#LancsBox を再度開いた際、コーパスとワードリストの素早い再ロードが可能です。

1. コーパス・ワードリストをインポートすると、画面下部のインポートしたコーパス・ワードリストの隣に表示 (🔍) または (📄) が現れます。これらは'delete'を左クリックすることで消すことができます。画面右下にて、コーパスの構造（コーパスを構成する個々のテキストファイル）を見ることができます。
2. 画面左下では、デフォルトのコーパスが表示されます。デフォルトのコーパスは #LancsBox がデフォルトの選択として、個々のモジュールで提供するコーパスになります。デフォルトのコーパスは名前を左ダブルクリックすることで表示でき、塗りつぶされた長方形 (■) がデフォルトのコーパスの名前の隣に現れます。
3. #LancsBox を閉じると、コーパスとワードリストはインポートされたままになりますが、再度読み込まれます。コーパスとワードリストを起動（再読み込み）するには、コーパス・ワードリストを左ダブルクリックします。
4. コーパス・ワードリストを右クリックすることでファイルをプレビューが可能です。これは Text ツールにて使用されます（セクション 8 を参照）。ファイルの一覧（サイズに関する情報も含め）はスプレッドシートやワード文書にコピー（Ctrl/Command+C）、貼り付け（Ctrl/Command+V）することができます。
5. コーパスは 5 つのモジュール(KWIC、Whelk、GraphColl、Words、Text)を使って分析することができます。ワードリストは Word ツールにて使用されます。

2.6 コーパスを保存する

#LancsBox はコーパスを横長書式もしくは縦長書式で保存します。

1. 保存したいコーパスを右クリックします。
2. 適切なオプションを選択します。



3. 'Save'をクリックします。

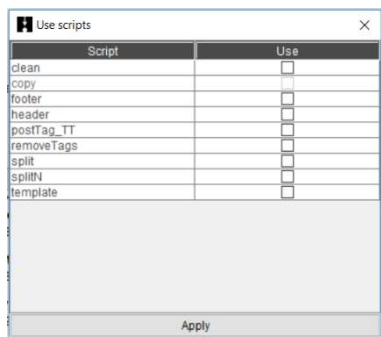
2.7 コーパスの前処理（上級者向け）

#LancsBox はインポートの手順の一部としてデータの前処理ができます。これは'Import options'の'Pre-processing'でセットアップが可能です。データは種類豊富な Groovy スクリプトを使って修正・変更が可能です。(Groovy スクリプト自体についてもカスタム可能です)

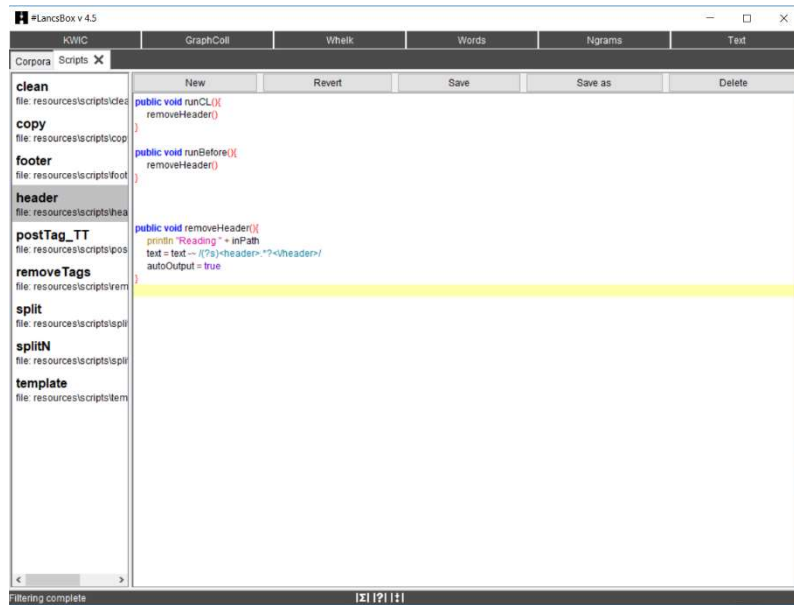
1. 'Pre-processing'では 3 つのオプションが利用可能です。



2. 'Download' では#LancsBox ウェブサイトから利用可能の最新版のスクリプトをダウンロードできます。
3. 'Use' は現在利用可能なスクリプトの一覧を表示します。それぞれのスクリプトの隣に表示されるチェックボックスはどのスクリプトが前処理段階で使われるかを示します。



4. 'Edit'はスクリプトをスクリプトエディターで表示します。ここでは既存のスクリプトを修正、新しいスクリプトを作成することができます。



5. スクリプトの構造は以下の通りです。グルービースクリプト言語に関する詳細は下記のリンクから↓

<http://groovy-lang.org>.

スクリプト	コメント
<pre>public void runCL(){ println "Ran on the command line." }</pre>	<p>スクリプトはコマンドラインを通じて作動します。</p>
<pre>public void runBefore(){ println "Ran as a pre-process script." }</pre>	<p>ファイルが読み込まれているときにスクリプトが作動します。これはファイルの分、テキストの削除・変更、要素（xml タグ等）の構造化を可能にします。</p>
<pre>public void runAfter(Token token){ println "Ran after the tagging step."</pre>	<p>品詞タグ付けの後にスクリプトを作動します。これはタグ付けのエラー修正など Tree</p>

<pre> } public void removeHeader(){ println "Reading " + inPath text = text --~ /(?s)<header>.*?</header>/ autoOutput = true } </pre>	<p>Tagger のアウトプットの修正を可能にします。</p> <p>テキストに”<header></header>”タグで示されたヘッダーを削除する簡単なスクリプトの一例。</p>
---	--

▶ 豆知識

ブラウンコーパスとランカスター-オスロ/ベルゲンコーパス(LOB)はコンピュータに保存・処理された最初の近代コーパスのひとつです。それぞれのコーパスは 100 万語で構成されていて、その規模はその当時では壮大なものでした。ブラウンコーパスは 1960 年代に米国ブラウン大学のヘンリー・クチェラと W.ネルソン・フランシスによって編集されました。元々は IBM のパンチカードで保存・処理されました。1970 年初期、ブラウンコーパスに対応するイギリス版のコーパスが英国ランカスター大学とノルウェーのオスロ大学、ベルゲン大学の協力で編集されました。そのプロジェクトはランカスター大学のジェフリー・リーチによって始められました。

