

# 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.....p.2
  1. 2. CQL 検索 .....p.5
2. さまざまなオプション機能
  2. 1. Word list.....p.8
  2. 2. Word sketch.....p.11
  2. 3. Thesaurus / Sketch diff.....p.13
3. コーパスの作成
  3. 1. ファイルのアップロード.....p.15
  3. 2. WebBootCaT.....p.19

# 1. Concordance 画面を使いこなす

## 1. 1. 検索の基礎

### (1) Query Type を選ぶ

The screenshot shows the top part of the Concordance search interface. It includes a 'Simple query' input field, a 'Make Concordance' button, and navigation links for 'Query types', 'Context', and 'Text types'. The 'Query type' section has radio buttons for 'simple', 'lemma', 'phrase', 'word', 'character', and 'CQL'. Below this are input fields for 'Lemma', 'Phrase', 'Word form', and 'Character', each with a 'PoS' dropdown menu. There is also a 'match case' checkbox and a 'Default attribute' dropdown set to 'lc = word (lowercase)'. At the bottom, there are 'Make Concordance' and 'Clear All' buttons, and a 'Tagset summary' link.

- **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**…表層形のみフレーズ検索（=活用形は含まれない）。
- **Word**…表層形のみ単語検索。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

This screenshot shows the 'Context' section of the Concordance search interface. It features two filter sections: 'Lemma filter' and 'PoS filter'. Both filters have a 'Window' dropdown set to 'both' and a token count dropdown set to '5'. The 'Lemma filter' has a 'Lemma(s)' input field and a dropdown set to 'all'. The 'PoS filter' has a list of parts of speech (adjective, adverb, conjunction, determiner, noun, noun singular) with checkboxes and a dropdown set to 'all'. The top part of the interface, including the query type selection, is also visible.

Lemma Filter…前後○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○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

Query write 9,875 > Sort Left 9,875 > Sort bncdoc.id/ 0>0 9,875 > Shuffle 9,875 > Sort Left 9,875 (88.03 per million)

Page 1 of 494 Go Next Last Concordance is sorted. Jump to: ▾

HR9	this Quigley gets too much,' he said, `write here. Give me a bit of time to get settled
BNA	Madam'. </p><p> NB If an advertisement says, `write for application form' then keep the letter
HH7	and had consequently been treated as a `write off' by its insurer. </p><p> According to
EUS	the control unit of the computer sends a `write ' signal to the store. After some delay
G00	original Canon laser engine is called `write black' because it charges up those areas
FT0	longstanding tradition that Mozart could `write down whole compositions, previously composed
HWF	times that the file has been opened for `write '. the File Protection is updated to include
HAC	the Source disk or at least that it is `write ' protected. If you leave it unmarked and
H7X	read only' memory. More strictly it is `write once, read many times' memory. The pattern
J25	they find it. A bestseller in its own `write ' - and no work of fiction either - Guinness

↑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○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○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○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 — 対象コーパス：BNC◇

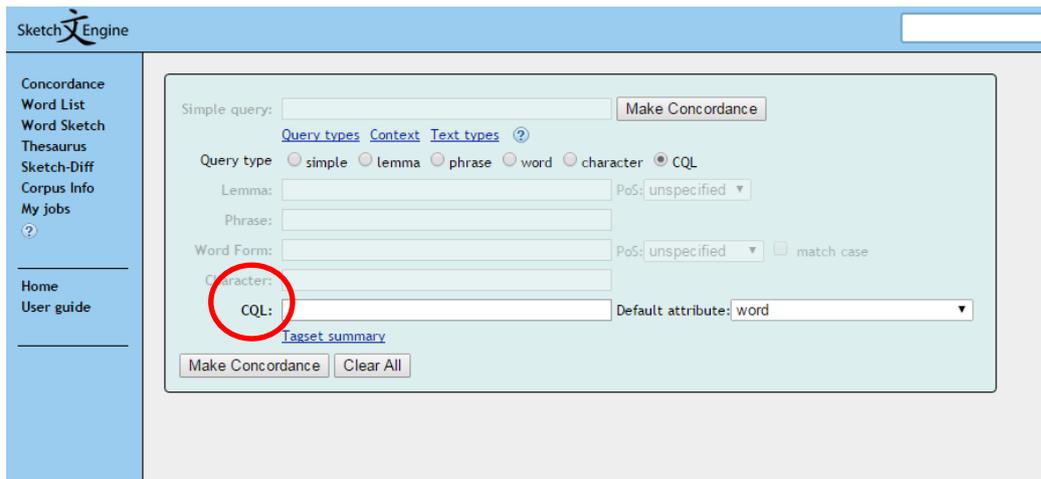


図1 検索画面

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1 CQL 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・・・

