Sketch Engine 操作マニュアル

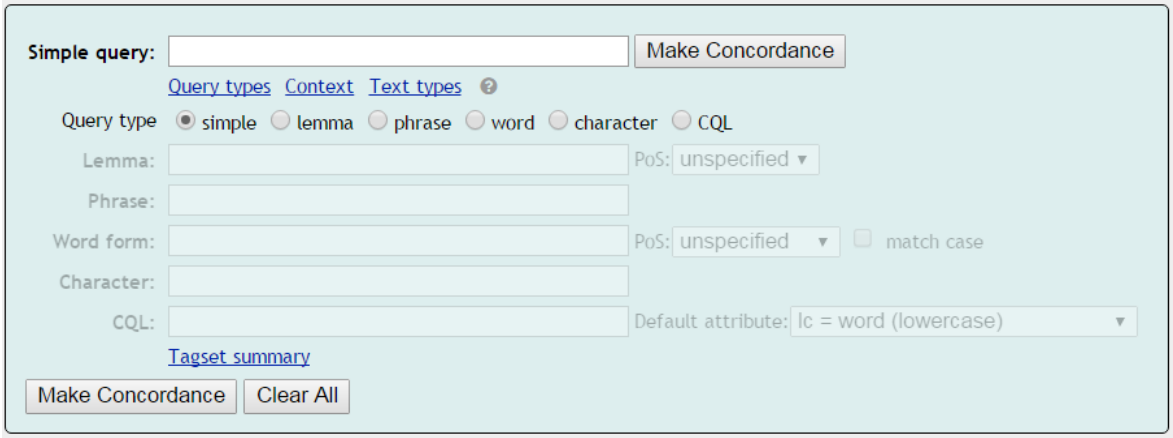
東京外国語大学　投野研究室

＜目次＞

１．　Concordance 画面を使いこなす

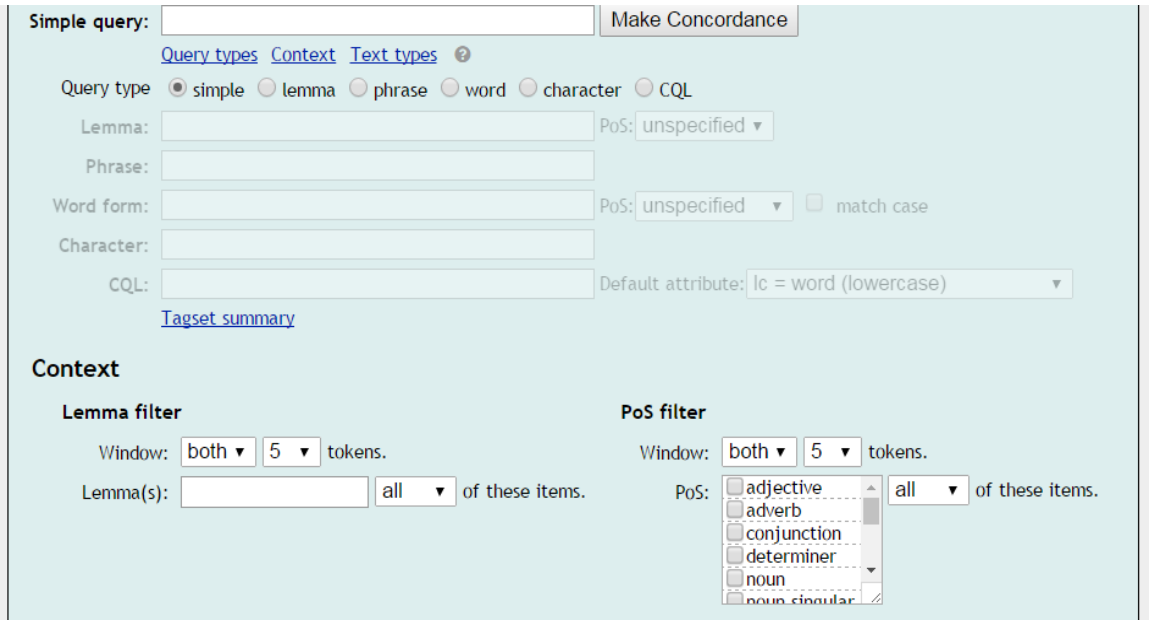
１．１．　検索の基礎

（１）Query Type を選ぶ



・Simple query…単語、フレーズ検索（各単語の活用形も含む＝lemma 検索）
・Lemma…単語のみ。活用形も含む。品詞指定可。
・Phrase…表層形のみのフレーズ検索（＝活用形は含まれない）。
・Word…表層形のみの単語検索。品詞指定可。大文字・小文字指定可。
・Character…特定のアルファベットの並びで検索（＝接頭・接尾語の検索可）。

（２）Context



Lemma Filter…前後〇語以内の指定の単語（lemma で）の有無によって絞り込み。

PoS Filter…前後〇語以内の指定の品詞の有無によって絞り込み。

※いずれも複数指定可で、all にするとその全てが含まれたもの、any にするとそのうち少なくとも 1 つが含まれたもの、none にするとそのうちどれも含まれないものがコンコーダンス上に現れる。

（3）Text types

…サブコーパスによる絞り込みができる。サブコーパスを編集することもできる。

（4）Query 画面の活用

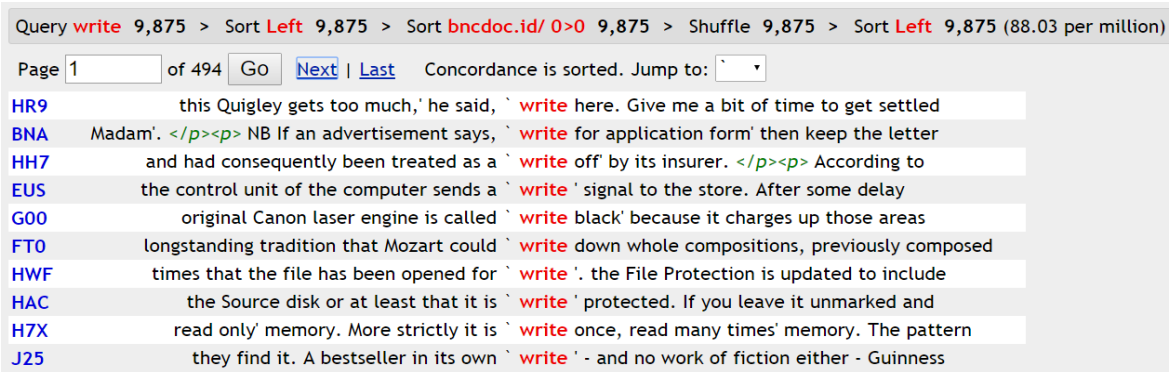
・Sort（並べ替え）

Left…中心語の左隣（L1）の語のアルファベット順で並べ替え

Right…中心語の右隣（R1）の語のアルファベット順で並べ替え

Node…中心語のアルファベット順で並べ替え

References…ファイル情報で並べ替え



↑ファイル情報　　↑L1　↑node　↑R1

・Sample（標本抽出）…ランダムに指定した数のコンコーダンスを出すことができる。

・Filter（絞り込み）

指定した語が中心語の前後〇語に現れる（positive）もしくは現れない（negative）ものでコンコーダンスをさらに絞り込むことができる。

・Frequency（頻度集計）

Frequency…中心語のみだけでなく、前後〇番目にある単語の指定した形での頻度集計ができる。

Node tags…中心語の時制ごとに頻度集計ができる。

Node forms…中心語の表層形ごとに頻度集計ができる。

Doc IDs…ファイルの種類ごとに頻度集計ができる。

Text Types…テキストの種類ごとに頻度集計ができる。

→さらに P/N (Positive/Negative)でコンコーダンスを絞り込むことができる。

・Collocation（共起語）

Attribute…共起語を word/tag/lempos/lemma…のどれによって分類するか決定。

Range…共起するのが中心語の前後〇語以内にするか決定。

Minimum Frequency in corpus…共起語のコーパス内の総頻度の下限を決定。

Minimum Frequency in given range…中心語＋共起語の頻度の下限を決定。

→さらに P/N (Positive/Negative)でコンコーダンスを絞り込むことができる。

・Visualize（グラフ化）

…全コーパス内における中心語の分布をグラフ化できる。全コーパスをどれくらいの束に分けるかは、グラフ下の数値で調節。

１．２．CQL 検索

機能：CQL(= Corpus Query Language)を使った検索式により、柔軟な検索ができる
CQL：1990 年代初期に独 University of Stuttgart の IMS が開発