3 キーファンクション

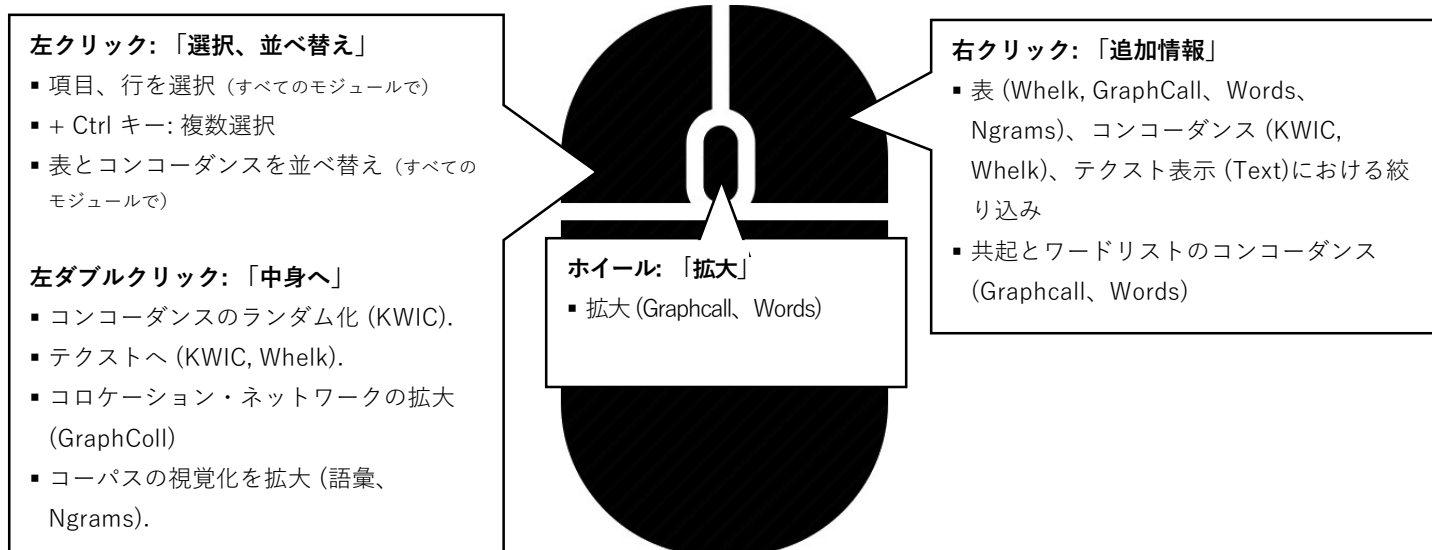
この章では、複数の#LancsBox モジュールにおいて共通するキーファンクションについて解説します。

3.1 マウス・クリック

#LancsBox はドロップダウン・メニューを使いません。代わりに、すべてのコマンドをワンクリックで完了することができます。



カーソルをホバーすることでツール情報（キーファンクション、用語についての簡単な説明）を表示



▶ **注意:** Mac を使用している場合、マウスクリックについての設定を確認する必要があります。初期設定の段階では、右クリックが Ctrl+クリックとして設定されています。また別の方法として 2 ボタン、ホイール付きのマウスを Mac に接続することも可能です。

3.2 ショートカットキー

#LancsBox ではテキストを読みやすいサイズへと変えることができます。この機能はグラフと表についても同様です。

テキストの拡大	Ctrl と +
テキストの縮小	Ctrl と -

3.3 ツールとタブ

#LancsBox では複数のコーパスについて、同時分析が可能です。#LancsBox には 5 つのモジュール (ツール): 「KWIC」、「Whelk」、「GraphColl」、「Words」、「Text」があります。各ツールは別々のタブにおいて繰り返し使用されます。#LancsBox のモジュールは相互接続的であり、ツールはモジュールの中においてポップアップとして起動されます。

1. #LancsBox のトップバーにおいては、ボタンとともに個々のモジュールにおいて複数のタブが開かれていることを確認できます。



2. #LancsBox のモジュールは以下の機能を有します。

KWIC はコンコーダンスを作成します。

Whelk はコーパスのファイルの中における検索語の配分について示します。

GraphColl はコロケーションの特定、視覚化を行います。


Words はワードリストの作成、キーワードの視覚化を行います。

Ngrams は n-gram リストの作成、キーとなる n-gram の視覚と特定を行います。

Text は検索語の文脈を表示します。

3.4 分割スクリーン

#LancsBox は分割スクリーンにおいて、二つの別々の分析について表示することが可能です。パネルの上部にあるいずれか、そして下部のいずれかについて表示が可能です。

3. 分割スクリーンを使用するには、三つの三角形（▲▲▲）を左クリックします。これはパネルの下部を引き出します。
4. 分割スクリーン表示を起動するには、パネルを左クリックします。起動されたパネルは水色のボーダーの囲みによって表示されます。()
5. 分割スクリーンを閉じる際は左のような三つの三角形（▼▼▼）を左クリックします。これは下部のパネルを隠しますが、結果を削除することはありません。必要に応じて下部のパネルを後に引き戻すことが可能です。

3.5 分析結果の保存

#LancsBox では簡単にコンコーダンス、ワードリスト、表、そして視覚化されたものなど、様々な形で結果を保存することができます。

1. #LancsBox で産出された結果は、右上にある保存アイコンを左クリック (📁) することで保存が可能です。
2. 結果を保存する保存先ファイルを選択します
3. 保存 ('Save') をクリックします。

3.6 指定した結果のコピー/ペースト

#LancsBox では指定した結果を簡単にコピー/ペーストすることができます。

1. コピー/ペーストしたい結果を左クリックで選択します。(選択した結果はハイライトで示されます。) 連続しない結果を選択するときは Ctrl を押したまま選択を続けます。すべての結果を選択する際は Ctrl + A のショートカットを使用します。[Mac: Command + A].

Index	File	Left	Node	Right
1	A_Press_rep1	and Juliet" was the irresponsibility of young	love	pushed into tragedy by Shakespeare." Othello" is
2	A_Press_rep1	a cultivated, brave man who comes to	love	too late, and does not know what
3	A_Press_rep1	not to know what to do with	love "	Zeffirelli does not mention the colour of
4	A_Press_rep1	Logue writes fierce, noisy poems about war,	love,	and Logue. Son of a Southampton civil
5	A_Press_rep1	go up in flames one day. In	love,	he wrote:—" I can not see Smiles
6	C_Press_rev1	Byron's leaving him, the scandal of his	love	affair with his half-sister, Augusta Leigh, the
7	C_Press_rev1	him to a point that looks like	love,	had fanned the enthusiasm which had sent

2. Ctrl + C を押します [Mac: Command + C].
3. 新たな場所にて (e.g.テキストファイル、スプレッドシート) Ctrl + V を押します [Mac: Command + V]

4 KWIC ツール (文脈の中におけるキーワード: key word in context)

KWIC ツールはコンコーダンスの形でコーパスの中における検索語の例のリストを作成します。
これは以下のように使用することができます。

- コーパスの中における、語彙、フレーズの頻度算出
- 名詞、動詞、形容詞などの異なる品詞の頻度の算出
- 受け身、分離不定詞など複雑な言語構造を'smart searches'を使って見つける
- コンコーダンスラインの並べ替え、絞り込み、ランダム化
- 二つのコーパスにおける検索語の用例についての統計的な比較分析の実施

4.1 KWIC タブの要覧

The screenshot shows the KWIC tool interface with several annotations:

- 結果を保存** (Save results): Points to the 'Save' button in the top right.
- 統計分析** (Statistical analysis): Points to the 'Statistics' button in the top right.
- 'Index'を左ダブルクリックで、コンコーダンスラインをランダム化** (Randomize concordance lines by double-clicking 'Index'): Points to the 'Index' column header.
- コンコーダンス ヘッダーを左クリックで並べ替え** (Sort concordance headers by left-click): Points to the 'Index' column header.
- コンコーダンス ヘッダーを右クリックで追加の絞り込み** (Filter concordance headers by right-click): Points to the 'Index' column header.
- コンコーダンス表示を左クリックでテキストを閲覧** (Browse text by left-clicking concordance display): Points to the 'Text' column header.
- 右クリックで絞り込みを反映** (Reflect filter by right-click): Points to the 'Text' column header.
- 下部パネルを引き出し** (Expand bottom panel): Points to the bottom panel.

