



#LancsBox X マニュアル

先端的な XML 機能を搭載、
大型コーパスに向けた設計。

British National Corpus 2014 で試用可能。

#LancsBox X の引用については:

Brezina, V., Platt, W. (2021). #LancsBox X 1.0 [software packag

目次

1	#LancsBox X のダウンロード、運用	3
2	データのインポート	6
2.1	データのインポートについての要覧.....	6
2.2	コーパスのロード	6
3	KWIC ツール (文脈の中のキーワード: 英 - key word in context).....	8
3.1	KWIC についての要覧	8
4	#LancsBox で検索を行う	11
5	用語集.....	15

1. #LancsBox X: ライセンス

#LancsBox は BY-NC-ND Creative commons license の下登録されています。#LancsBox の商業的使用については自由となっています。


ライセンスは下記のリンクから↓

<http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

1 #LancsBox X のダウンロード、運用

#LancsBox は新世代のコーパス分析ツールです。バージョン X は最大のパフォーマンスを引き出すため、64 ビットのシステム (Windows 64 ビット、OS X、Linux) に向けて設計されています。

❶ **選択とダウンロード:** お使いのコンピューターのシステムに最適なダウンロード・インストーラーを選択します。



The screenshot shows the website interface for #LancsBox. At the top, it says "#LancsBox: Lancaster University corpus toolbox". Below that, it says "Download version X: suitable for large corpora such as the British National Corpus 2014". There is a horizontal line. Below the line, it says "#LancsBox X for Windows". Underneath, there is a button with the Windows logo and the word "Download". Below this, there are three circular icons: the first has the Windows logo and the word "Windows", the second has the Apple logo and the word "Mac", and the third has the Linux penguin logo and the word "Linux".

❷ インストーラーの起動

使用するコンピューターのセキュリティ警告で「同意 (Agree)」を選択します (#LancsBox は安全なソフトウェアです)。この後はインストーラーの示すステップを進めます。#LancsBox は必ず、読み書きの権限の与えられているフォルダーへの保存をお願いします; Windows においてはプログラム・ファイルへの保存はしないよう、お願いします。