✧実際に見てみましょう　ー　対象コーパス：BNC✧

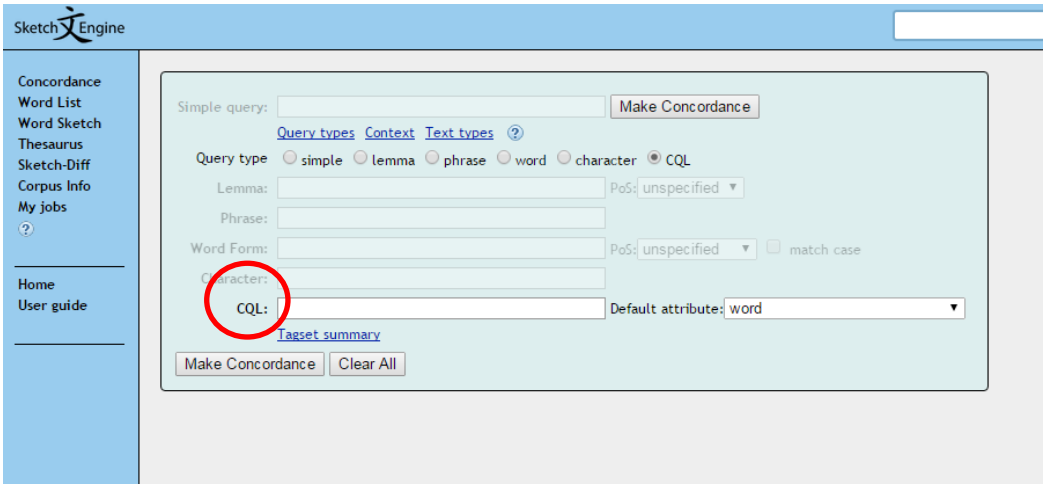


図１　検索画面

CQL の一般式：　[attribute=”value”]
※attribute に word/lemma/tag/lempos がはいる
※value に検索したい正規表現を含む文字列がはいる

turn（表層形）を検索する場合の式
[word=”turn”]
″[tT]hank″

turn（レマ）を検索する場合の式
[lemma=”turn”]

turn の名詞を検索する場合の式
[lemma=”turn” & tag=”N.*”]　または　[lempos=”turn-n”]

※タグの記号に関しては図１CQL のすぐ下の Tagset summary を参照
※　.(ピリオド)：任意の１文字
※　*：０以上ｎ個　/　+：１以上ｎ個　/　？：０個または１個　/！：～以外

tag=”　” を使った他の複雑な例

″confuse.*″ [tag=″IN″ | tag=″PP″]

″confuse.*″ ([tag=″IN″] | [tag=″PP″])

″confuse.*″ [tag=″IN|PP″]

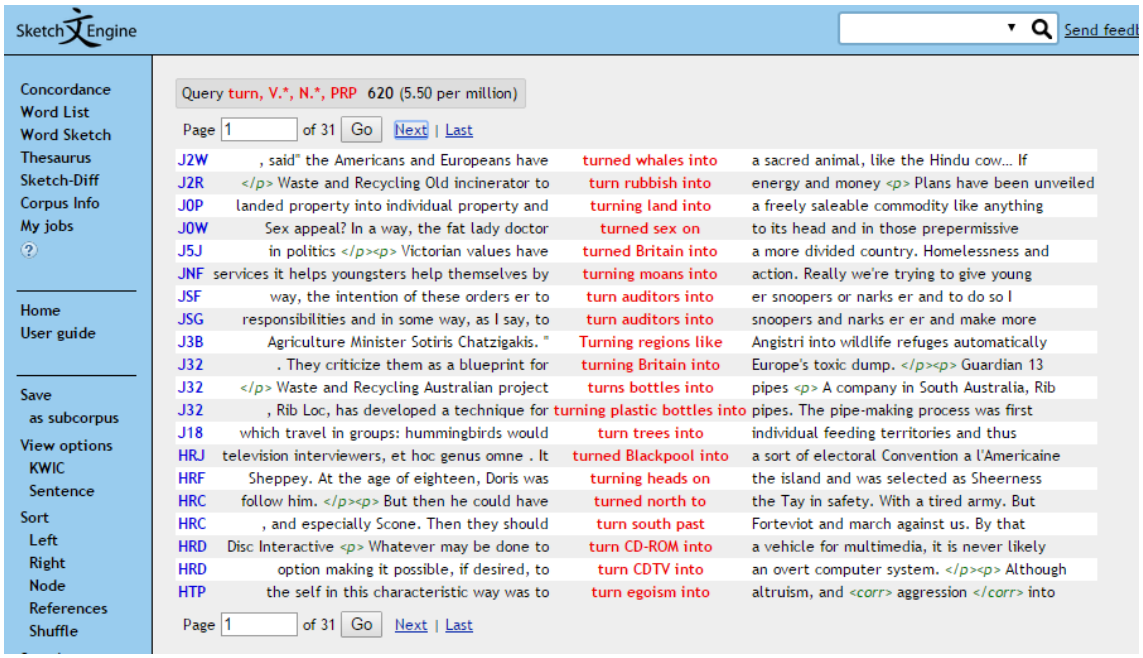
turn + 名詞 + 前置詞　の検索式

[lemma=” turn” &tag=” V.*”][tag=” N.*”]{1,2}[tag=” PRP”]

※[ ]と[ ]の間にはスペースがあってもなくても ok

※{ }で検索スパンを指定（{1,2}は１語から２語という意味）

この式で CQL 検索をしてみると・・・



でた～～～　図２

✧複雑な検索式✧

“help to do” vs. “help do”のコンテクスト差を調べる為、文を抽出する式

[lemma=″help″&tag=″V.*″][word=″to″]？[tag=″V.I″]

※動詞の不定詞形タグは VBI (be)、VDI (do)、VHI (have)、VVI (lexical verbs)

名詞 + be + -ed 形で終わる動詞　の検索式

[tag=” N.*”][lemma=” be”][tag=” V.*” &word=” .*ed”]

look/bring + up/down の検索式

[lemma=” look|bring” &tag=” V.*”][tag!=” V.*”]{0,5}” up|down”

OR の意味のバー（｜）の例

[tag=″JJ.*″] [tag=″N.*″] ″and|or″ [tag=″N.*″]

✧within を使った検索式✧

文境界を指定した式

[word=” confus.*”][tag!=” V.*”]*[word=” by”]within<s/>

動詞で始まり、動詞で終わる連鎖の中にあるすべての名詞句を抽出する式

[tag=” N.*”]+within[tag=” VB.*”][]*[tag=” VB.*”]