メイン検索ボックス

ここでは以下の通りを行うことができます。

語、フレーズの検索

特定の数値の領域における検索 e.g. >1930&<=1945

「新しい(new)」といった検索

追加検索ボックス

ここでは以下の通りを行うことができます。

異なるアノテーションのレベルにおいての検索

様々なレベルにおける検索語彙の結合

通常の表現の使用, e.g. /N.*/

特定の表現のケースに特化した検索 e.g. /[abc].*/
特定でない、表現のケースについての検索, e.g. /dog|cat/i
句点の検索, e.g. /.*/p
‘Smart searches’の使用, e.g. PASSIVE, NOUN
コーパス検索言語の使用 (CQL)

検索群の定義づけ

▶ 豆知識

1992 年、コーパス言語学の最先端を鑑みたとき、Leech (1992) はコンコーダンスプログラムを、「コーパスに基づく研究において、最も簡潔かつ広く使われているツールである」と考えました (p.114)。25 年後、KWIC のようなコンコーダンスプログラムは今でもコーパス言語学者にとって必要不可欠な道具です。コンコーダンスプログラムが可能にする簡潔かつ直接的なデータへのアクセスは並べ替え、絞り込み、ランダム化などのより洗練された機能とともに、力強い分析手法を提供しています。

Leech, G. (1992). Corpora and theories of linguistic performance. In: *Directions in corpus linguistics*, 105-122.

5 Whelk ツール

Whelk ツールは検索語がコーパスファイルの中でどのように配分されているかについての情報を提供します。

これは以下のように使用することができます。

- コーパスファイルの中における検索語の絶対的、相関的な頻度について算出する
- 異なる基準に基づいた結果による絞り込みを行う
- 検索語の絶対的、相関的な頻度によってファイルを並べ替える。

5.1 Whelk タブの要覧

The screenshot shows the Whelk software interface. The top panel displays search results for the word 'love' across various corpus files. The bottom panel shows a table of corpus distribution, listing files, tokens, frequency, and relative frequency per 10k.

File	Tokens	Frequency	Relative frequency per 10k
P_Romance.txt	58197	75	12.887262
C_Press_review.txt	34289	39	11.37391
K_Fiction_gen.txt	58515	60	10.253781
L_Fiction_myst.txt	48259	15	3.1082284
F_Pop_love.txt	88742	26	2.9298415
N_Adventure.txt	58322	16	2.7433903
G_Belle_lett_nov.txt	155271	35	2.2541234
E_Skills.txt	76613	16	2.0884185
M_Science_fict.txt	12037	2	1.6915435
D_Religion.txt	34257	4	1.1676446
J_Acad_writing.txt	161289	10	0.6200051
A_Press_report.txt	88805	5	0.5630314
R_Humour.txt	18087	1	0.55288327
B_Press_edit.txt	54367	0	0.0
H_Misc_non_fict.txt	60627	0	0.0

上部パネル: コーパスの検索

ここでは以下の通りを行うことができます:

- 検索、並べ替え、絞り込み
- 追加の検索機能、通常の検索機能の使用
- 'Smart' 検索の使用

下部パネル: 配分の表示

ここでは以下の通りを行うことができます:

- 個々のファイルにおける検索語彙の配分表示
- 並べ替え、絞り込み、コピー/ペースト

5.2 上部パネル: KWIC

Whelk の上部パネルは KWIC のツールとして、同じように力強い検索、並べ替え、絞り込み機能を有しています。これは下部パネルへと直接つながっており、上部パネルでのいかなる変更、アップデートも下部パネルへとすぐに反映されます。

5.3 下部パネル: 頻度配分

Whelk の下部パネルは検索語の配分についての詳細情報が供給されています。

1. 「ファイル(File)」の列が個々のコーパスのファイルの名前をリスト化します。
2. 「トークン(Tokens)」の列は各ファイルの中で分析されている語彙数（トークン）の大きさについての情報を提供します。
3. 「頻度 (Frequency)」の列は検索語の絶対的な頻度についての情報を示します。i.e., 各ファイルにおいていくつかの例があるかどうか、について指し示します。
4. 「一万当たりの相関頻度 (Relative frequency per 10k)」は 1 万トークンという基準のもとに正規化された相関的な頻度情報についての情報を提供します。値はファイル、コーパスを通して比較可能なものです。

▶ 豆知識

Whelk ツールは（機能、そして名前の両点で）Kilgarriﬀ's (1997: 138ff) の「つぶ貝問題 (Whelk Problem)」の概念からヒントを得ています。つぶ貝（カタツムリ形の海の生き物）テキスト(本)がコーパスのなかにあったことを想像してみてください。この *Whelks*（つぶ貝）という語は特定の一つの文脈に限って使われているにもかかわらず、テキストの中でが多く現れるため、結果的にコーパス全体において頻度の高い語として表されてしまいます。この問題を克服し、より語の配分について正確な情報を提供するために、Whelk ツールは個々のコーパスファイルにおける検索語の頻度配分についてを示すことができます。

Kilgarriﬀ, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2), 135-155.

6 GraphColl

GraphColl ツールはコロケーションの特定、そしてそれをコロケーショングラフ、コロケーションネットワーク、表といった形での表示を行います。

これは以下のように使用することができます。

- 語やフレーズの共起を探す
- コリゲーション（文法カテゴリーの共起）を特定
- コリゲーションとコロケーションの視覚化
- 語やフレーズの共通する共起の特定
- 談話の「おおよそさ（'Aboutness'）」について概括

6.1 GraphColl タブの要覧

結果を保存

オプションを表示

コロケーション設定の変更

表に共起を表示

コロケーショングラフ、ネットワークの表示

下部パネルの引き出し

Index	Status	Post	Collocate	Count	Score	Log P
1	O	L	eat	10	1072	1072
2	O	R	long-kungo	10	894	894
3	O	R	short	10	1088	1088
4	O	R	spare	10	1059	1059
5	O	L	eat	10	1033	1033
6	O	M	much	10	876	876
7	O	L	life	10	1287	1287
8	O	R	years	10	7033	7033
9	O	R	most	10	1821	1821
10	O	L	where	10	2808	2808
11	O	L	how	10	1287	1287
12	O	R	first	10	1287	1287
13	O	O	on	10	7033	7033
14	O	O	my	10	1821	1821
15	O	R	their	10	2808	2808

6.2 コロケーショングラフの作成

GraphColl はコロケーショングラフを進行的に作成しています。適切な設定を選択したのち、ネットワークの結び目（ノード）、そしてその共起についての検索を行うことができます。

1. コロケーション検索に際して適切な設定を選択:
 - i) Span（スパン）：共起の検索において、ノード（検索語）の左側（L）、右側（R）の何語が考慮に入れられているか [初期設定: 5L, 5R].
 - ii) Statistics（統計）：コロケーションの強さを算出するための関連性の尺度 [初期設定: 頻度 – 尺度は研究の問いにより異なるため、ここでは推奨なし].
 - iii) Threshold（しきい値）：共起として考えられるもの（レマ、語彙、POS）の最低頻度、統計的しきい値。
 - iv) Corpus（コーパス）：検索されるコーパス
 - v) Unit（ユニット）：共起に使用されるユニット（タイプ、レマ、POS タグ）
2. 検索する語を検索ボックス（左上）に入力、「検索 (search)」をクリック
3. これでコロケーションの表、コロケーショングラフが作成されます。（右）