The screenshot shows the Sketch Engine interface with a search query: "turn, V.\*, N.\*, PRP" (620 results, 5.50 per million). The results table shows various examples of the word "turn" used in different contexts, with the phrase "turn ... into" highlighted in red. The left sidebar contains navigation options like Concordance, Word List, and Thesaurus. The bottom of the page shows pagination: Page 1 of 31.

Code	Text	Phrase	Text
J2W	, said" the Americans and Europeans have	turned whales into	a sacred animal, like the Hindu cow... If
J2R	</p> Waste and Recycling Old incinerator to	turn rubbish into	energy and money <p> Plans have been unveiled
J0P	landed property into individual property and	turning land into	a freely saleable commodity like anything
J0W	Sex appeal? In a way, the fat lady doctor	turned sex on	to its head and in those prepermissive
J5J	in politics </p><p> Victorian values have	turned Britain into	a more divided country. Homelessness and
JNF	services it helps youngsters help themselves by	turning moans into	action. Really we're trying to give young
J5F	way, the intention of these orders er to	turn auditors into	er snoopers or narks er and to do so I
J5G	responsibilities and in some way, as I say, to	turn auditors into	snoopers and narks er er and make more
J3B	Agriculture Minister Sotiris Chatzigakis. "	Turning regions like	Angistri into wildlife refuges automatically
J32	. They criticize them as a blueprint for	turning Britain into	Europe's toxic dump. </p><p> Guardian 13
J32	</p> Waste and Recycling Australian project	turns bottles into	pipes <p> A company in South Australia, Rib
J32	, Rib Loc, has developed a technique for	turning plastic bottles into	pipes. The pipe-making process was first
J18	which travel in groups: hummingbirds would	turn trees into	individual feeding territories and thus
HRJ	television interviewers, et hoc genus omne . It	turned Blackpool into	a sort of electoral Convention a l'Americaine
HRF	Sheppey. At the age of eighteen, Doris was	turning heads on	the island and was selected as Sheerness
HRC	follow him. </p><p> But then he could have	turned north to	the Tay in safety. With a tired army. But
HRC	, and especially Scone. Then they should	turn south past	Forteviot and march against us. By that
HRD	Disc Interactive <p> Whatever may be done to	turn CD-ROM into	a vehicle for multimedia, it is never likely
HRD	option making it possible, if desired, to	turn CDTV into	an overt computer system. </p><p> Although
HTP	the self in this characteristic way was to	turn egoism into	altruism, and <corr> aggression </corr> into

でた〜〜〜 図2

◇複雑な検索式◇

“help to do” vs. “help do” のコンテキスト差を調べる為、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。

Word list options

Subcorpus: None (whole corpus) [info](#)

Search attribute: word

use n-grams. Value of n: from 2 to 2

hide/nest sub-n-grams

**Filter options:**

Filter word list by: Regular expression:

Minimum frequency: 5

Maximum frequency: 0 (0 = no maximum frequency)

Whitelist:  選択されていません

Blacklist:  選択されていません  [format](#)

Include non-words

**Output options:**

Frequency figures:  Hit counts  Document counts  ARF

Output type:  Simple  Keywords

Reference (sub)corpus: English Web 2013 (enTenTen13) (whole corpus)

Prefer: rare words  common words 1

Change output attribute(s)

--- --- ---

You can select one or more output attributes. Please note that this option can be time-consuming.