（動詞句を一緒にとってこないための式）

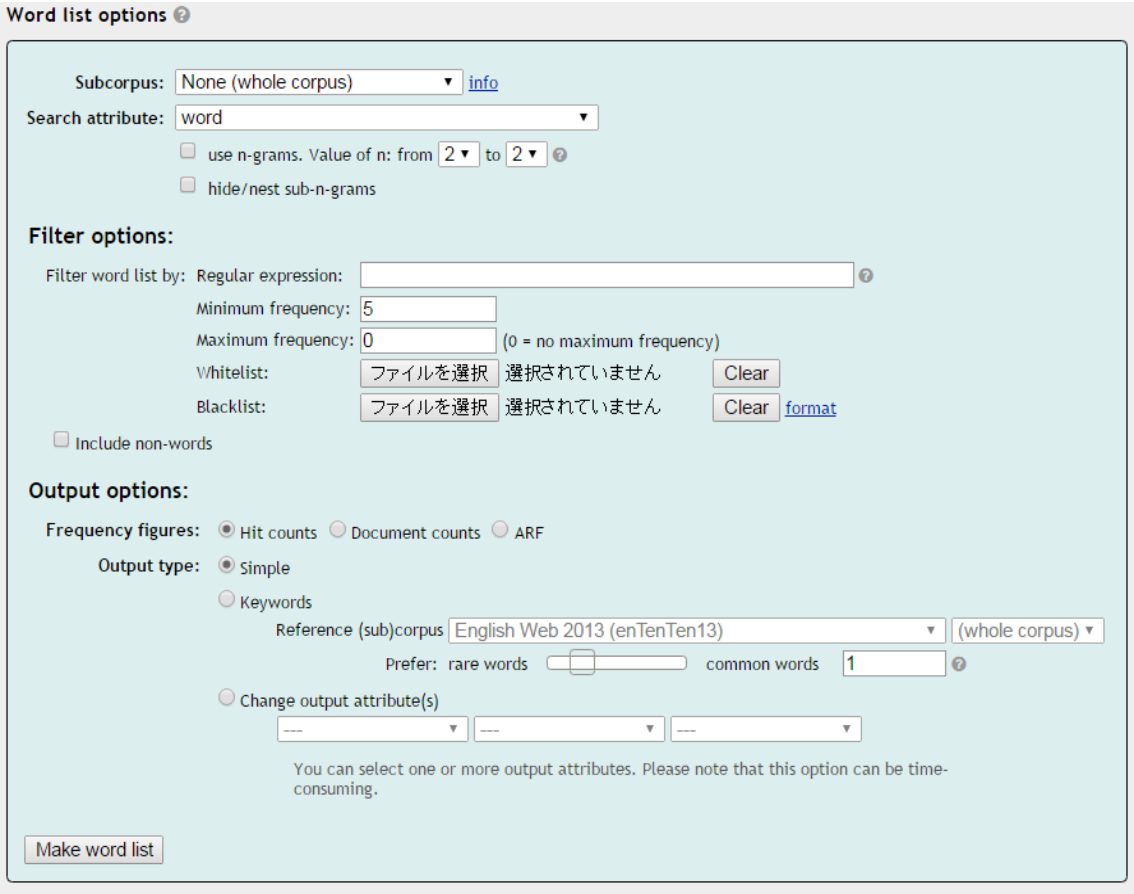
※ Query within Query という式も可能である

※ Containing もある（cf. Sketch Engine）

２．さまざまなオプション機能

２．１．Word list

コーパス（サブコーパス）内から語彙表を抽出できる。



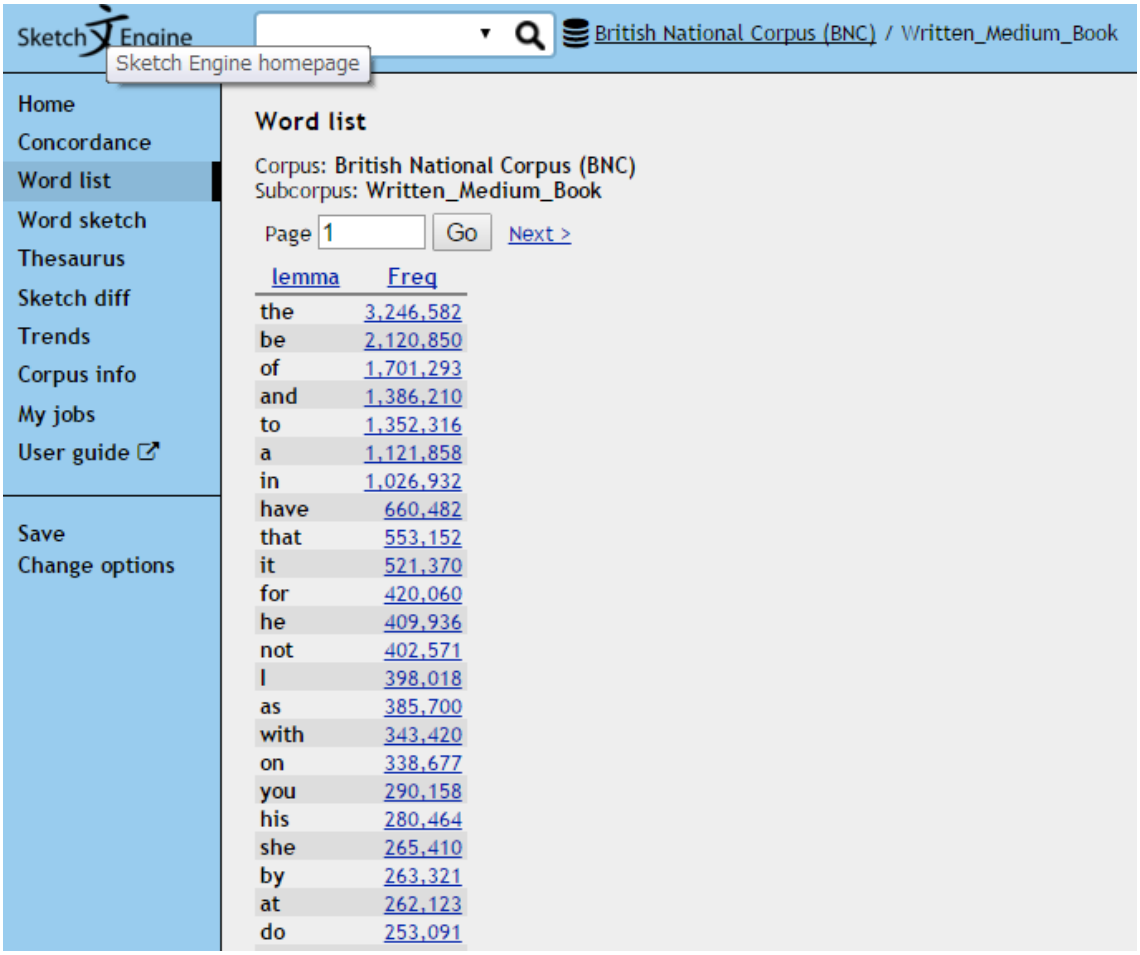
上から順に

●Subcorpus：サブコーパスを選択、新たに作成できます。

●Search attribute: word, lemma, tag (POS)などが選べます。

Use n-grams では n 語の連鎖の語彙表を作成できます。

ここまでで検索してリストを作成できます。また以下のオプションを使うこともできます。ここで、コーパスを BNC、サブコーパスを Written_Medium_Book、Search attribute を lemma にして word list を作成すると以下のようになります。



（頻度が高い順に lemma を並べた語彙表）

【Filter options】

●Regular expressions: 正規表現で検索できます。 .* がワイルドカード（何が何文字入っても OK）を表すので、「th.*」で検索すると the, that, this 等の語彙表が作成されます。（その他、+、?、！などがあります）

●Minimum frequency: 最小頻度を指定できます。

●Maximum frequency: 最大頻度を指定できます。

●Whitelist: 語彙表に含めたい特定の単語リストがある場合、アップロードできます。

●Blacklist: 語彙表に含めたくない特定の単語リストがある場合、アップロードできます。

●Include non-words: 句読点や記号などを含めたいときに使います。

【Output options】

- Frequency figures: Hit counts　→粗頻度 (= raw frequency)

  Document counts　→語彙表中の単語を含むドキュメントの数

  ARF (Average Reduced Frequency)　→ひとつの単語が近距離（e. g. 同じドキュメント内）で複数回現れるときに頻度を調節する機能です。