▶ 豆知識

GraphColl という名前は視覚的コロケーションツール (*Graphical Collocation Tool*) の略語です。GraphColl は#LancsBox の最初のモジュールであり、後の段階で追加されたツールと協働します。コロケーションの視覚的表示、コロケーションネットワークは、小さな特殊コーパスにおける語彙ネットワークの概念を提示した Philips (1985) から着想を得たものです。(Philips の語ではコロケーションネットワーク 'collocation networks') GraphColl はこの概念を掘り下げ、小さなコーパスから大きなコーパスまでについて、異なる統計的選択を提供し、進行的にコロケーションネットワークを構築しています。

Phillips, M. (1985). *Aspects of text structure: An investigation of the lexical organisation of text*. Amsterdam:

7 Words ツール

Words ツールではタイプ、レマ、そして POS といったカテゴリーの詳細な頻度分析、そしてキーワード技術を使ったコーパスの比較を行うことができます。

これは以下のように使用することができます。;

- タイプ、レマ、POS タグの頻度、拡散度合いの算出
- コーパスの中の拡散、頻度の視覚化
- キーワード技術を使ったコーパスの比較
- キーワードの視覚化

7.1 Words ツールの要覧

The screenshot displays the #LancsBox v 4.0 Words tool interface. The top menu bar includes options: KWIC, GraphColl, Wheelk, Words, Ngrams, and Text. The main window is divided into two panes. The left pane shows a table for the 'L-O-B' corpus, and the right pane shows a table for the 'Brown' corpus. Both tables have columns for Type, Frequency, Dispersion, and Type. The right pane also features a visual comparison chart with two circles labeled 'L-O-B' and 'Brown'. Annotations with arrows point to specific parts of the interface:

- 「表のヘッダーを右クリックで絞り込みを起動」 (Click the table header to start filtering)
- 「コーパスをドラッグしてキーワードを作成」 (Drag the corpus to create keywords)
- 「コーパスを左ダブルクリックで内部構造を表示」 (Double-click the corpus to show internal structure)
- 「右クリックでコーパスの統計を表示」 (Right-click to show corpus statistics)
- 「テーブル内を右クリックで Wheelk ポップアップを起動」 (Right-click in the table to start the Wheelk pop-up)

The bottom status bar indicates 'Filtering complete' and shows a search bar with the text 'L-O-B'.

左: 頻度情報の作成、キーワードと拡散の算出

右: 頻度情報、拡散、キーワードの視覚化

▶ 豆知識

キーワード分析の統計的技術は元々 Mike Schott (1997)によって開発され、WordSmith ツールで行われました。これはカイ二乗検定、ログ尤度検定を使ったコーパスの比較によるものでした。Kilgarriff が指摘した通り、カイ二乗検定とログ尤度検定はこの種の比較に完全に適したものではありません。Kilgarriff の解決法はスケッチエンジンを使った『簡単な数学 (simple math)』、比較するコーポラ中の語彙の相対的頻度の単純な比較、によってなされるものでした。この「簡単な算数」に加え、#LancsBox はコーパスの比較において他の種類の手段をも提供します。

Scott, M. (1997). PC analysis of key words—and key key words. *System*, 25(2), 233-245.

Kilgarriff, A. (2009, July). Simple maths for keywords. In *Proceedings of the Corpus Linguistics*

8 Ngram ツール

Ngram ツールは隣接するコンビネーションの種類、レマ、POS によって決定される n-gram の頻度についての詳細な分析を可能にします。ツールはキーワードに類似した技術を使い二つのコーパスの比較することで、キーとなる ngram を算出します。

これは以下のように使用することができます：

- N-gram、単語連鎖、p-frame、skip gram の特定
- Ngram の種類、レマ、POS タグ用の頻度、拡散の尺度の算出
- コーパスの ngram の頻度、拡散の視覚化
- キーワード技術を使用した二つのコーパスの ngram の比較
- キーとなる ngram の視覚化

8.1 Ngram ツールの要覧

The screenshot shows the #LancsBox v 4.0 Ngram tool interface. The main window has a menu bar with options: KWIC, GraphColl, Wheelk, Words, Ngrams, and Text. The 'Ngrams' menu is open, showing 'Corpora' and 'Ngrams: L-O-B, Brown'. Below this is a search bar and a table with columns: Corpus, L-O-B, Frequency, Dispersion, Type, and Grams. The table is divided into two sections: 'L-O-B' and 'Brown'. The 'L-O-B' section shows a list of words and their frequencies and dispersions. The 'Brown' section shows a list of words and their frequencies and dispersions. Annotations with callouts point to various parts of the interface:

- 「複数のコーパスをドラッグでキーとなる ngram を作成」 (Create key ngram by dragging multiple corpora)
- 「左ダブルクリックでコーパスの内部構造を表示」 (Display internal structure of corpus by left double-click)
- 「右クリックでコーパスのコーパス統計を表示」 (Display corpus statistics by right-click)
- 「表のヘッダーを右クリックで絞り込みを起動」 (Start filtering by right-clicking the table header)
- 「表の内部を右クリックで Wheelk ポップアップを起動」 (Start Wheelk pop-up by right-clicking the table content)

Corpus	L-O-B	Frequency	Dispersion	Type	Grams
of the	9518.000000	0.381724			
in the	5961.000000	0.224633			
to the	49.000000	0.149529			
on th	00	0.140736			
and t	00	0.270452			
it is	00	0.652103			
for th	00	0.343772			
to be	1912.000000	0.224275			
at the	1745.000000	0.211144			
that the	1651.000000	0.551571			
it was	1555.000000	0.553916			
with the	1525.000000	0.258497			
from the	1509.000000	0.159117			
of a	1501.000000	0.254168			
by the	1486.000000	0.503977			
in a	1259.000000	0.247329			

Corpus	Brown	Frequency	Dispersion	Type	Grams
your expenses	1.000000	3.741657			
owe additional	1.000000	3.741657			
foundation during	1.000000	3.741657			
surprise he	3.000000	2.176043			
health hazard	1.000000	3.741657			
parables being	1.000000	3.741657			
with lipstick	1.000000	3.741657			
sullam that	1.000000	3.741657			
drank slowly	1.000000	3.741657			
horsemanship	1.000000	3.741657			
have fashioned	1.000000	3.741657			
for lunch	3.000000	2.055493			
themselves from	3.000000	2.102494			
unlikely synonyms	1.000000	3.741657			
noble or	1.000000	3.741657			

左: 頻度リストの作成、拡散とキーとなる ngram 右: 頻度、拡散、キーとなる ngram の視覚化の算出

▶ 豆知識

複合語表現は言語を記述する上で極めて重要なものです。コロケーション、n-gram、単語連鎖、p-frame など複合語表現を説明する上で様々な用語があります。GraphColl モジュールで特定されるコロケーションは通常、非隣接的表現を代表する一方で、n-gram における複合語表現は隣接する語彙文法パターンを代表します。これは以下のように定義されます。

- N-gram: テキスト、またはコーパス内の連続する POS、レマ、n-types
- 単語連鎖: 配分的性質や歩いて位の頻度を持つ n-gram (e.g., 相対的頻度 1000 万の範囲で 5 以上)
- P-frame (加えて skip gram): 一つ、または複数の箇所において変異点を可能にする n-gram; it would be * to など

これらの複合語表現のすべての種類は#LansBox の Ngram ツールで特定することができます。

Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139-173.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus - based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, 67 (S1), 155-179.

