

1、アップロードしたいファイルに xml タグをつける

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 >

(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)

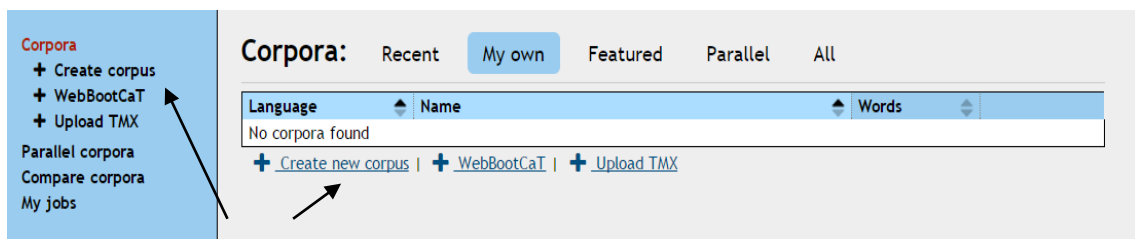
</text>

</doc>

</xml>

2、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



いずれかをクリック

(ii) コーパスの名前と扱う言語の設定

The screenshot shows the Sketch Engine interface. At the top, there is a navigation bar with links for 'About', 'Home', 'Settings', 'Change password', and 'Log out'. Below this, a search bar is visible. The main content area is titled 'Create new corpus'. It features a form with two input fields: 'Corpus name' and 'Language', which is currently set to 'English'. There are 'Cancel' and 'Create' buttons at the bottom of the form. On the left side, there is a sidebar menu with various options like 'Corpora', 'Parallel corpora', and 'Support'.

3、ファイルのアップロード

(i)

The screenshot shows the 'English: Add new file: Step 1' page. It provides information about supported file types: .doc, .docx, .htm, .html, .pdf, .ps, .tar.bz2, .tar.gz, .tgz, .tmx, .txt, .vert, .xml, .zip. It also explains that multiple documents can be uploaded in an archive file and lists supported archive types: .zip, .tar, .tar.gz, and .tar.bz2. A warning states that files with unknown extensions will be ignored and that vertical files need to contain tab-separated values: word, tag, lempos. A disclaimer notes that the user is responsible for copyright and other intellectual property issues. The main form has three radio buttons: 'Upload from disk' (selected), 'Download from location', and 'Use file or directory on the server'. The 'Upload from disk' option has a button labeled 'ファイルを選択' (Select file) with a tooltip that says '選択されていません' (Not selected). An arrow labeled 'クリック' (Click) points to this button. Below the radio buttons, there is a text input field for 'http://', a dropdown menu, and a checkbox for 'show files in subdirectories'. At the bottom, there is a 'Paste text' option.

→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

(ii)



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.