- Output type: Simple Keywords　→他の（サブ）コーパスを参照して比較する場合にキーワードを抽出できます。

- Reference (sub)corpus　→比較するサブコーパスを選択できます。

- Prefer: rare/common words　→頻度が高い・低い単語に高いスコアが当てられるように調節できます。

- Change output attribute(s):

【実際に検索してみよう】

①BNC 全体で search attribute を pos、 minimum frequency を 0

②BNC の中のサブコーパス Written_Domain_Imaginative で search attribute を lemma、regular expressions を wh.*、minimum frequency を 1、maximum frequency を 0

③BNC の中のサブコーパス Written_Domain_Informative で search attribute を word、regular expression を .*ing、Frequency figures を Document counts

④BNC 全体で search attribute を word、use n-grams で n=4

⑤BNC の中のサブコーパス Written_Medium_Book で search attribute を word、Output type を Keywords にして、Reference subcorpus を BNC のサブコーパス Written_Medium_To-be-spoken

⑥BNC 全体で search attribute を word、regular expression を .*ing、Output type を Change output attributes にして、lemma, pos, word を選択

２．２．Word sketch

◎Word sketch とは？

ある語がコーパス内でどのような語と共起しているかを検索できる機能。

◎Word sketch の基本操作

１．検索したいコーパスを選択。

２．Lemma に検索したい単語、Part of speech で品詞を選択。

３．通常はこのまま Show word sketch をクリック（Advanced options で詳細な設定も可能）。

４．青いラベル（図１）：順に

・文法関係

・用例全体における頻度

・構文全体の中でどのくらい特徴的な構造なのかを示すスコア値（他の動詞の同一パターンと比較）→スコア値が大きいほどその単語における特徴的な文法構造といえる。

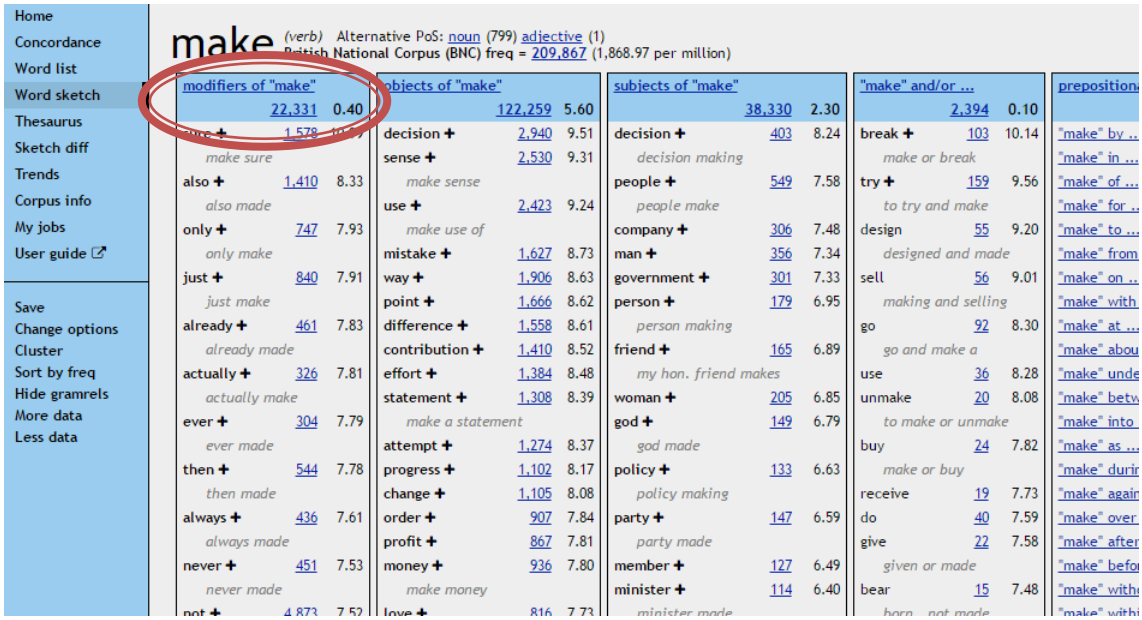


図１：make の検索結果

５．その他のオプション（図１の左側の列）：上から

・Change options…Word Sketch のホーム画面に戻る

・Cluster…クラスター分析を行う