9 Text ツール

Text ツールでは語やフレーズが使われている文脈について深く洞察することが可能です。

これは以下のように使用することができます：

- 検索語を完全な（断片的でない）文脈の中で表示
- テキストのプレビュー
- 起動中のテキストとしてコーパスをプレビュー
- テキスト、コーパスにおける異なるレベルでのアノテーションの確認

9.1 要覧

The screenshot displays the Text tool interface. At the top, the search term is 'new', and the corpus is 'LOB'. The results are sorted by 'Occurrences' (181 total). The main pane shows a list of text excerpts with line numbers. The word 'new' is highlighted in several instances. A callout box points to the word 'new' in the first excerpt, stating '10万語あたりの絶対的、相対的頻度' (Absolute and relative frequency of about 100,000 words). Another callout box points to the up and down arrow icons on the right side of the list, stating '上(↑)下(↓)の矢印 事例の間を移動' (Up and down arrows to move between examples).

Line	Text
3387	Mr. Ormesby-Gore has now resigned... in Office, while another reason for the reshuffle is the... for Technical Co-operation."
3237	Mr. Ormsby-Gore, asked if he was... atmosphere is not conducive to pro... view of the Berlin crisis, said: "I think the general political Soviet Union at the present time."
4156	Mr. P. S. Watson and Mr. J. G. Nutman have been appointed directors of Smith and Nephew.
282	Mr. Pearson is now talking about "his new and dynamic liberalism" and this week will show perhaps how far "Mike" will go.
4434	Mr. Platts-Mills breeds prize pigs— there are about 300 of them,— and they respond admirably to his farming techniques."
4414	Mr. Platts-Mills's career details read like a plot for a schoolboy adventure story.
115	Mr. Powell devoted half his speech to giving details of plans for improving the hospital service, on which indeed the Government is making progress.
108	Mr. Powell finds it easier to take it out of mothers, children and sick people than to take on this vast industry," Mr. Brown commented icily."
80	Mr. Powell, white-faced and outwardly unemotional, replied with a statistical statement— and ended by inciting Labour M P s to angry uproar.
1670	Mr. R. Southern, chairman of the C.W.S. Board's retail trading committee and one of four directors nominated by the wholesale societies to act as a caretaker Board in the early stages, said to-day: "Behind the new organisation will be the vast financial and technical resources of the C.W.S. and the C.W.S. Co-ops will be an attractive and modern co-op in the country."

検索語に関してすべての例がハイライトで表示される

上(↑)下(↓)の矢印
事例の間を移動

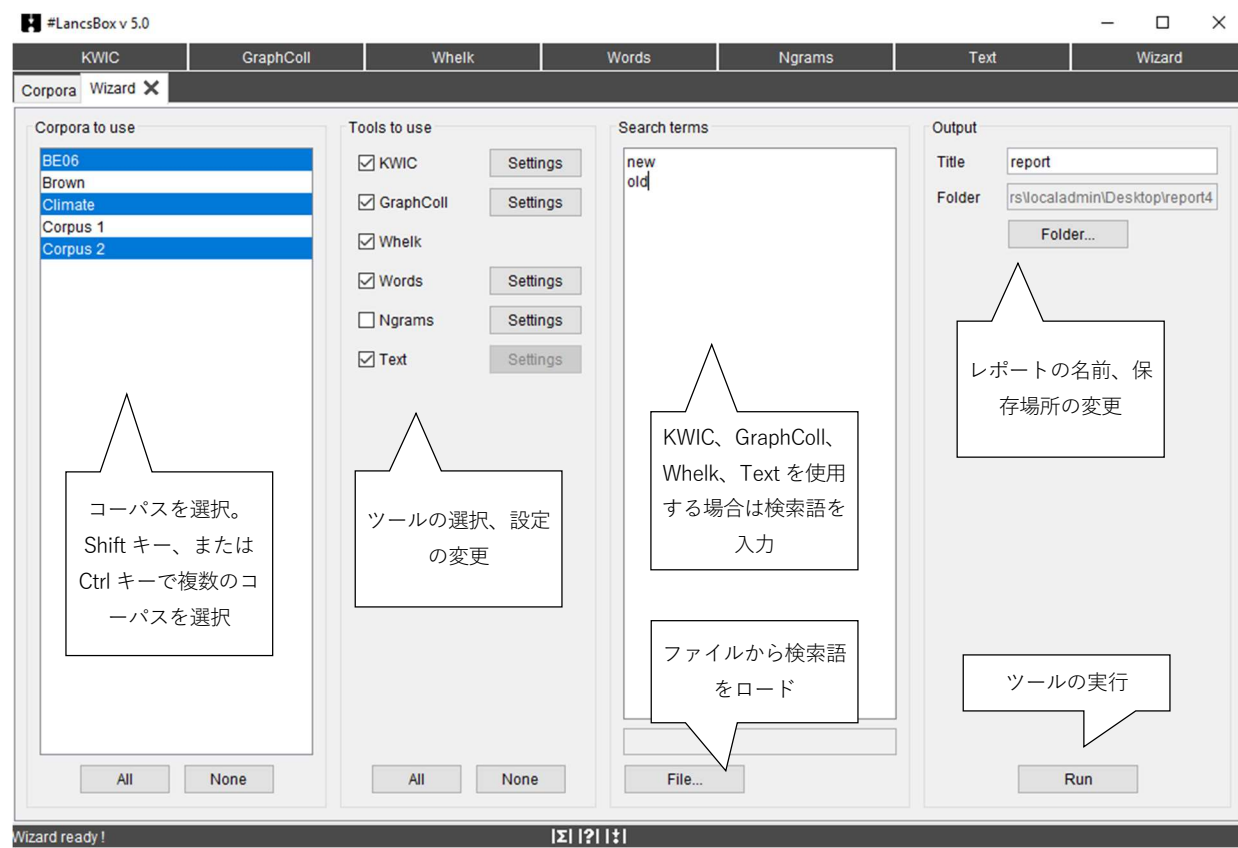
10 Wizard ツール

Wizard ツールは#LancsBox のすべてのツールの力を合わせたもので、web、プリント様式でのレポート作成、コーパスの検索を行います。

これは以下のように使用することができます：

- 複雑、単純な研究の実施
- レポートのドラフト作成
- 関連データのダウンロード

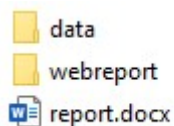
10.1 要覧



10.2 研究レポート

Wizard ツールは二通りの様式 (docx と html) で形式だったデータレポートを作成します。

1. Docx. レポートはワード、Writer などのワードプロセッサで簡単に編集が可能です。



2. レポートの長さ、コンテンツは選択したコーパスの大きさ、ツールなどに依拠します。
3. レポートは学術研究におけるレポートの形式に従って作成されます。

Created by #LancsBox Wizard

June 9, 2020 - 22:07

Comparison of British and American English

1 Introduction

This research report was automatically produced by #LancsBox (Brezina et al. 2020), a corpus analysis tool developed at Lancaster University. It uses cutting-edge technology and statistical sophistication (Brezina 2018) to analyze and visualize corpus data. For more information and tips on research report writing see the [Research Report Guide](#).