注意: システム権限

お使いのシステムに合った説明に沿って、以下の手順を進めてください。

Windows 10

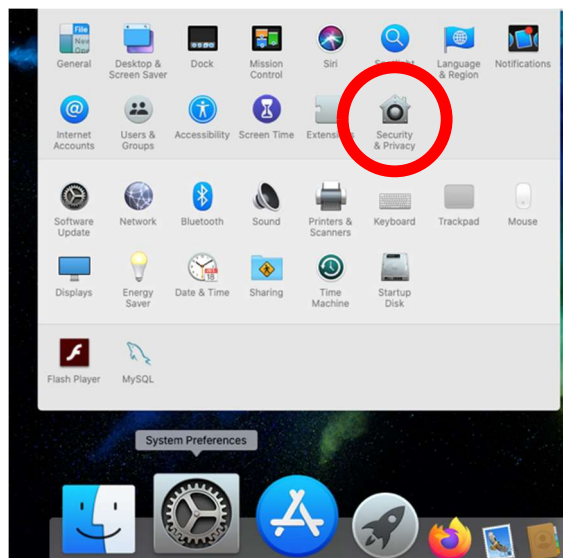
Windows 10 は下記のようなメッセージを表示します。

「インストールしようとしているアプリは、Microsoft 検証済みアプリではありません。」この警告メッセージが表示されたら、「インストールする」を選択します。

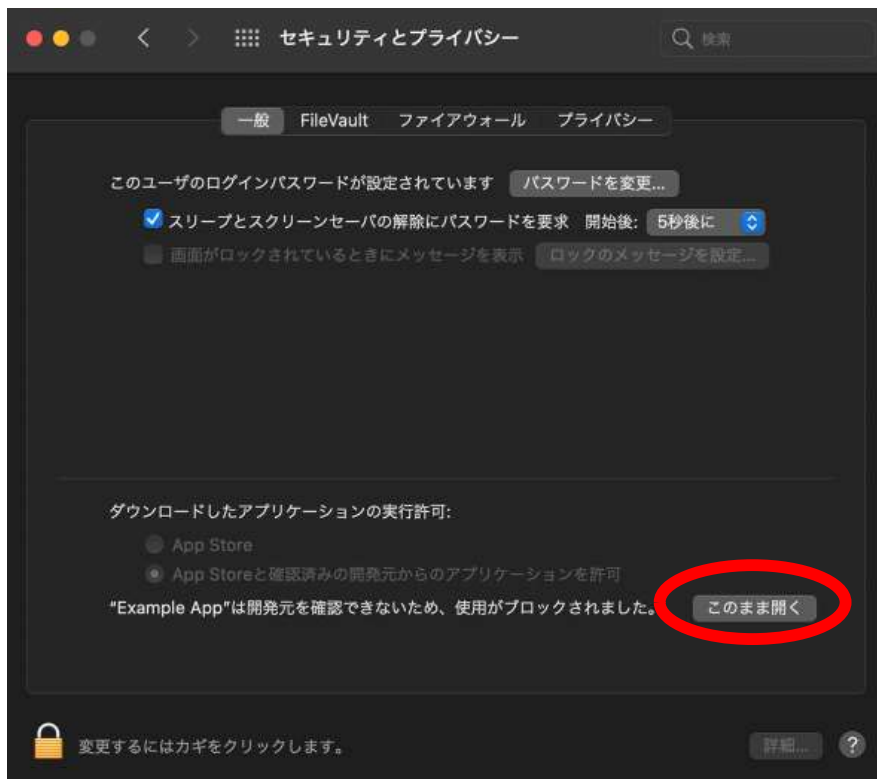


OS X (Mac)

ドックの「システム設定」を開き、「セキュリティとプライバシー」を選択します。



「LancsBox X Installer.app (写真の中では Example App) の開発元を確認できないため、使用がブロックされました」というメッセージの横にある「このまま開く」を選択。



(Retrieved from Apple 2021)

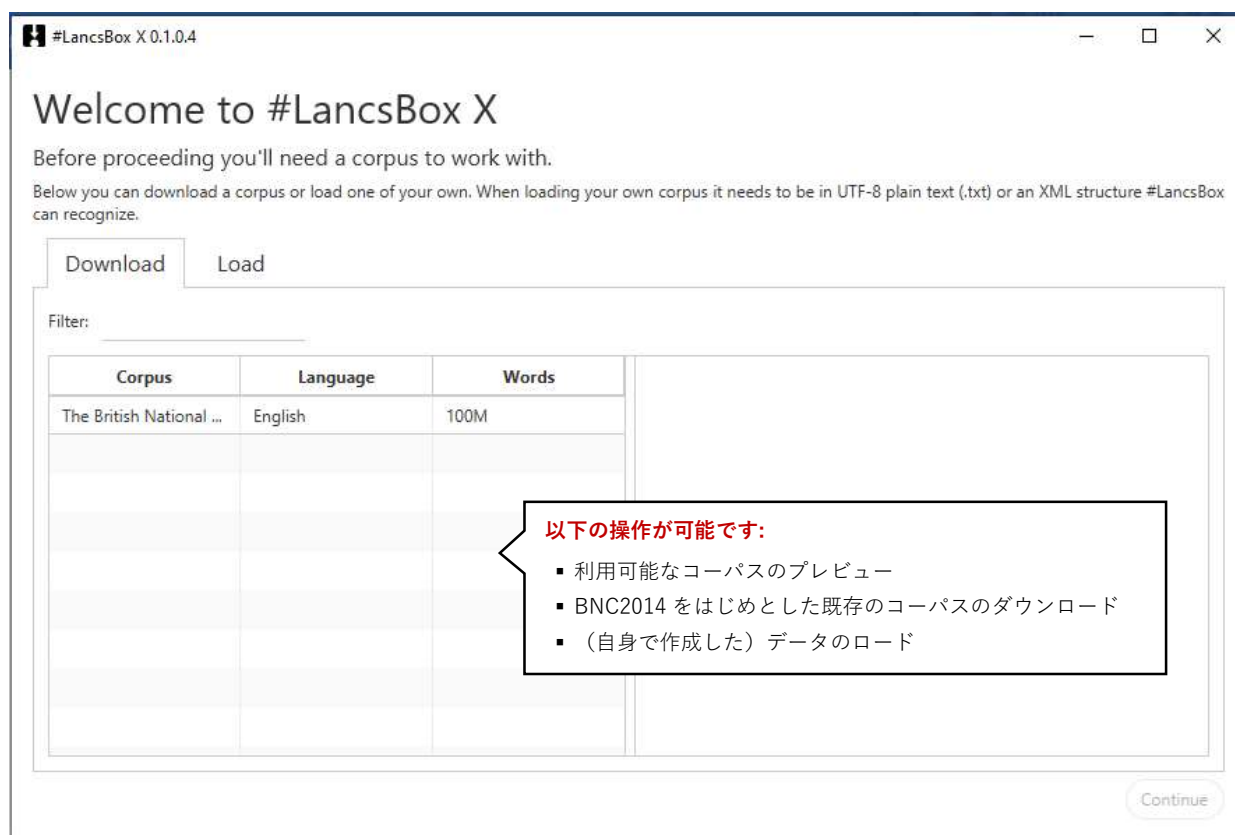
以下のような、「"LancsBox V5 Installer app"が悪質なソフトウェアかどうかを Apple で確認できないため、このソフトウェアは開けません。」というメッセージが表示されたら「開く」をクリック。