・Sort by freq/Sort by score…頻度順あるいはスコア順でソートを行う

・Hide gramrels…文法項目関係なしに全体をランキング化

・More data…1 column における表示データの量が増加

・Less data…1 column における表示データの量が減少

◎Word sketch の具体的な使用例

★動詞 make の文法関係を自動抽出する。

１．コーパスは BNC を選択。

２．Lemma に make、Part of speech で verb を選択。

３．検索結果は図２参照。

４．adjectives after “make” and noun のスコア値が最も高い。

→make は make+O+C の構文が特徴的といえる。

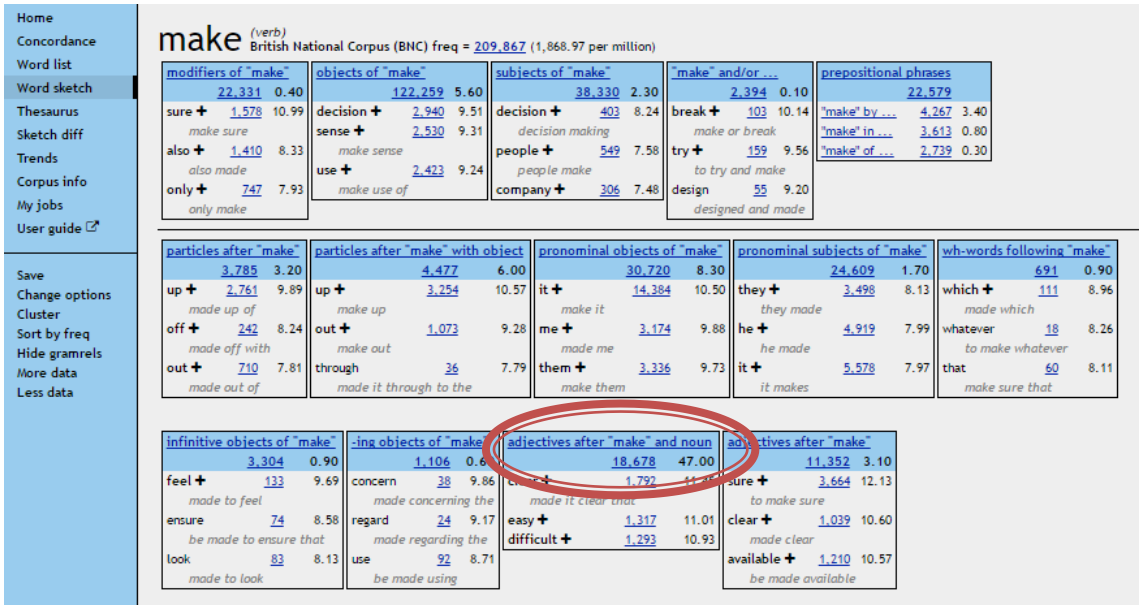


図２：make の検索結果（※Less data で表示データ数を減らしたもの）

２．３．Thesaurus / Sketch diff

機能：同義語、あるいは異なる２語間の共起関係を調べることができる

✧ 実際に見てみましょう ― 検索対象コーパス：BNC ✧

検索画面（名詞の"love”で検索、POS を指定）

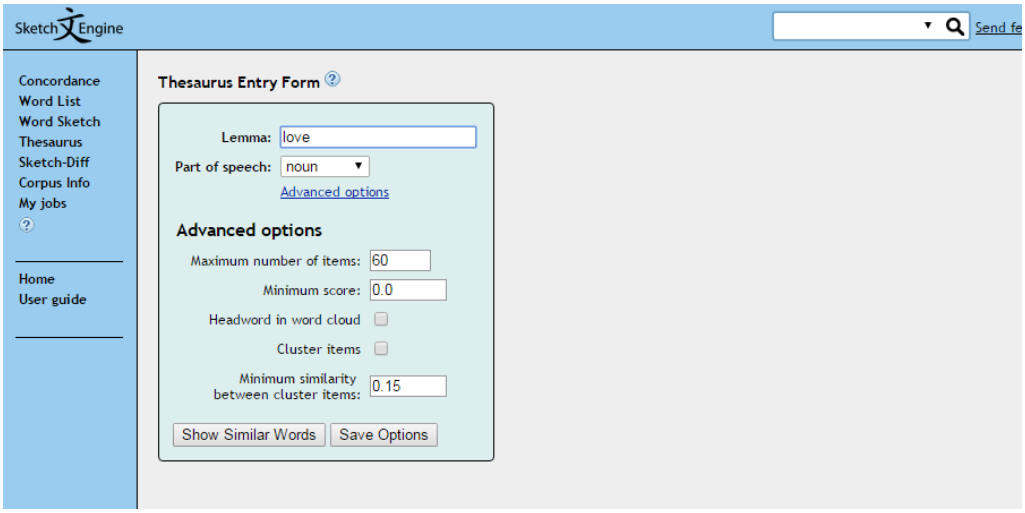


図１

検索結果

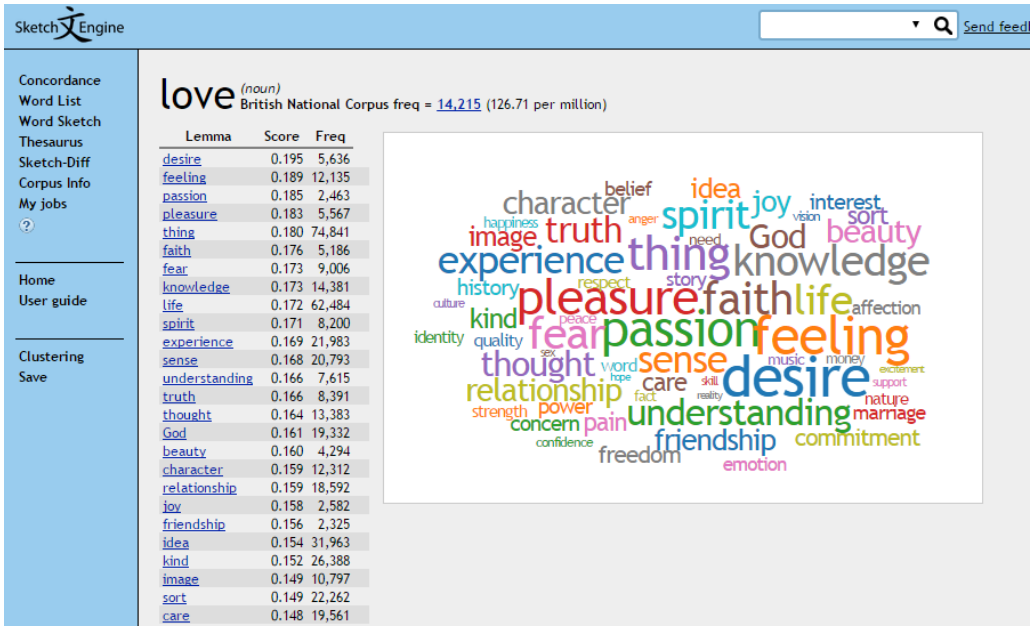


図２（下に結果が続く）

検索をすると、図２にあるように名詞 love の類義語リストが表示されるだけでなく、視覚的にも分かりやすく同義の語彙を表示してくれる。（大きさが大きい単語ほど、検索語彙により近い）
必ずしも類義語とは言えない単語も入っているが、単語の振る舞いが似ているものが提示される。

リストの一番上、かつ大きさ最大の desire のリンク先

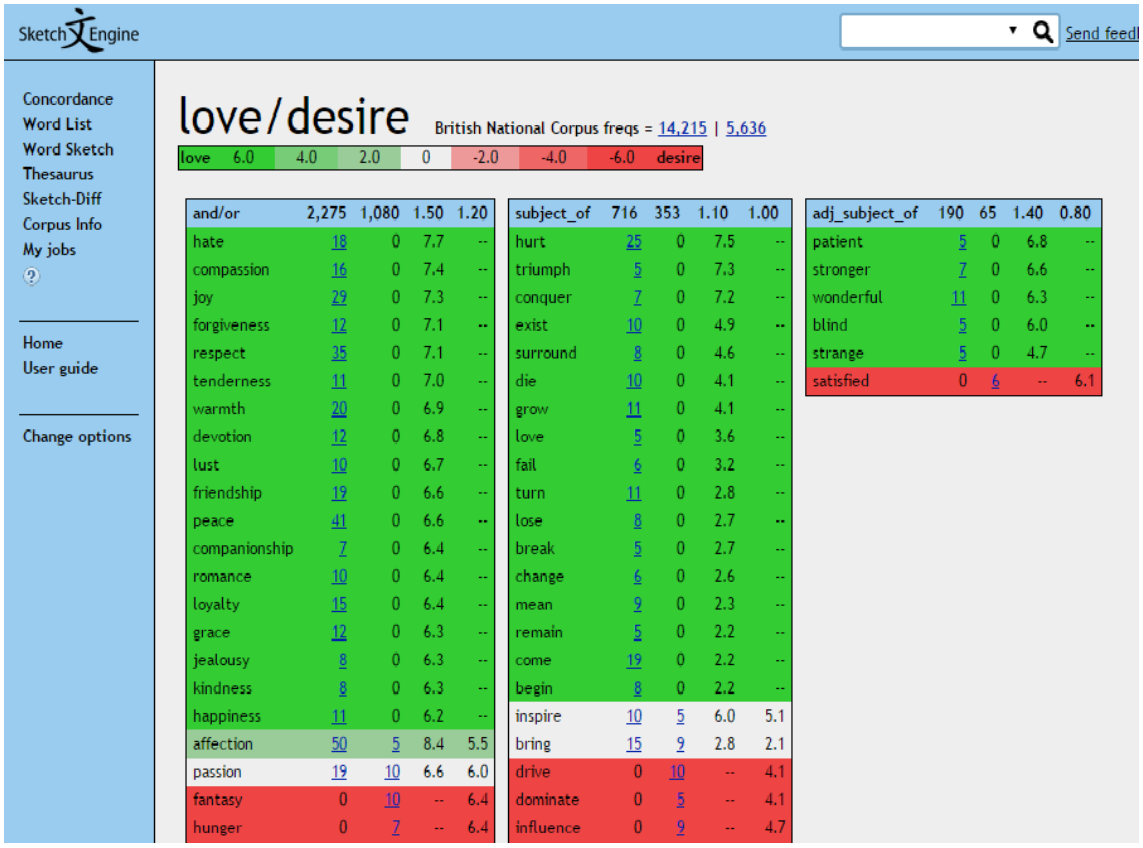


図３（下に結果が続く）

Word sketch 機能の２単語比較バージョンで、Thesaurus からだけでなく、メニュー左の Sketch-Diff からもこの機能を使うことができる。

この図３では、名詞 love と desire の違いをコロケーションの観点から学び取られる。

✧ 左から、 and/or で並列共起する名詞、主語のときに共起する動詞、そして修飾を受けて共起する形容詞の３つの観点から love と desire の違いが分かる
（他にもたくさんの違いを示す表が下に続く）

✧ 緑であればあるほど love との共起が多く、赤であるほど desire との共起が多い
（ここで love には satisfied は修飾されていない・・・ということは・・）

３．コーパスの作成

３．１．ファイルのアップロード

☆自分がコーパス化したいパソコン上のテキストファイルをスケッチエンジンにアップロードすると、既存のコーパスと同じようにスケッチエンジン上で検索できるようになります。