2 Method

2.1 Data

The study analyzed the following corpora:

Table 1. Corpora used

Name	Language	Texts	Tokens	Additional information
Brown	English	15	1,014,361	Types: 49,686 Lemmas: 44,622
L-O-B	English	15	1,007,677	Types: 48,349 Lemmas: 43,920

In the study, 2 corpora were used of the total size of 2,022,038 running words (tokens) in 30 texts. A full description of the corpora is available in [data\tsv\corpora](#).

2.2 Procedure

#LancsBox (Brezina et al. 2020) software package was employed to analyse the data. The following tool from the package was used: KWIC. The KWIC tool generates a list of all instances of a search term in a corpus in the form of a concordance. The following search terms were used: "new", "old" and "some".

3 Results

11 #LancsBox における検索機能

ツールを通して、#LancsBox は様々な形で (i) シンプル検索、ii) ワイルドカード検索、iii) スマート検索、iv) 正則表現検索、v) まとめ検索) で異なるレベルのコーパスアノテーションにおける力強い検索を提供します。これに加え、バージョン 5.1 からは CQL (コーパス検索言語- Corpus Query Language)を含む複数の慣習的な用語を使った複雑な検索が可能です。

1. シンプル検索 (Simple Searches)では文字通りに語 (例: *New*)、フレーズ (例: *New York Times*) についての検索を行います。
2. ワイルドカード検索 (Wildcard searches) は以下の 3 種類の特殊な字句の検索を行います。
(*、<、>、=)

特殊な字句	意味	使用例
*	0 もしくはそれ以上の複数文字 、語 [スペースを含む]	new* [<i>new, news, newly, newspaper...</i>] new * [<i>new car, New York, new ideas...</i>]
>	大なり	
<	小なり	
=	等号 [あるいは>と合わせて使われる場合あり]	

3. スマート検索 (Smart searches) はユーザーにとって複雑な検索への簡単なアクセスを提供するため、すでにツールの中で定義された検索です。スマート検索は#LancsBox 特有であり、品詞 (NOUN, VERB など)、複雑な文法パターン (PASSIVE, SPLIT INFINITIVE など)、そして意味カテゴリー (PLACE ADVERB) などの検索に使用されます。

スマート検索は特にツールの中において特定の言語について決定されています。現在、この検索の「慣用」はすでにリソースフォルダーに定義されています。

: resources¥languages¥[name of language]¥Searches.txt. ユーザーは検索/削除でこのリストを編集することが可能です。

英語では以下のようなスマート検索が可能です:

!	BOOSTER
,	COLLECTIVE NOUN
.	COMPARATIVE
?	COMPLEX NOUN PHRASE
ADJECTIVE	CONDITIONAL
ADVERB	CONNECTOR
BE	CONTRACTION

DEGREE ADVERB
 DETERMINER
 DO
 DOWNTONER
 EXISTENTIAL THERE
 GERUND
 HAVE
 INFINITIVE
 HYPHENATED WORD
 INDEFINITE PRONOUN
 INFINITIVE
 INTERJECTION
 LINKING ADVERB
 LONG WORD
 MODAL
 NEGATION
 NOMINALIZATION
 NOUN
 NUMBER

PARTICLE
 PASSIVE
 PAST TENSE
 PAST PARTICIPLE
 PERFECT INFINITIVE
 PHRASAL VERB
 PLACE ADVERB
 PREPOSITIONAL PHRASE
 PRESENT TENSE
 PRONOUN
 PROPER NOUN
 REFLEXIVE PRONOUN
 REPETITION
 SHORT WORD
 SPLIT INFINITIVE
 SUPERLATIVE
 SWEARWORDS
 TIME ADVERB
 VERB

4. 正規表現検索 (Regex searches) は字句のどのような組み合わせの検索についても可能にする高度な検索です。斜線で囲まれた表現はすべて正規表現として解釈されます。`#LancsBox` は通常の表現の周辺についての検索についてもサポートしています。

Regex	Explanation	Regex	Explanation
Word	字句の連続した列（大文字、小文字について判別あり）	a{3}	丁度 3 の a
/word/i	字句の連続した列（大文字、小文字について判別あり）	a{3,}	3 より上の a
/word¥./p	句点検索: 終点の前にある字句の連続（大文字、小文字について判別あり）	a{3,6}	3 から 6 の間の a
[abc]	a、b、c いずれかの字	¥d	数字のいずれか
[^abc]	a、b、c 以外いずれかの字句	¥D	数字でないもののいずれか

[a-z]	a から z までの領域のいずれかの字句	¥w	字句（文字、数字、下線を含む）のいずれか
[a-zA-Z]	a から z、または A から Z までの領域のいずれかの字句	¥W	すべての文字でない字句
[0-9]	0 から 9 の域のいずれかの数字		
.	単字句のいずれか		
(a b)	a もしくは b		
a?	0、もしくは 1 の a		
a*	0、もしくはそれ以上の a		
a+	1、もしくはそれ以上の a		

5. まとめ検索 (Batch searches) は複数の検索語彙について、繰り返しの検索、自動での結果の保存を行います: #LancsBox は複雑、そして単純なまとめ検索をサポートしています。バッチ検索はコーパスにタグ付けがされている際に KWIC、GraphColl、Whelk モジュールでの使用が可能です。まとめ検索は以下のように使用されます、

- a) 詳細検索のオプションを起動するには、検索ボックス中の下向き矢印をクリックします。最後のオプションがまとめ検索になります。「まとめ('Batch')」をクリックします。


The screenshot shows a search interface with a dropdown menu. The menu options are: Headword, POS, <Select file>, [Clear], and Batch. The 'Batch' option is highlighted, indicating it is the selected option for batch searching.

- b) 適切な検索語とともに、一行ごとにテキストファイルのロード、読み込みを行います。シンプル検索語は検索された語の形態のリストを含みます。複雑な検索語は POS タグ、語形、見出し語といった条件の組み合わせを通して定義されます。連続した条件は以下のようにタブ (¥t) によって分離し、同行にて提示される必要があります: ラベル - 語形 - 見出し語 - POS - ユーザータグ (の順。) これは Excel や Calc による詳細まとめ検索語彙とともに作られたファイルによって最も大きな効果を発揮します。下記はこれについてのシンプル検索、複雑検索の例になります。