2 データのインポート

#LancsBox X は大型コーパスに向けて作られています: メタデータを使った作業を可能にする XML を初期段階でサポートしています。また、データは簡単にインポート、ロードすることができます。

2.1 データのインポートについての要覧



ヒント: キーボードのショートカット (Ctrl -/ Ctrl+) を使ってズームのレベルを変更できます (Mac では Cmd -/ Cmd +)。

2.2 コーパスのロード

#LancsBox では自身で作成したコーパスにおいても作業が可能です。現在、#LancsBox は通常のテキスト・ファイル (UTF-8 でエンコード)、もしくは XML をサポートしています。

1. フォルダーでファイルを準備。
2. データが #LancsBox が認識できる形式であるか確認する。

.txt (UTF-8): テキストファイル	XML: w エレメントを含む
<p>We can pick up on the last comment. Once we are in the grip of reflective thinking it is very hard, if not impossible, for us to see our ethical justifications of our ethical concepts, say, in a genuine way: we will always be drawn to the thought that this is all local. In addition, we will no longer see such judgements as embodying any sort of knowledge.</p>	<pre><?xml version="1.0" encoding="utf-8"?> <text id="AcaHumBk20" mode="writing" genre="academic prose" subgenre="academic prose: humanities" subsubgenre= "academic prose: humanities: NA" publication="book" section ="NA" sample="end" source="NA" author="NA" pubDate="NA" words="6635"> <p n="1"><s n="1"><w pos="PPIS2" hw="we" class="PRON" usas= "Z8">We</w> <w pos="VM" hw="can" class="VERB" usas="A7">can </w> <w pos="VVI" hw="pick" class="VERB" usas="M2">pick</w> <w pos="RP" hw="up" class="ADV" usas="M2">up</w> <w pos= "II" hw="on" class="PREP" usas="Z5">on</w> <w pos="AT" hw= "the" class="ART" usas="Z5">the</w> <w pos="MD" hw="last" class="ADJ" usas="N4">last</w> <w pos="NN1" hw="comment" class="SUBST" usas="Q2:1">comment</w><c>.</c></s> <s n="2" ><w pos="CS" hw="once" class="CONJ" usas="Z5">Once</w> <w pos="PPIS2" hw="we" class="PRON" usas="Z8">we</w> <w pos= "VBR" hw="be" class="VERB" usas="A3">are</w> <w pos="II" hw ="in" class="PREP" usas="Z5">in</w> <w pos="AT" hw="the" class="ART" usas="Z5">the</w> <w pos="NN1" hw="grip" class= "SUBST" usas="A1:1:1">grip</w> <w pos="IO" hw="of" class= "PREP" usas="Z5">of</w> <w pos="JJ" hw="reflective" class= "ADJ" usas="X2:1">reflective</w> <w pos="NN1" hw="thinking" class="SUBST" usas="X2:1">thinking</w> <w pos="PPH1" hw= "it" class="PRON" usas="Z8">it</w> <w pos="VBZ" hw="be" class="VERB" usas="A3">is</w> <w pos="RG" hw="very" class= "ADV" usas="A13:3">very</w> <w pos="JJ" hw="hard" class=</pre>