〈テキストの整形〉アップロードしたいファイルに xml タグをつける（適宜）。
xml → 情報を記述するためのタグ

```
<bookstore>
  <book category="COOKING">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>
  <book category="CHILDREN">
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
  <book category="WEB">
    <title lang="en">Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
```

＜w3schools.com(http://www.w3schools.com/xml/xml_tree.asp)より＞

【タグ付与例】

(i)ヘッダー（開始タグ）

<xml>

<doc Level=”○○” Title=”テキスト名”>

<text Types=”○○”>

(ii) フッター（終了タグ）

</text>

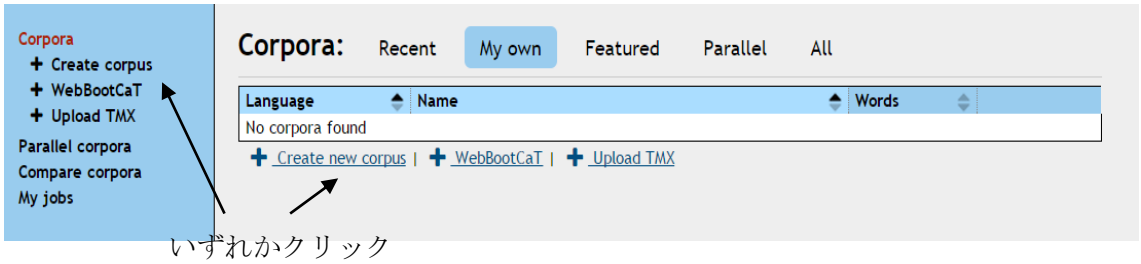
</doc>

</xml>

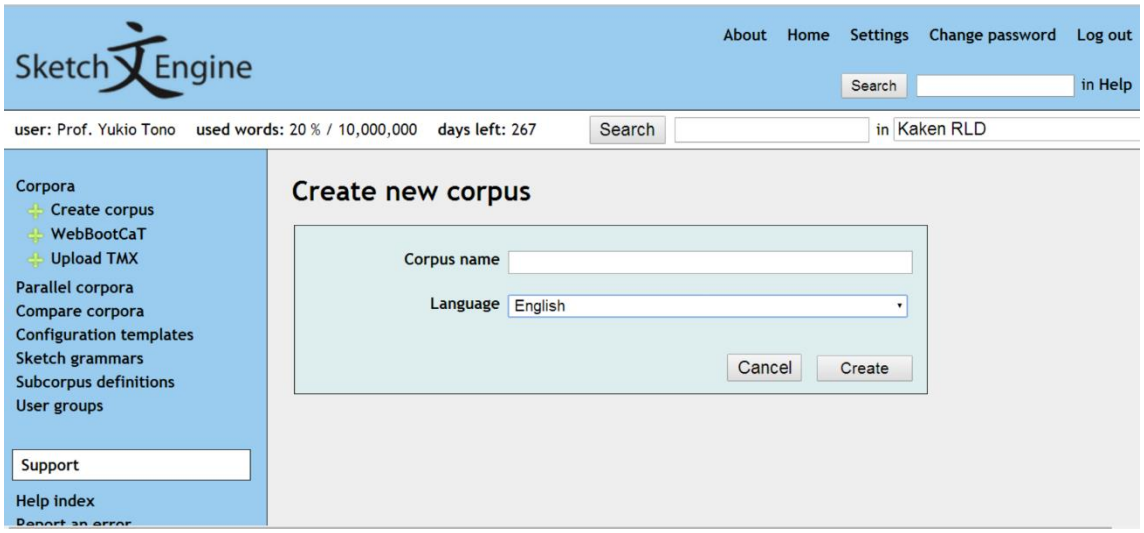
このようにテキスト内にコンピューターが理解できるような形で情報を付与すると、アップロードした後に自動的にサブコーパスに分けてくれたり、検索結果画面でテキストの細かい情報が表示できたりするようになります！