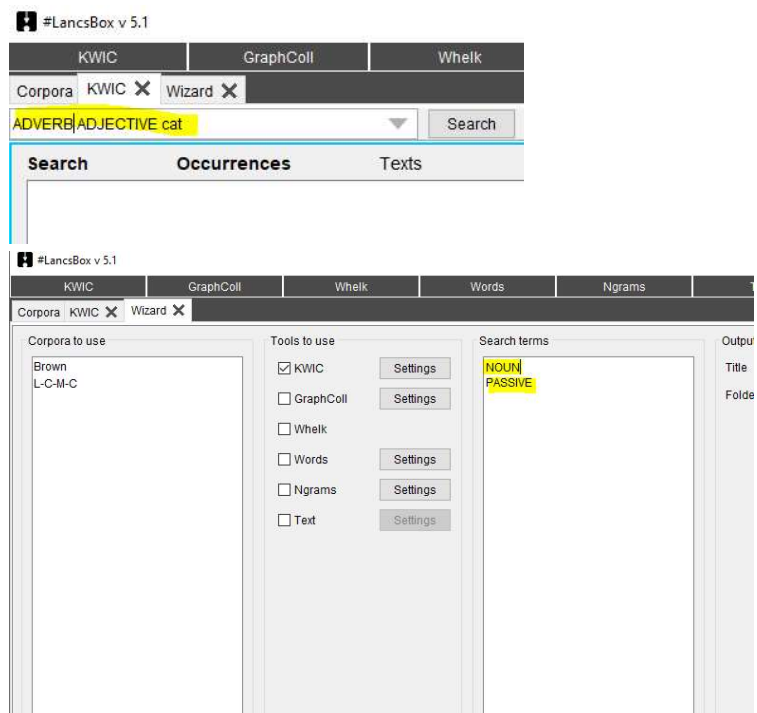
シンプルまとめ検索: 別々の行に各検索語	複雑まとめ検索: ラベル - 語形 - 見出し語 - POS - ユーザータグ (タブによって区分けあり)
----------------------	---

my					
cat					
go					
went					

	A	B	C	D	E
1	multiple-simple	/cat dog mouse/			
2	headword		go		
3	pos			N*	
4	user				Mytag
5	combination	/going went/	go	/V.*/	Mytag

c) 検索語彙のファイルがロードされたら、検索ボタンをクリックし()、結果が保存される場所を指示。

6. 複雑検索 (Complex searches) 。#LancsBox バージョン 5.1 から検索ファンクションが使えるようになりました。#LancsBox は自動で異なる検索方法を特定し、自動でそれに合った検索を行います。



#LancsBox は CQL (Corpus Query Language) のエラー修正も行います – 以下がこの一例についての解説です。

例えば、間違った検索語の入力

[headvor="cat

は [headword="cat"] のように正しい形で読み取られます。

複雑検索では以下のような例が使用可能です。

a) 複数のスマート検索が同検索で使用することができます: スマート検索はシンプル検索と組み合わせることも可能です。

- 1) ADVERB ADJECTIVE
- 2) PRONOUN PASSIVE
- 3) ADVERB ADJECTIVE NOUN was

b) シンプル検索では他の候補を指し示す際に'OR' を使うことができます: 3)と 4)にある通り、括弧を使ってどの語彙が直接接するかについて示すことも可能です。

- 1) cat OR dog
- 2) car OR dog OR mouse
- 3) my (cat OR dog)
- 4) (my cat) OR (my dog)

注意: 'OR'の操作句は異なる長さ、語数の複合表現で使うことができません。e.g.
(my cat) OR dog

c) 'NOT' の操作句はシンプル検索において特定の検索語を否定する際に使用されます (Xでないもののいずれか、という意味合い) : 3), 4)と 5)にある通り、括弧を使ってどの語彙が直接接するかについて示すことも可能です。

- 1) NOT my
- 2) NOT my friend
- 3) NOT (my friend)
- 4) NOT (a good) idea
- 5) NOT (a good or bad) idea NOT me

d) #LancsBox は CQL (コーパス検索言語 'Corpus Query Language) についてもサポートしています。これは異なるレベルでのアノテーション (1-4)、そしてその組み合わせについての複雑検索を定義する上で使用されます。CQL 中の内部の二重引用符についての問いはすべて、大文字小文字の判別をする通常の表現であると読み取られます: 大文字小文字の判別については二重の等号(==)が必要になります (e.g., 5)。

CQL は複数のレベル (: i) 語、ii) 見出し語 (レマ)、iii) POS、iv) タグ) のアノテーションについて検索が可能です。i) から iii) については言語の文法によるものになりますが、iv) はユーザーによって任意で決められたタグを示します。例えば、CQL ではある項目が以下のように定義されます。

```
[word="goes" & headword="go" & pos="V.*"]
```

ここでは 'Goes' という語の形態が、見出し語 'Go'、そして POV タグ Vとして読み取られます。ここで注意すべきはアンパサンドが角括弧で囲まれた異なるレベルのアノテーションを区分けするのに使われているという点です。もしアノテーションのレベルが特定されていない場合、レベルの制限は適用されません。

CQL では角括弧 [] がフレーズの中のスロットの区分けを行います。そのため、例えば、次のような CQL 表現は↓

```
[pos="VB.*"] []{0,3} [pos="V.N"]
```


制限のない 0 から 3 語 ($\{0,3\}$)、過去分詞系 ($V.N$) が続く、動詞となるもの ($VB.*$)、
として読み取られます。

- 1) [word="cat"]
- 2) [headword="go"]
- 3) [pos="V.*"]
- 4) [tag="XX"]
- 5) [word== "Cat"]
- 6) [word="go" & headword="go" & pos="N.* "]
- 7) [headword="go" & pos="V.*"] [word="to"]
- 8) [headword="very" & pos="R.*"]{2} [pos="J.*"]

12 #LancsBox と統計

#LancsBox は i) 頻度情報、ii) 拡散、iii) キーワード、そして iv) コロケーションの値を算出するために統計を使います。これらの算出に用いられる数式は「統計 'Stats」タブにて編集、閲覧することができます（Σ ボタンをクリック）。

12.1 頻度の算出

1. 絶対頻度（度数） = o11
2. 相対頻度（度数） = (o11/r1) x 10,000

12.2 拡散の算出

1. CV 値（変動係数） = SD 値（標準偏差） / Mean（平均）
2. SD 値（標準偏差） = $\sqrt{\frac{\sum(x - \text{平均})^2}{n}}$
3. Range（範囲） = 検索語が最低一度は現れるファイルの数
4. Range（範囲） % = $\frac{\text{範囲}}{\text{ファイルの数}} \times 100$
5. D（差分） = $1 - \frac{\text{変動係数}}{\sqrt{\text{ファイルの数} - 1}}$
6. DP（データ処理） = $\frac{(\text{観測された値} - \text{予期された値}) \text{の絶対値の合計}}{2}$

12.3 キーワード算出

1. 基礎パラメーター = $\frac{C + k \text{ における } w \text{ の相関頻度}}{R + k \text{ における } w \text{ の相関頻度}}$
2. log 尤度関数_{short} = $2 \times \left(O_{11} \times \log \frac{O_{11}}{E_{11}} + O_{21} \times \log \frac{O_{21}}{E_{21}} \right)$
3. % DIFF（差分） = $\frac{(C \text{ における相関頻度} - R \text{ における相関頻度}) \times 100}{R \text{ における相関頻度}}$
4. Log 比 = $\log_2 \left(\frac{C \text{ の相関頻度}}{R \text{ の相関頻度}} \right)$
5. Cohen の d = $\frac{\text{平均 } \ln C - \text{平均 } \ln R}{\text{プールされた SD}}$