3. 「ロード (Load)」 タブではコーパスの情報を入力し、「参照 (Browse)」 でファイルのフォルダーを指示。

4. 「ロード (Load)」 をクリック
5. 「続ける (Continue)」 をクリック

KWIC ツールからはコーパスの名前、をクリック、「Add Corpora (コーパスの追加)」を選択することでコーパスを追加することができます。



3 KWIC ツール (文脈の中のキーワード: 英 - 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の検索語について、全例のリストを作成します。これは以下のような用途で使うことができます:

- コーパスの中における語、またはフレーズの頻度について知る
- 形容詞、動詞、名詞などの異なる品詞の頻度について知る
- コンコーダンス列の並べ替え
- 複数の分析について、並べての比較

3.1 KWIC についての要覧

語、フレーズ、構文について検索

結果を保存

#LancsBox X 0.1.0.4

cat

BNC2014 magazines 15M

cat Hits: 428 (0.29)

Left	Node	Right
dual - mode LTE (up to	Cat	4 at 150 Mbps). While
; but they killed that	cat	in his thirties. I soon
ircassia s (CIR) novel	cat	allergy medicine failed to reduce
med bay. Adventure	Cat	tours offer a day or
ffers reward to catch	cat	killer Black Sabbath bassist disgusted
most combative rider, two first	cat	climbs, a special prime on
Convention, Nick Drake and even	Cat	Stevens, also enjoyed a certain
's Binky Felstead speaks to	Cat	Sarsfield about beauty, boys and
Chelsea's Lucy chats to	Cat	Sarsfield about finding her perfect
was just too hard a	cat	for me. It took all
win Eurovision 2014 20. A	cat	saved a little boy from
their garden bushes into a	cat	and has since created a
a traditional curse - a mutilated	cat	on the doorstep. Anger spent