〈スケッチエンジンへのアップロード〉

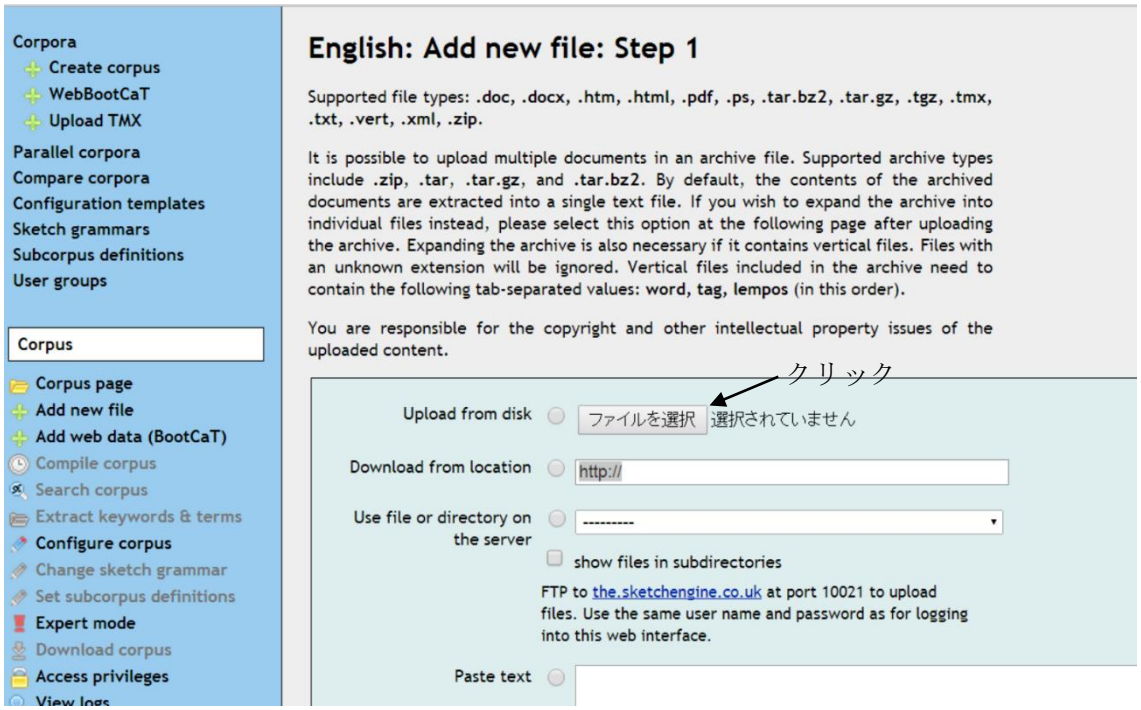
(i) 新規のコーパスを作成する



(ii) コーパスの名前と扱う言語の設定

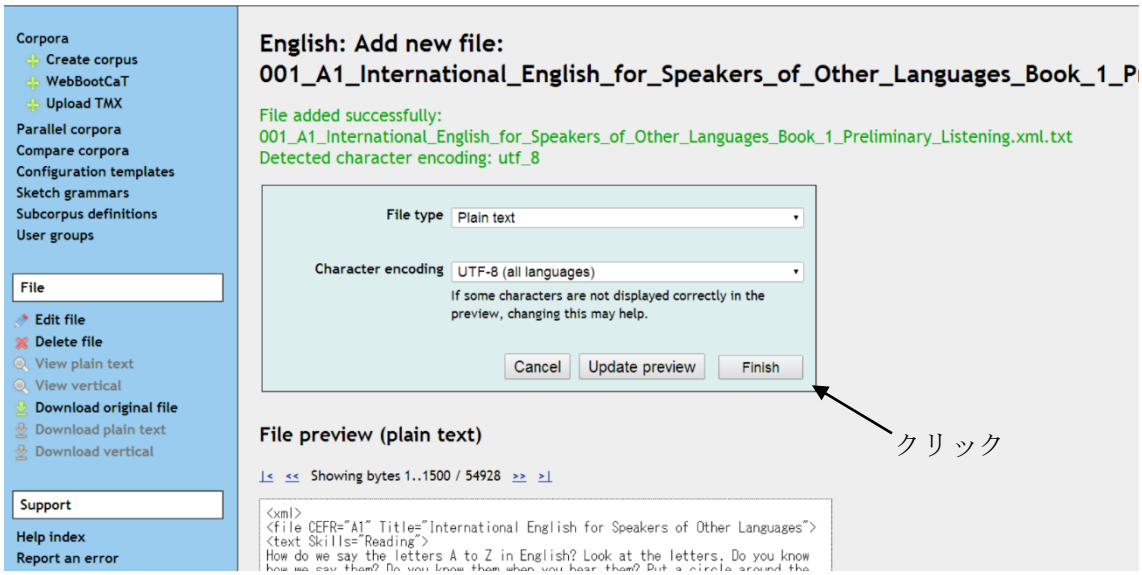


(iii)ファイルの選択

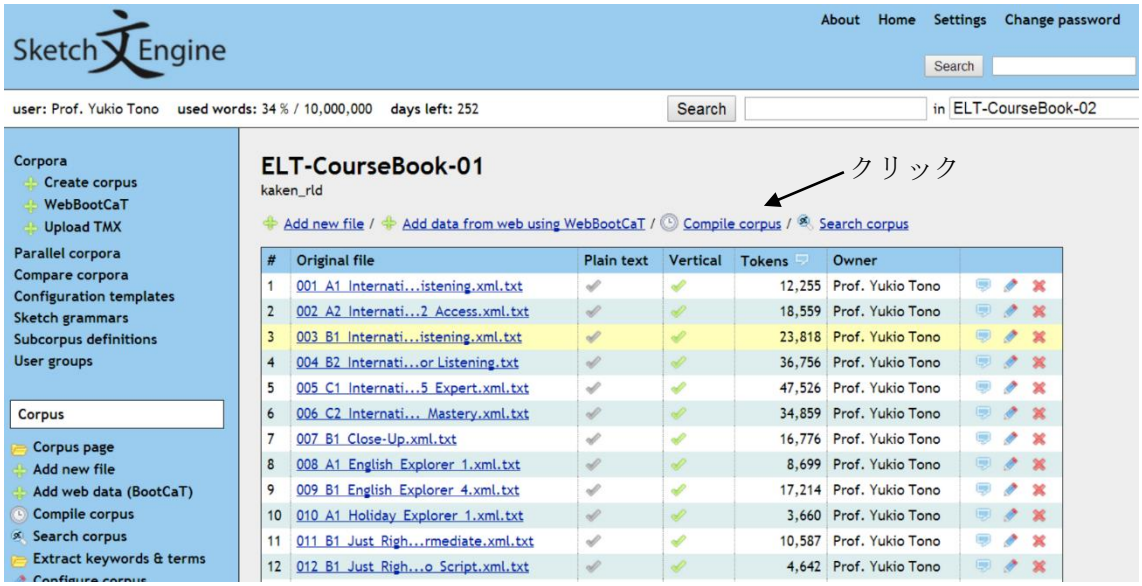


→ライブラリからファイルを選択し、Next をクリックしてアップロード

(iv)ファイルの種類と文字コードを選択



(v) すべてのファイルをアップロードし終えたらコーパスを compile する



３．２．WebBootCaT

◎WebBootCaT とは？

インターネットをクロールしてテキストを自動収集し、コーパスを作成する機能。

◎WebBootCaT の基本操作（図１参照）

１．Home 画面の左側のメニューから WebBootCaT をクリック。

２．コーパス名と言語を設定。

３．Input type で Seed words/URLs のどちらかを選択。（後述）

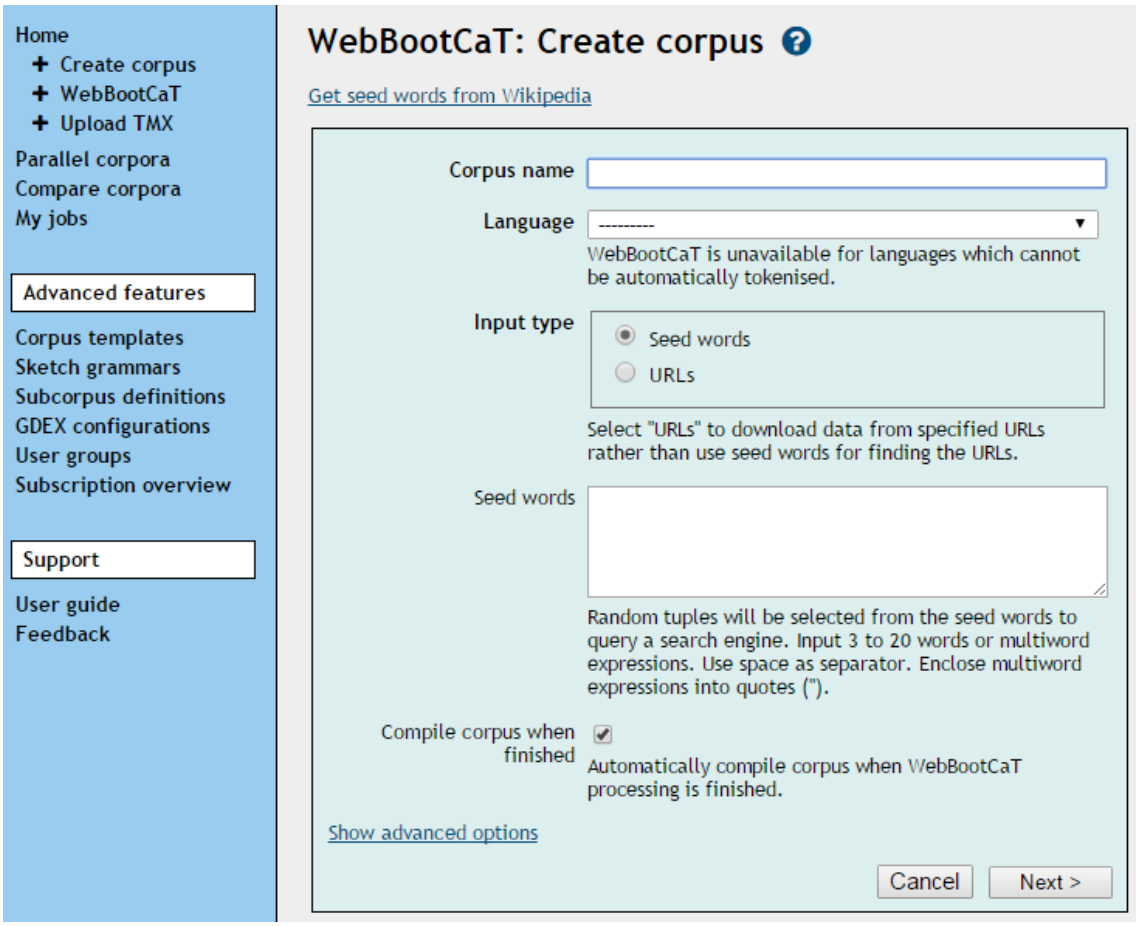


図１：WebBootCaT の基本画面