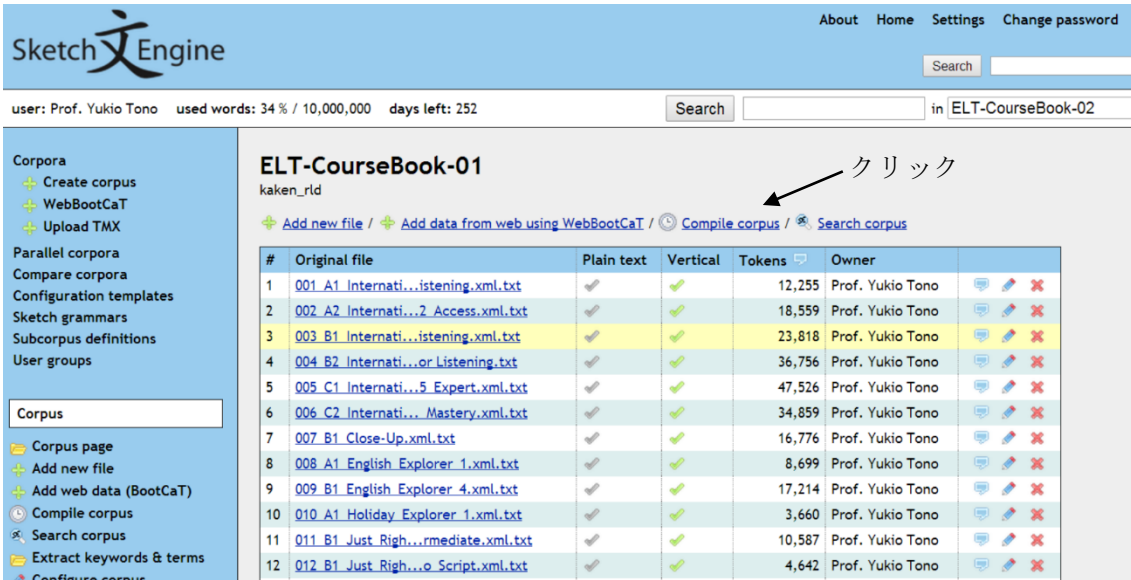
Buttons: Cancel, Update preview, Finish

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
```

クリック

(iii) 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



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

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

ELT-CourseBook-01  
kaken\_rld

Buttons: 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

クリック

## (1) 検索の基礎

- 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) 検索の応用

### • Context

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。

- 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。

## (2) コンコーダンスの活用

### • Sort

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  of 494  [Next](#) | [Last](#) Concordance is sorted. Jump to:

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

↑ファイル情報

↑ L1 ↑ node ↑ R1

- Sample…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

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