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 を作成すると以下ようになります。

Sketch Engine   [British National Corpus \(BNC\) / Written\\_Medium\\_Book](#)  
 Sketch Engine homepage

Home  
 Concordance  
**Word list**  
 Word sketch  
 Thesaurus  
 Sketch diff  
 Trends  
 Corpus info  
 My jobs  
 User guide ↗

Save  
 Change options

**Word list**  
 Corpus: British National Corpus (BNC)  
 Subcorpus: Written\_Medium\_Book  
 Page   [Next >](#)

<u>lemma</u>	<u>Freq</u>
the	<a href="#">3,246,582</a>
be	<a href="#">2,120,850</a>
of	<a href="#">1,701,293</a>
and	<a href="#">1,386,210</a>
to	<a href="#">1,352,316</a>
a	<a href="#">1,121,858</a>
in	<a href="#">1,026,932</a>
have	<a href="#">660,482</a>
that	<a href="#">553,152</a>
it	<a href="#">521,370</a>
for	<a href="#">420,060</a>
he	<a href="#">409,936</a>
not	<a href="#">402,571</a>
I	<a href="#">398,018</a>
as	<a href="#">385,700</a>
with	<a href="#">343,420</a>
on	<a href="#">338,677</a>
you	<a href="#">290,158</a>
his	<a href="#">280,464</a>
she	<a href="#">265,410</a>
by	<a href="#">263,321</a>
at	<a href="#">262,123</a>
do	<a href="#">253,091</a>

(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。

Home		make (verb) Alternative PoS: <a href="#">noun</a> (799) <a href="#">adjective</a> (1)		British National Corpus (BNC) freq = <a href="#">209,867</a> (1,868.97 per million)	
Concordance	Word list	<b>modifiers of "make"</b>	<b>objects of "make"</b>	<b>subjects of "make"</b>	<b>"make" and/or ...</b>
Word sketch	Thesaurus	22,331 0.40	122,259 5.60	38,330 2.30	2,394 0.10
Sketch diff	Trends	sure + 1,578 10.77	decision + 2,940 9.51	decision + 403 8.24	break + 103 10.14
Corpus info	My jobs	make sure	sense + 2,530 9.31	decision making	make or break
User guide	Save	also + 1,410 8.33	make sense	people + 549 7.58	try + 159 9.56
Change options	Cluster	also made	use + 2,423 9.24	people make	to try and make
Sort by freq	Hide gramrels	only + 747 7.93	make use of	company + 306 7.48	design 55 9.20
More data	Less data	only make	mistake + 1,627 8.73	man + 356 7.34	designed and made
		just + 840 7.91	way + 1,906 8.63	government + 301 7.33	sell 56 9.01
		just make	point + 1,666 8.62	person + 179 6.95	making and selling
		already + 461 7.83	difference + 1,558 8.61	person making	go 92 8.30
		already made	contribution + 1,410 8.52	friend + 165 6.89	go and make a
		actually + 326 7.81	effort + 1,384 8.48	my hon. friend makes	use 36 8.28
		actually make	statement + 1,308 8.39	woman + 205 6.85	unmake 20 8.08
		ever + 304 7.79	make a statement	god + 149 6.79	to make or unmake
		ever made	attempt + 1,274 8.37	god made	buy 24 7.82
		then + 544 7.78	progress + 1,102 8.17	policy + 133 6.63	make or buy
		then made	change + 1,105 8.08	policy making	receive 19 7.73
		always + 436 7.61	order + 907 7.84	party + 147 6.59	do 40 7.59
		always made	profit + 867 7.81	party made	give 22 7.58
		never + 451 7.53	money + 936 7.80	member + 127 6.49	given or made
		never made	love + 816 7.73	minister + 114 6.40	bear 15 7.48
		not + 4,873 7.52	love + 816 7.73	minister made	born not made

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。

modifiers of "make"		objects of "make"		subjects of "make"		"make" and/or...		prepositional phrases	
sure +	1,578 10.99	decision +	2,940 9.51	decision +	403 8.24	break +	103 10.14	"make" by ...	4,267 3.40
also +	1,410 8.33	sense +	2,530 9.31	people +	549 7.58	try +	159 9.56	"make" in ...	3,613 0.80
only +	747 7.93	use +	2,423 9.24	company +	306 7.48	design	55 9.20	"make" of ...	2,739 0.30

particles after "make"		particles after "make" with object		pronominal objects of "make"		pronominal subjects of "make"		wh-words following "make"	
up +	3,785 3.20	up +	4,477 6.00	it +	30,720 8.30	they +	24,609 1.70	which +	691 0.90
off +	242 8.24	out +	3,254 10.57	me +	14,384 10.50	he +	3,498 8.13	whatever	111 8.96
out +	710 7.81	through	1,073 9.28	them +	3,174 9.88	it +	4,919 7.99	that	18 8.26
			36 7.79		3,336 9.73		5,578 7.97		60 8.11

infinitive objects of "make"		-ing objects of "make"		adjectives after "make" and noun		adjectives after "make"	
feel +	3,304 0.90	concern	1,106 0.6	clear +	18,678 47.00	sure +	11,352 3.10
ensure	74 8.58	regard	38 9.86	easy +	1,317 11.01	clear +	3,664 12.13
look	83 8.13	use	24 9.17	difficult +	1,293 10.93	available +	1,039 10.60