◎WebBootCaT の具体的な使用例（Seed words/URLs）

★ “Seed word”で Web corpus を作成する。

１．上記の 1~2 を経て、Input type の Seed words にチェックを入れる。

２．Seed words の欄にキーワード（３～２０個）を入力。

３．Seed words を３語ずつランダムに組み合わせたものをインターネット検索にかけた結果が表示される（図２参照）。

４．Next をクリック→自動的にテキストをダウンロード。

５．OK をクリック→コーパスの完成。

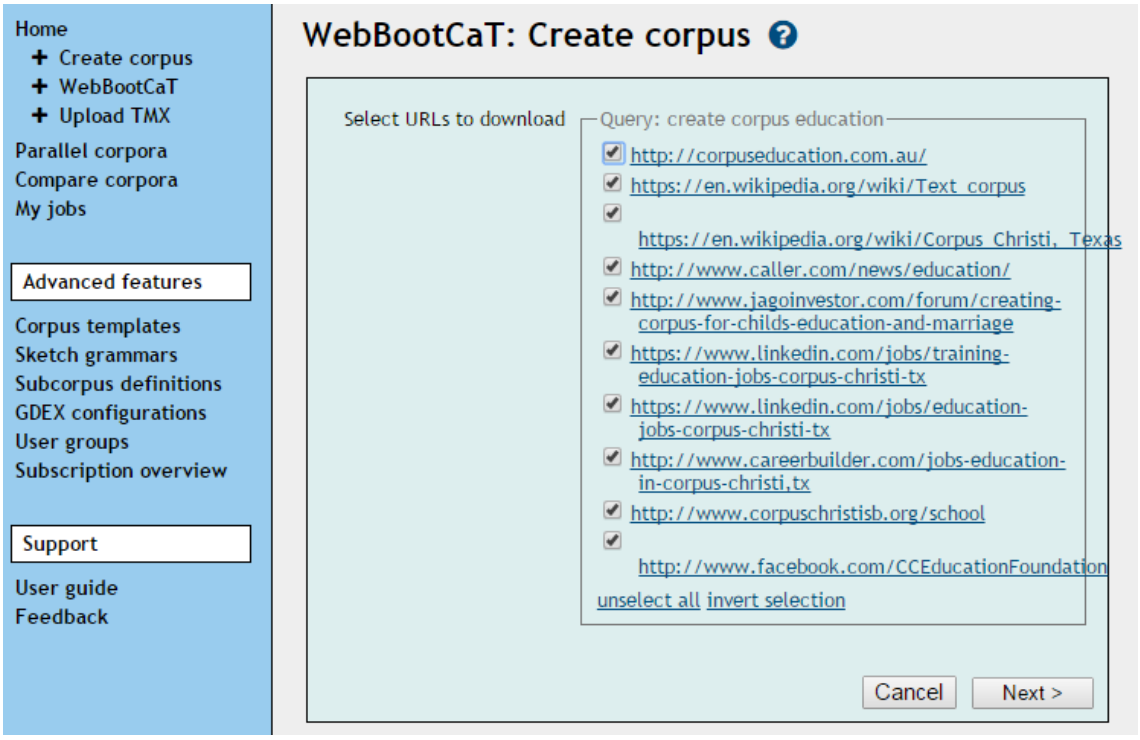


図２：URL の一覧

★ “URLs”で Web corpus を作成する。

１．上記の 1~2 を経て、Input type の URLs にチェックを入れる。

２．URLs の欄に指定する URL を入力（図３参照）。

３．Next→OK→コーパスの完成（図４参照）。

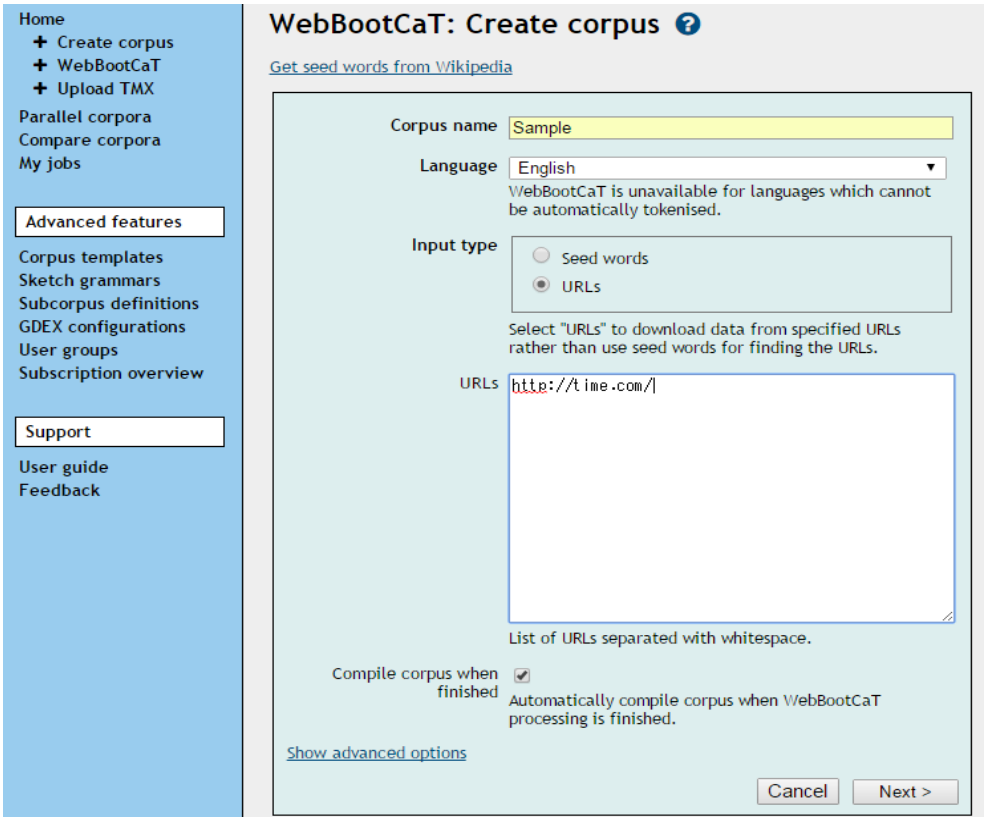


図３：URL の指定

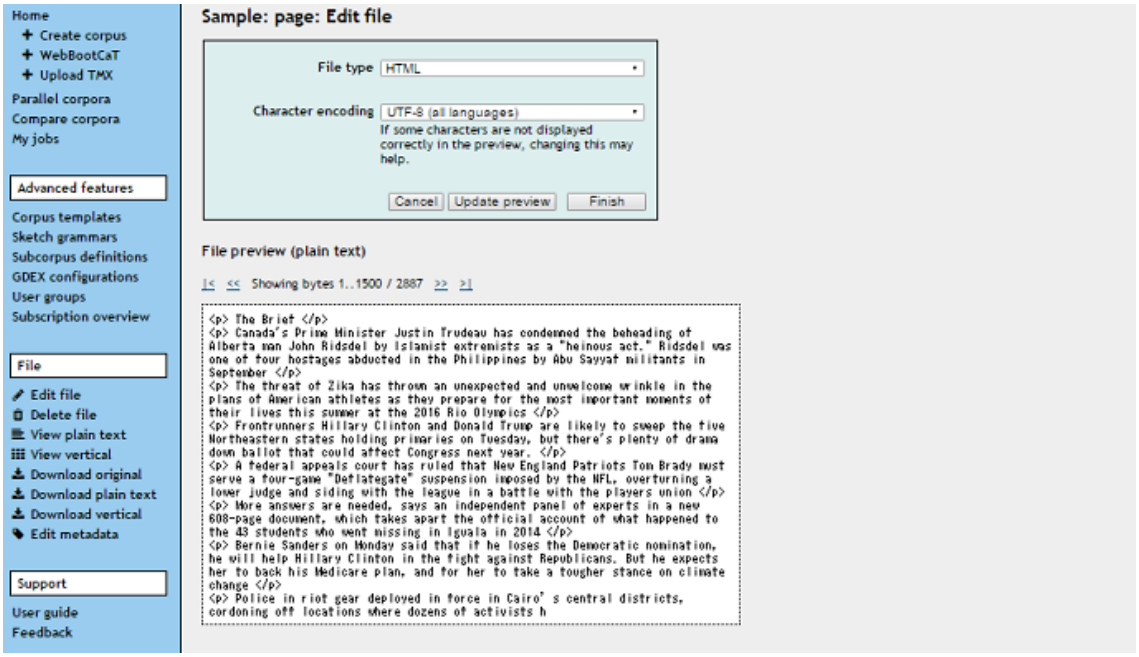


図４：完成したコーパス