**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…共起語を調べる

**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…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。

## 2 Concordance -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・・・

Sketch Engine

Query **turn, V.\*, N.\*, PRP** 620 (5.50 per million)

Page 1 of 31 Go Next Last

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 er and to do so l
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
HTP	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

Page 1 of 31 Go Next Last

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)

参考資料

投野由紀夫 (2010)「第 35 回英語コーパス学会ワークショップ Web コーパス検索  
ツール Sketch Engine の基本操作と活用」



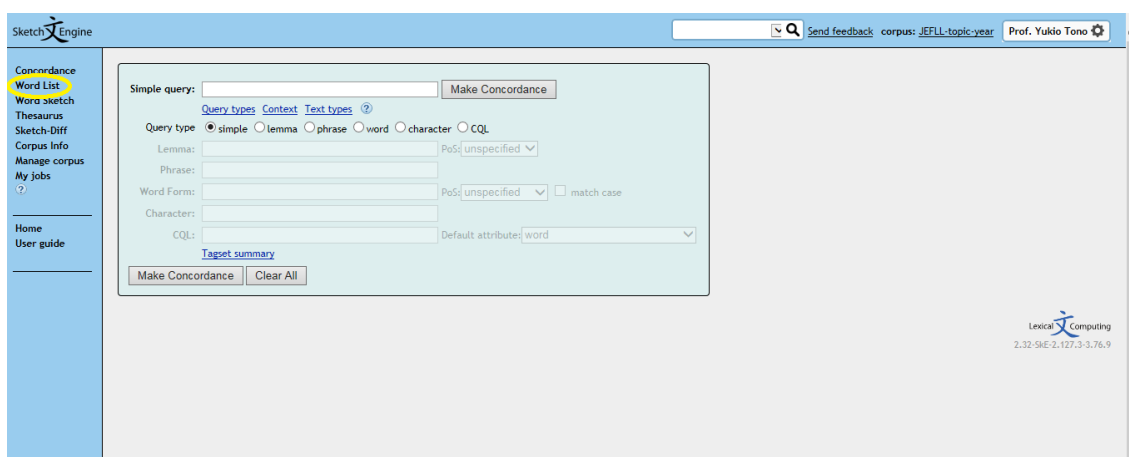
投野ゼミ ゼミ合宿 2015

Sketch Engine 使用マニュアル

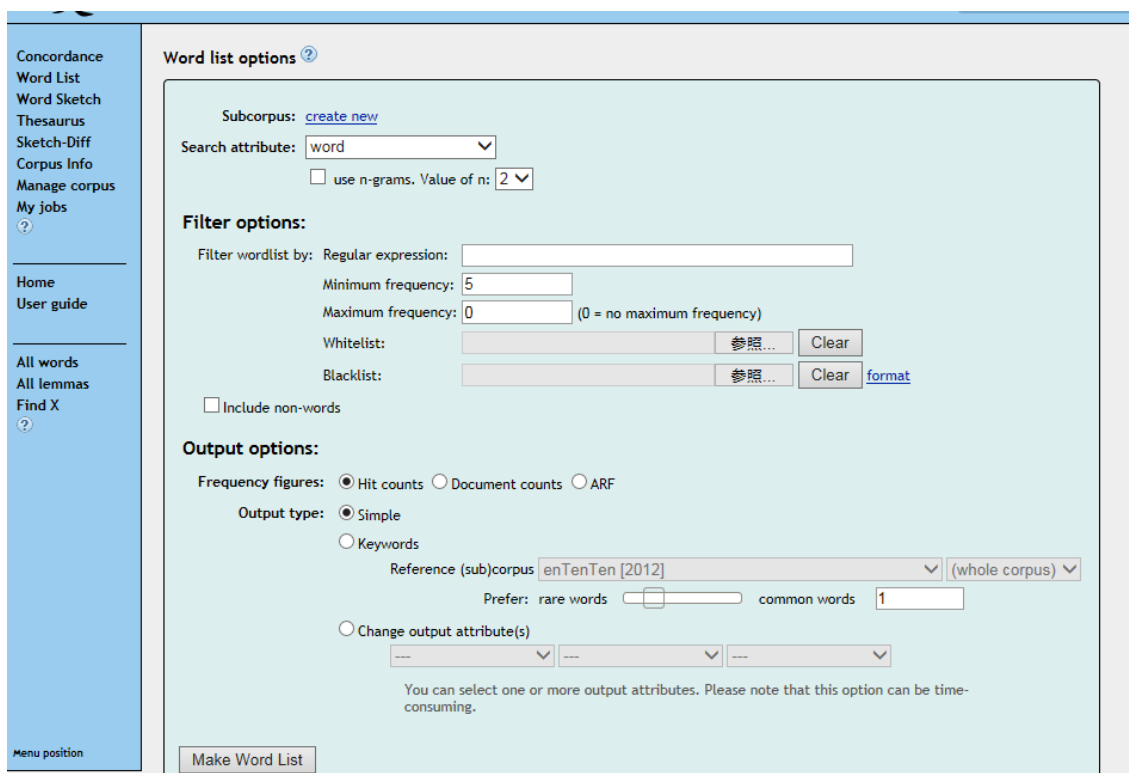
担当:中島詩菜

### 3. Word List

コーパス（サブコーパス）内から語彙表を抽出できる。



クリックすると↓

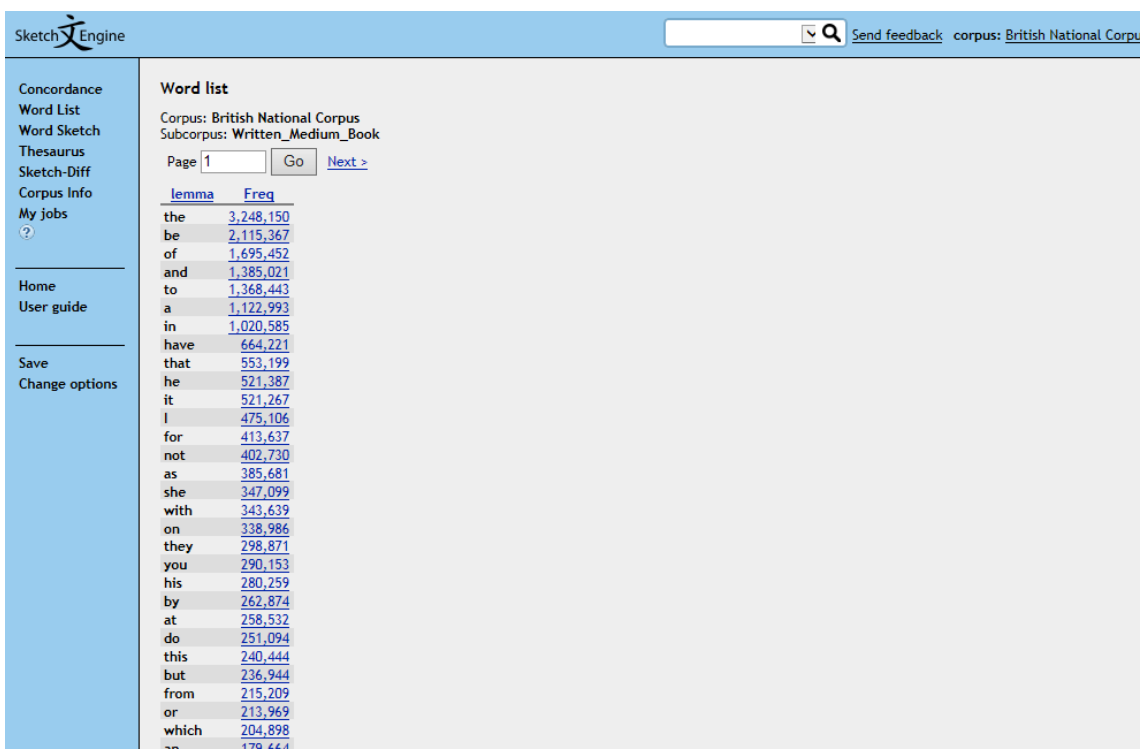


この画面が出てきます。

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 options 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