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

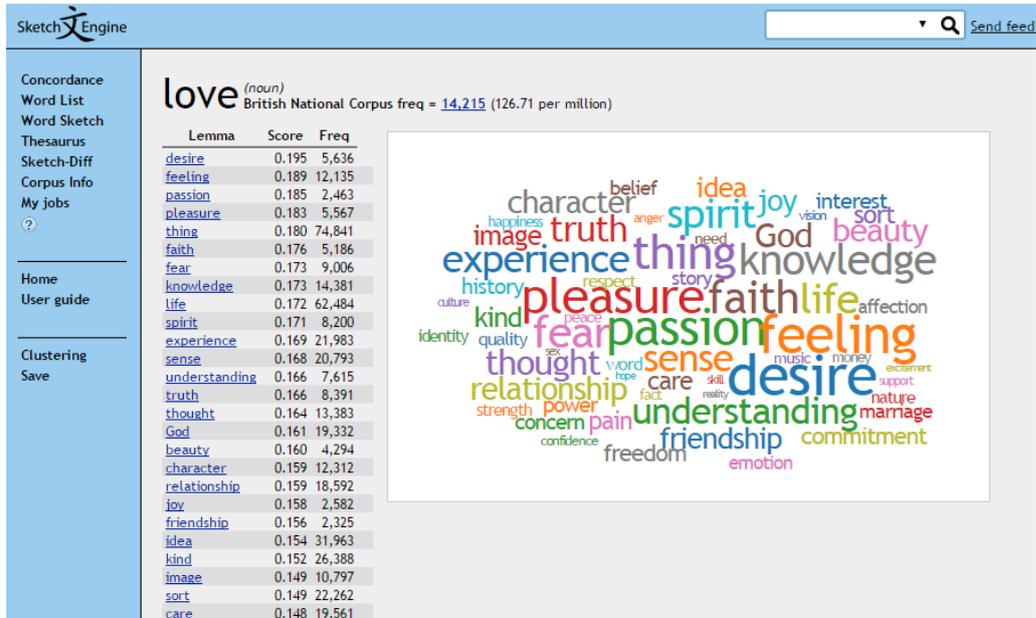


図 2（下に結果が続く）

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

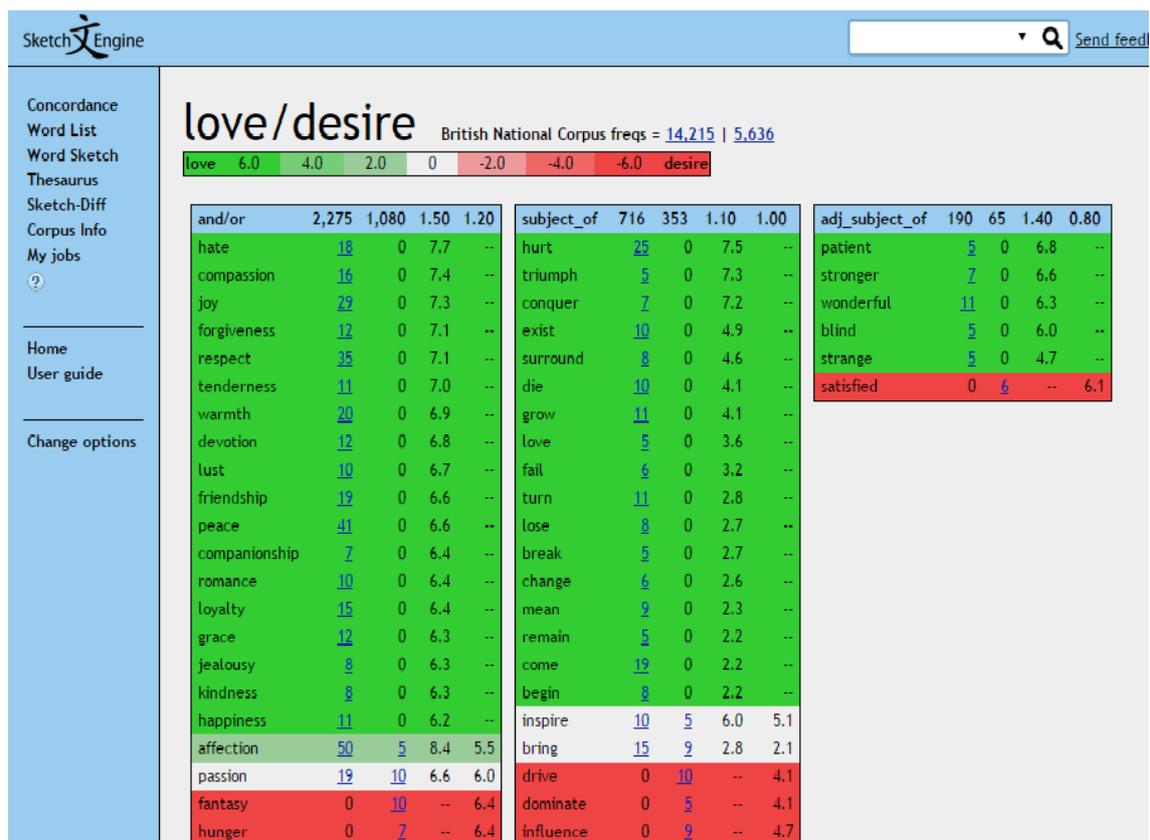


図3 (下に結果が続く)

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](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="〇〇" Title="テキスト名">

<text Types="〇〇">

(ii) フッター（終了タグ）

</text>

</doc>

</xml>

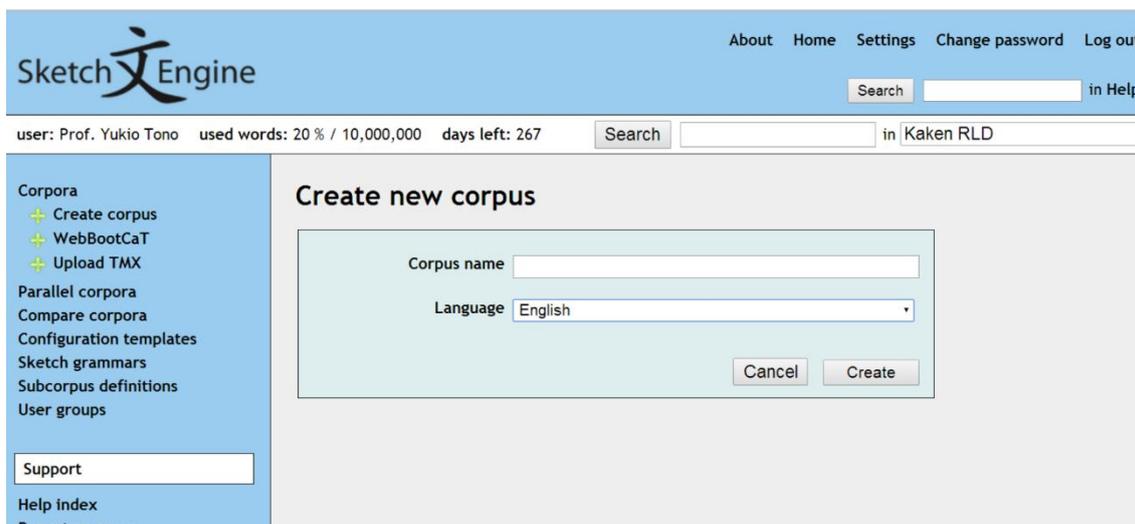
このようにテキスト内にコンピューターが理解できるような形で情報を付与すると、アップロードした後に自動的にサブコーパスに分けてくれたり、検索結果画面でテキストの細かい情報が表示できたりするようになります！

〈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

Corpora

- Create corpus
- WebBootCaT
- Upload TMX

Parallel corpora

- Compare corpora
- Configuration templates
- Sketch grammars
- Subcorpus definitions
- User groups

Corpus

- Corpus page
- Add new file
- Add web data (BootCaT)
- Compile corpus
- Search corpus
- Extract keywords & terms
- Configure corpus
- Change sketch grammar
- Set subcorpus definitions
- Expert mode
- Download corpus
- Access privileges
- View logs

### English: Add new file: Step 1

Supported file types: .doc, .docx, .htm, .html, .pdf, .ps, .tar.bz2, .tar.gz, .tgz, .tmx, .txt, .vert, .xml, .zip.