12.4 コロケーションの算出

ID	統計	計算式	ID	統計	計算式
1	共起の頻度	O_{11}	8	T-score (T 値)	$\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$
2	MU (母平均)	$\frac{O_{11}}{E_{11}}$	9	DICE	$\frac{2 \times O_{11}}{R_1 + C_1}$
3	MI (共通情報 – Mutual Information)	$\log_2 \frac{O_{11}}{E_{11}}$	10	LOG DICE	$1.4 + \log_2 \frac{2 \times O_{11}}{R_1 + C_1}$
4	MI2	$\log_2 \frac{O_{11}^2}{E_{11}}$	11	LOG 比	$\log_2 \frac{O_{11} \times R_2}{O_{21} \times R_1}$
5	MI3	$\log_2 \frac{O_{11}^3}{E_{11}}$	12	MS (最低感度)	$\min\left(\frac{O_{11}}{C_1}, \frac{O_{11}}{R_1}\right)$
6	LL (ログ尤度関数)	$2 \times \left(O_{11} \times \log \frac{O_{11}}{E_{11}} + O_{21} \times \log \frac{O_{21}}{E_{21}} + O_{12} \times \log \frac{O_{12}}{E_{12}} + O_{22} \times \log \frac{O_{22}}{E_{22}} \right)$	13	デルタ P	$\frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}; \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2}$
7	Z-score ₁ (Z 値)	$\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$	14	Cohen の d	$\frac{Mean_{in\ window} - Mean_{outside\ window}}{pooled\ SD}$

13 用語集

絶対 (Raw) 頻度 – 検索語についてのコーパス、またはその部分における頻度。

まとめ (Batch) 検索 – まとめ検索は複数の検索語を繰り返し検索、かつその結果を自動で保存する。#LancsBox はシンプル、複雑 (i.e., 見出し語、POS タグ、語の形態などの基準の組み合わせによって定義されるような) な検索の両方についてサポートする。

コリゲーション (連辞的結合) – テキスト中にて統計的に特定可能な文法カテゴリー (e.g., POS tag) の構造的共起。

共起 – 構造的にノード (検索語、考察する語、フレーズ) と共に、出現する語。

コロケーション – テキストの中において統計的に党く呈される語彙の構造的共起。

コロケーション・グラフ – 共起とノードの関連性について視覚化したもの。GraphColl を参照。

コロケーション・ネットワーク は談話 – 言語間の複雑な関連性 (コロケーション) について視覚化したもの。複数の関連したコロケーショングラフから成る。GraphColl を参照。

コンコーダンス ライン – KWIC においてノード (検索語) とその前後の語彙について表す行。

コンコーダンス はノード (検索語) について、その語を中心に左右の文脈を表示した形で、コーパスの言語使用について示したもの。コンコーダンスは KWIC 表示とも呼ばれることもある。

コーパス – コンピューターにて検索が可能な言語情報の集合。

拡散 – 拡散はデータ (コーパス) における変数の値 (e.g., 検索語の相対的頻度) の広がりのこと。拡散は標準偏差(SD)や変動係数(CV)、変域、Juilland's D、DP などの数式を用いて統計的に求められる。

頻度 – 検索語がコーパスの中に出現する回数。区分けは絶対 (検索結果のヒット数)、相対頻度 (トークン数に対して配分された頻度) という形でなされる。

頻度配分 – 頻度の配分はコーパスの異なる部分における語やフレーズについての情報を提供する。

GraphColl - #LancsBox のモジュールの一つ。共起の特定、コロケーションネットワークの進行的な作成を行う。

インポート(import) - #LancsBox のパッケージ中で、コーパスデータの計算をすべてのモジュールで可能にする。

KWIC - KWIC は「文脈の中のキーワード (Keyword in Context)」の略語である。これはコーパス中においてノード（検索語、考察する語、フレーズ）を中心として、その左右に文脈としていくつかの語彙を表示する、これは語の使用例の表示における通例である。KWIC はコンコーダンス (concordance) と呼ばれることもある。KWIC は#LancsBox のモジュールの一つである。

左の文脈 - 特定の検索語（ノード）を導く語彙。左の文脈の各位置は L1（直前）、L2, L3 というように呼ばれる。

レマ - 語のすべての屈折は語幹に基づくものである。これは#LancsBox においての標準として、見出し語と文法カテゴリーという組み合わせ (e.g., go+VERB) である。例えば 'go; というレマは以下のような語形態を含む: 'go', 'goes', 'went', 'going', and 'gone'。

語彙の結束 - 一定の頻度、配分（拡散）様態を持つ n-gram。E.g., 相対頻度 1000 万、領域>5

ロード - #LancsBox ではコーパスはロードされた際に分析が可能である。コーパスの再ロードにはコーパスの名前を左ダブルクリック。

モジュール - #LancsBox のツールは特色ある分析機能を有する。#LancsBox は異なる 5 つのモジュールを有する: KWIC、Whelk、GraphColl、Words、Text。

N-gram - コーパス、テキスト中における n タイプ、レマ、P O S の連続

ノード - 考察を行いたい語、フレーズ、文法構造。検索語を参照。

品詞 (POS) - 文法カテゴリー、語類。品詞は通常品詞タグ (POS tagging: 下記参照) をつかって自動でプロセスされる。#LancsBox においては幅広い言語における品詞タグを行う TreeTagger が用いられる。

品詞タグ (POS tagging) - テキスト、コーパスにおいて各語に文法カテゴリーについての情報を加えるプロセス。例えば、次のような文は品詞タグ付けされている: Automatically_RB annotates_VBZ data_NNS for_IN part-of-speech_NN.

P フレーム (skip gram と呼ばれる) - to など一つ以上位置の変化を持つ n-gram。

正規表現 (regex) – ユーザーのどのような組み合わせにおいても検索を可能にするメタ言語。

相対 (正規) 頻度 (RF) – 相対頻度はコーパスにて語の絶対頻度の割合をコーパスの総語数で割る形で算出される。数値は通常標準化に適するように掛け算される。

右の文脈 – 特定の検索語 (ノード) に続く語彙。右の文脈における各位置は R1 (直後) R2, R3 というように示される。

分割スクリーン – #LancsBox において比較を行う際にはスクリーンを2つのパネルへと分割することができる; 各パネルは異なる種類の分析を示すことが可能である。2つ目のパネルについては三つの小さな3角形(▲▲▲/▼▼▼)を左クリックすることで開く、閉じることが可能。

タブ – #LancsBox では新たな「ページ」を開くことで複数の分析を同時進行的に行うことができる。また、各モジュールでは制限なくタブを運用することが可能である。

タグ付け – テキスト、コーパス中の語彙について、言語情報を付加していく作業。自動、半自動で行われる。POS タグを参照。

テキスト (Text) – コーパスの基本的な構成素。前述の通り、コーパスは複数のテキストの集合である。Text は#LancsBox においてコーパス中のテキストの表示、検索を行うモジュールの名前でもある。

しきい値 – GraphColl と Words において関連のある共起、キーワードのみを表示するための設定。

トークン (延べ語数- Token) - テキスト、コーパスにおける総語数

TreeTagger - Helmut Schmid によって開発された品詞タグ用ソフトウェア。様々な言語において、品詞のタグ付けを行う。

Type (異なり語数) - テキスト、コーパスにおける異なる語の形態。

Whelk – Whelk は#LancsBox のモジュールの一つであり、複数のコーパスファイルの中でどのように検索語が配分されているかについて情報を提供する。

Words - Words は#LancsBox のモジュールの一つであり、キーワード技術を使ったコーパスの比較、また頻度の種類、レマ、品詞カテゴリーなどについての深い考察を行う。