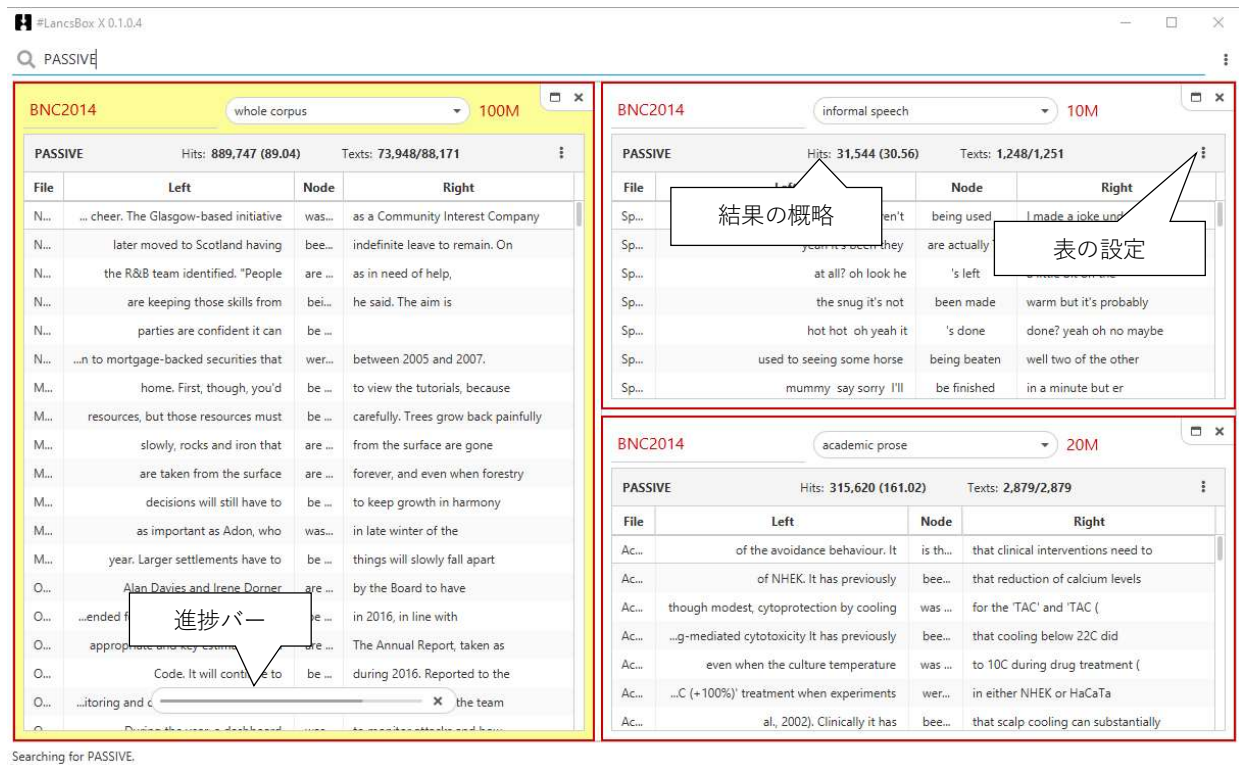
コーパスの選択

サブ・コーパスを選択

ヘッダーの左列をクリックして整列、ドラッグで並べ替え

+をクリックでパネルの追加

Search completed.



パネルはウィンドウ上部をクリック、ドラッグすることで並べ替えが可能です。

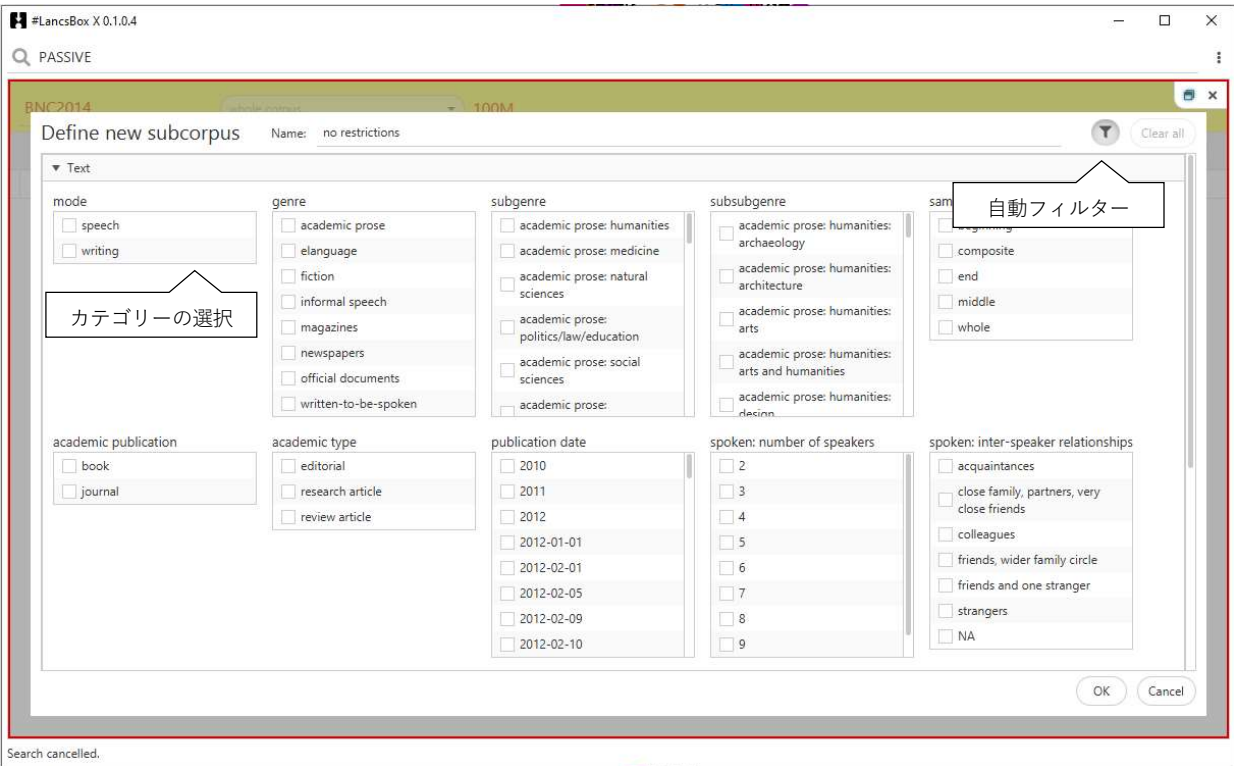
ツールをクリックする際、Ctrl、もしくはCmd キーを押しておくことで、複数のパネルを選択可能です。これによって、複数のパネルにおいて同じ検索を行うことが可能です。