It is possible to upload multiple documents in an archive file. Supported archive types include .zip, .tar, .tar.gz, and .tar.bz2. By default, the contents of the archived documents are extracted into a single text file. If you wish to expand the archive into individual files instead, please select this option at the following page after uploading the archive. Expanding the archive is also necessary if it contains vertical files. Files with an unknown extension will be ignored. Vertical files included in the archive need to contain the following tab-separated values: word, tag, lempos (in this order).

You are responsible for the copyright and other intellectual property issues of the uploaded content.

Upload from disk  ファイルを選択 選択されていません **クリック**

Download from location  http://

Use file or directory on the server  -----

show files in subdirectories

FTP to [the.sketchengine.co.uk](http://the.sketchengine.co.uk) at port 10021 to upload files. Use the same user name and password as for logging into this web interface.

Paste text

→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

Corpora

- Create corpus
- WebBootCaT
- Upload TMX

Parallel corpora

- Compare corpora
- Configuration templates
- Sketch grammars
- Subcorpus definitions
- User groups

File

- Edit file
- Delete file
- View plain text
- View vertical
- Download original file
- Download plain text
- Download vertical

Support

- Help index
- Report an error

### English: Add new file:

#### 001\_A1\_International\_English\_for\_Speakers\_of\_Other\_Languages\_Book\_1\_P

File added successfully:  
001\_A1\_International\_English\_for\_Speakers\_of\_Other\_Languages\_Book\_1\_Preliminary\_Listening.xml.txt  
Detected character encoding: utf\_8

File type: Plain text

Character encoding: UTF-8 (all languages)

If some characters are not displayed correctly in the preview, changing this may help.

Cancel Update preview Finish **クリック**

#### File preview (plain text)

Showing bytes 1..1500 / 54928

```
<xml>
<file CEFR="A1" Title="International English for Speakers of Other Languages">
<text Skills="Reading">
How do we say the letters A to Z in English? Look at the letters. Do you know
how we say them? Do you know them when you hear them? Put a circle around the
```

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

Sketch Engine

About Home Settings Change password

Search

user: Prof. Yukio Tono used words: 34 % / 10,000,000 days left: 252

Search in ELT-CourseBook-02

**ELT-CourseBook-01**  
kaken\_rtd

← クリック

[Add new file](#) / 
 [Add data from web using WebBootCaT](#) / 
 [Compile corpus](#) / 
 [Search corpus](#)

#	Original file	Plain text	Vertical	Tokens	Owner
1	001_A1_Internati...istening.xml.txt	✓	✓	12,255	Prof. Yukio Tono
2	002_A2_Internati...2_Access.xml.txt	✓	✓	18,559	Prof. Yukio Tono
3	003_B1_Internati...istening.xml.txt	✓	✓	23,818	Prof. Yukio Tono
4	004_B2_Internati...or_Listening.txt	✓	✓	36,756	Prof. Yukio Tono
5	005_C1_Internati...5_Expert.xml.txt	✓	✓	47,526	Prof. Yukio Tono
6	006_C2_Internati...Mastery.xml.txt	✓	✓	34,859	Prof. Yukio Tono
7	007_B1_Close-Up.xml.txt	✓	✓	16,776	Prof. Yukio Tono
8	008_A1_English_Explorer_1.xml.txt	✓	✓	8,699	Prof. Yukio Tono
9	009_B1_English_Explorer_4.xml.txt	✓	✓	17,214	Prof. Yukio Tono
10	010_A1_Holiday_Explorer_1.xml.txt	✓	✓	3,660	Prof. Yukio Tono
11	011_B1_Just_Righ...mediate.xml.txt	✓	✓	10,587	Prof. Yukio Tono
12	012_B1_Just_Righ...o_Script.xml.txt	✓	✓	4,642	Prof. Yukio Tono

## 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)

Home

- + Create corpus
- + WebBootCaT
- + Upload TMX

Parallel corpora  
Compare corpora  
My jobs

Advanced features

Corpus templates  
Sketch grammars  
Subcorpus definitions  
GDEX configurations  
User groups  
Subscription overview

Support

User guide  
Feedback

## WebBootCaT: Create corpus ?

[Get seed words from Wikipedia](#)

Corpus name

Language

WebBootCaT is unavailable for languages which cannot be automatically tokenised.