表の行はクリックで選択することができます。Ctrl、またCmd キーを押しながら複数行について選択することが可能です。選択された行はCtrl+C / Cmd+C のショートカット、右クリックで「コピー (Copy)」を選択することでコピーできます。

表の中において列の追加、削除を行う際は表を右クリックで、サブ・メニューから「列の表示 (Show Columns)」を選択。

#LancsBox X ではサブ・コーパスの選択が可能です。これにより、検索をコーパスの特定の箇所の中に絞ることができます。新しいサブ・コーパスの追加には、サブ・コーパスのドロップダウン・リストから「新しいコーパス (New Subcorpus)」のオプションを選択します。

サブ・コーパス決定の基準、名前の設定はオーバーレイを開くことで実行可能です。完了後は「OK」をクリックします。これで新しいサブ・コーパスが選択可能になります。



サブ・コーパスのドロップダウン・リストを使ってサブ・コーパスの変更が可能です。ドロップダウン・リストの編集と削除ボタンでは、決定したサブ・コーパスの変更、削除ができます。

4 #LancsBox で検索を行う

#LancsBox では 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) CQL 検索を使った異なるレベルでのコーパスアノテーションを検索することができます。

1. シンプル検索 (Simple Searches) は特定の語 (e.g., *new*)、フレーズ (*New York Times*) について文字通りの検索を行います。
2. ワイルドカード検索 (Wildcard Searches) はアスタリスクの記号 (*) を含んだ検索です。

記号	意味	使用例
*	そのまま、もしくは続く文字 語彙が続く [スペースの後]	new* [<i>new, news, newly, newspaper...</i>] new * [<i>new car, New York, new ideas...</i>]

3. スマート検索 (Smart Searches) は#LancsBox 特有の機能です: 複雑な検索へと簡単にアクセスできるように、あらかじめ設定された検索を提供します。これらの検索は異なる品詞 (名詞 [NOUN]、動詞 [VERB])、異なる複雑な文法パターン (受け身 [PASSIVE]、分離 [SPLIT]、不定詞 [INFINITIVE] など)、意味的カテゴリー (副詞 [PLACE ADVERB]) を調べるのに使用が可能です。

以下のスマート検索は英語でのみ使用可能です。:

ADJECTIVE	LINKING ADVERB
ADVERB	LONG WORD
BE	MODAL
BOOSTER	NEGATION
COLLECTIVE NOUN	NOMINALIZATION
COMPARATIVE	NOUN
COMPLEX NOUN PHRASE	NUMBER
CONDITIONAL	PARTICLE
CONNECTOR	PASSIVE
CONTRACTION	PAST TENSE
DEGREE ADVERB	PAST PARTICIPLE
DETERMINER	PERFECT INFINITIVE
DO	PHRASAL VERB
DOWNTONER	PLACE ADVERB
EXISTENTIAL THERE	PREPOSITIONAL PHRASE
GERUND	PRESENT TENSE
HAVE	PRONOUN
INFINITIVE	PROPER NOUN
HYPHENATED WORD	REFLEXIVE PRONOUN
INDEFINITE PRONOUN	SHORT WORD
INFINITIVE	SPLIT INFINITIVE
INTERJECTION	SUPERLATIVE

SWEARWORDS
TIME ADVERB
VERB

4. CQL (コーパス検索言語検索: Corpus Query Language) 検索: #LancsBox は CQL を使った力強い検索をサポートしています。

CQL 検索は異なるレベルでのアノテーションにおいて、複雑な検索を定義するのに使用されます。

アノテーションと文構造のレベルはコーパスのタグ付けによりますが、XML のコーパスには通常 i) 語、ii) ヘッド・ワード / レマ (hw)、iii) POS (part-of-speech)、そして iv) ユーザーによって定義されたタグがあります。

```
[語(word) ="goes" ヘッド・ワード (hw) ="go" pos="V.*"]
```

上記はヘッド・ワード「Go」と「V (Verb: 動詞)」という POS タグを含む「Goes」という語の各例について、該当します。アノテーションのレベルが指定されていない場合、そのレベルでは制限なし、という適用になります。二重の括弧は大文字、小文字の判別をしない表現であるとみなされます。

複数のトークンについても、下記のように一連の中に収めることが可能です。空の大括弧 [] についてはどのトークンにも当てはまります。トークンは {X} という式を用いて X 回、{Y, Z} の式では Y から Z 回の繰り返しを示します。[0, 1] の省略は ? マークです。このため、以下の CQL コードは

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

制限のない 0 から 3 のトークン ([]{0,3}) が続く、動詞となるもの (VB: a verb to be) として読み込まれ、場合によっては過去分詞形 (V.N) が続きます。

検索の部分は括弧 () で括弧することもできます: これにより {1,2} のような数量表示を複数のトークンに適用することができます。(例: ([pos="N.* "] [word="and"]){2}。) 語、フレーズ、スマート検索は CQL トークンの有効な場所であれば、どこでも使用が可能です。(例: very{2} ADJECTIVE{1,2} [hw="year"]。)

CQL は XML 構造の検索についてもサポートしています。この検索では、各 <u>/<u> ユニットに当てはまります (<u>=発話 (Utterance)を示す)。以下は n について 1、国籍についてイギリス/アメリカという条件に該当します。

```
<u n="1" nationality="British|American"/>
```

これらの部分検索は、構造内において、ほかの種類の検索と組み合わせることができます。

```
[pos="D.* "] green NOUN within <text genre="newspapers"/>
```

以下の検索は新聞記事の中で、名詞に続く「green」の前に現れる限定詞の例に該当します。*Within* の左右にある検索についての指定はありません。これらは他の検索にもなりえます。

(<emoji/> within please) within (<e/> within <text genre="elanguage"/>)

5 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。

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。

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。

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'。

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。

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。

トークン (延べ語数- Token) – テキスト、コーパスにおける総語数

XML – Extensive Markup Language (拡張可能なマークアップ言語) の略称。これはテキストファイルにおいて、機械で読み込み可能であり、情報にアノテーション、構造的な側面を与えています。XML は語について POS 情報とともにアノテーションすることができ、例えばセクション、段落ごとといった構造による区分けを行うことができます。