Input type

Seed words  
 URLs

Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs.

Seed words

Random tuples will be selected from the seed words to query a search engine. Input 3 to 20 words or multiword expressions. Use space as separator. Enclose multiword expressions into quotes ("").

Compile corpus when finished

Automatically compile corpus when WebBootCaT processing is finished.

[Show advanced options](#)

Cancel Next >

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。

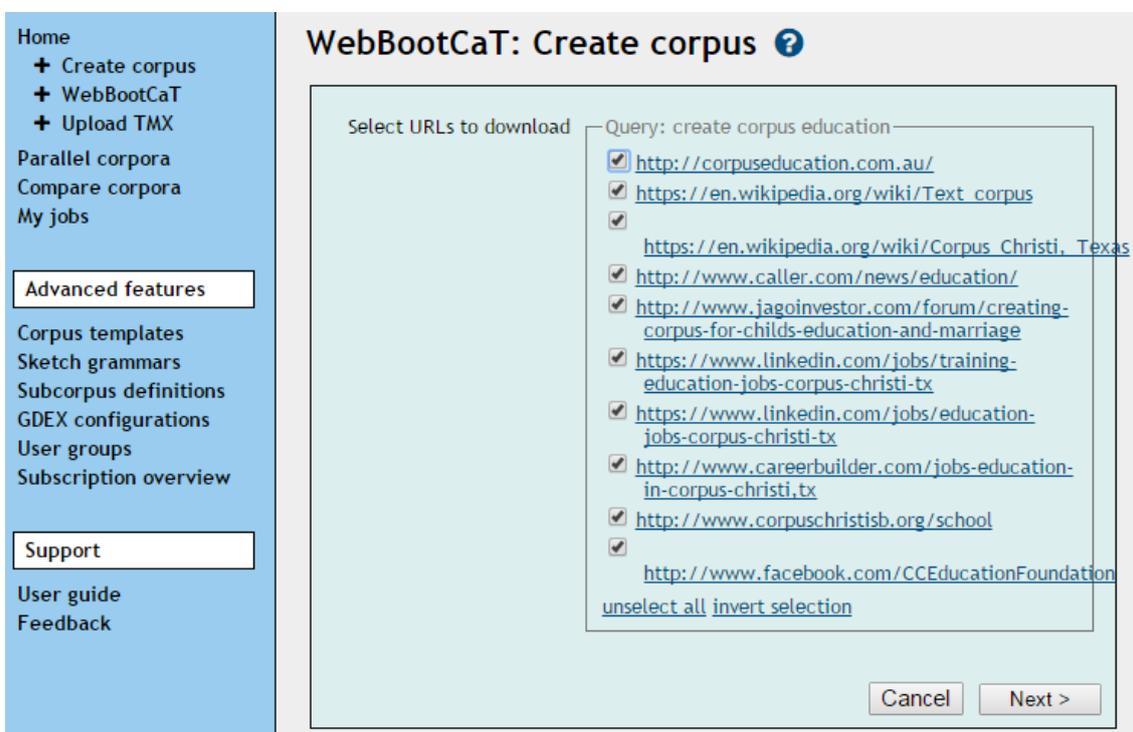


図 2 : URL の一覧

★ “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)。

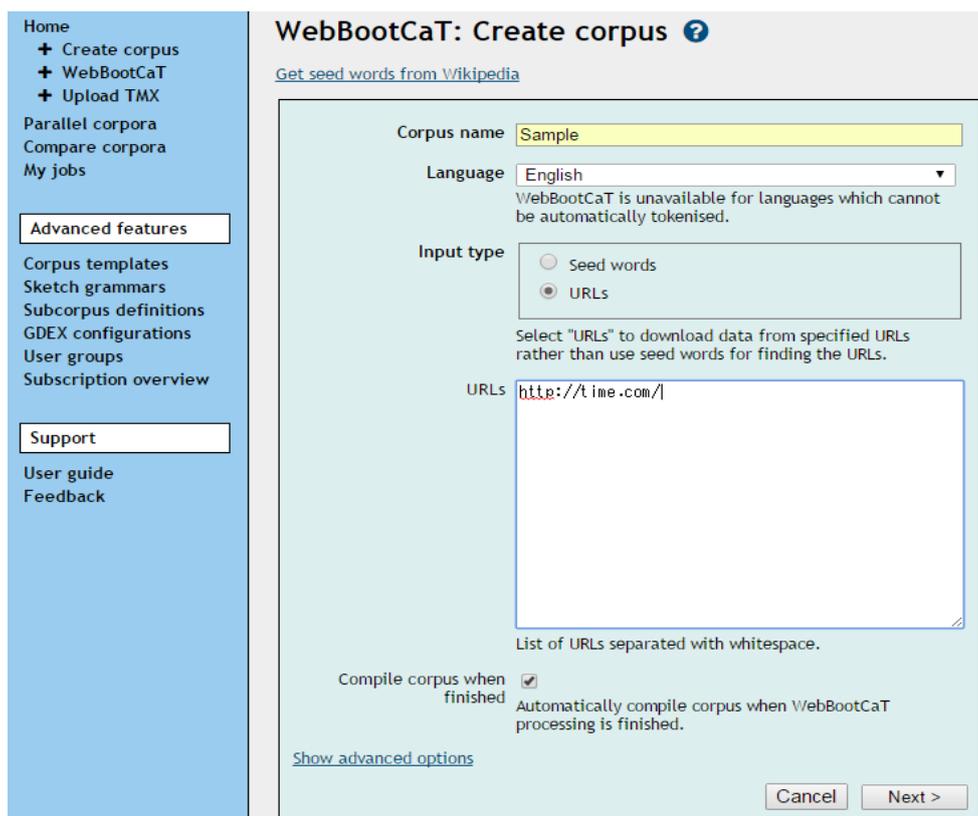


図 3 : URL の指定

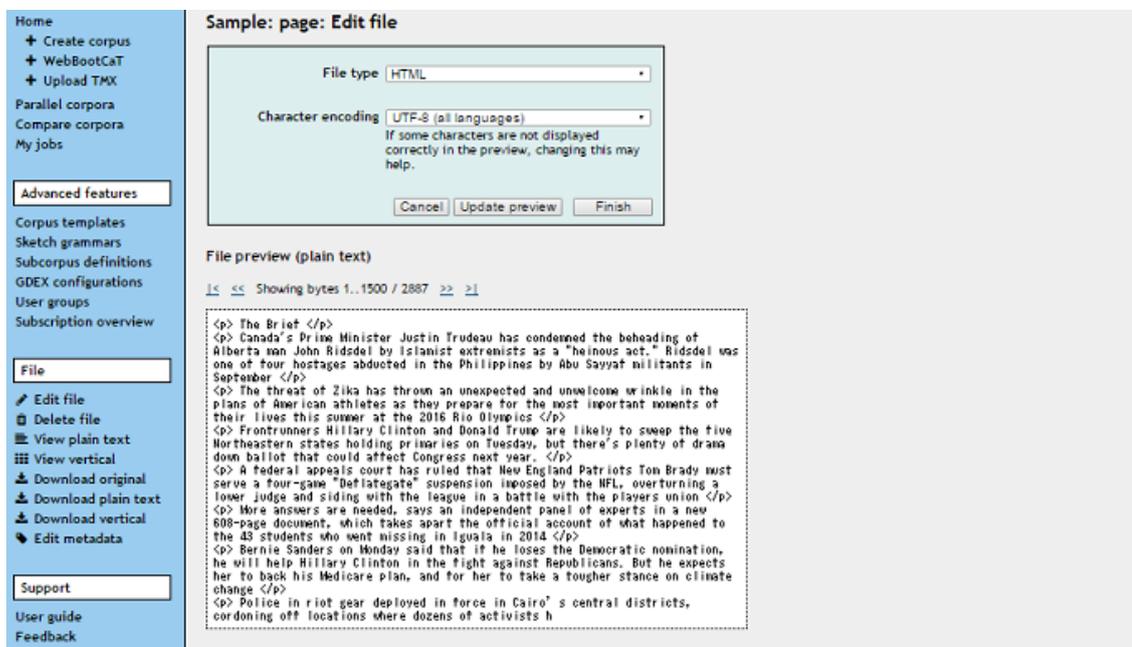


図 4 : 完成したコーパス