Send feedback corpus: British National Corpus

Concordance  
Word List  
Word Sketch  
Thesaurus  
Sketch-Diff  
Corpus Info  
My jobs  
Home  
User guide  
Save  
Change options

**Word list**

Corpus: British National Corpus  
Subcorpus: Written\_Medium\_Book

Page 1 Go Next >

lemma	Freq
the	3,248,150
be	2,115,367
of	1,695,452
and	1,385,021
to	1,368,443
a	1,122,993
in	1,020,585
have	664,221
that	553,199
he	521,387
it	521,267
I	475,106
for	413,637
not	402,730
as	385,681
she	347,099
with	343,639
on	338,986
they	298,871
you	290,153
his	280,259
by	262,874
at	258,532
do	251,094
this	240,444
but	236,944
from	215,209
or	213,969
which	204,898
an	179,664

(頻度が高い順に lemma を並べた語彙表)

### 【Filter options】

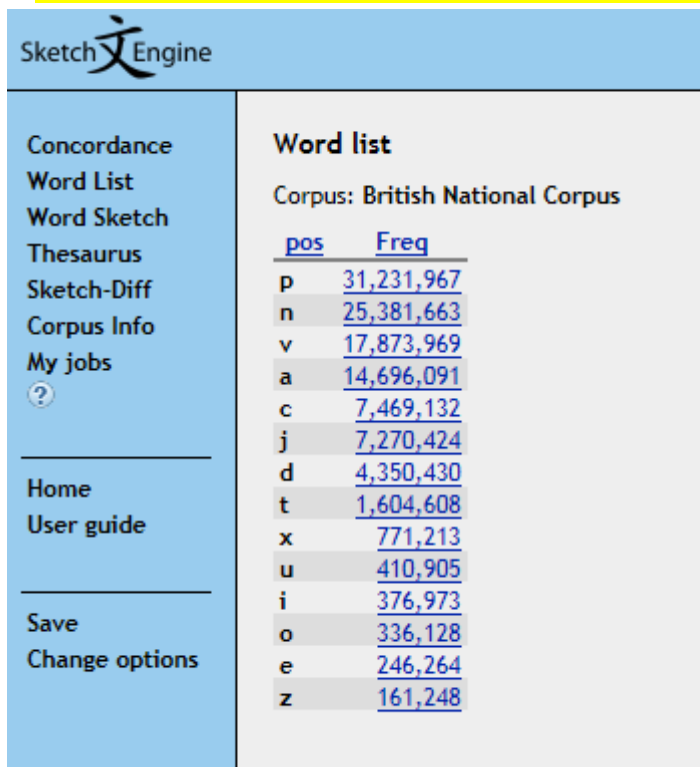
- Regular expression: 正規表現で検索できます。".\*"がワイルドカード（何が何文字入っても OK）を表すので、"th.\*"で検索すると the, that, this 等の語彙表が作成されます。  
(その他、+、?、!などがあります)
- Minimum frequency: 最小頻度を指定できます。
- Maximum frequency: 最大頻度を指定できます。
- Whitelist: 語彙表に含めたい特定の単語リストがある場合、アップロードできます。
- Blacklist: 語彙表に含めたくない特定の単語リストがある場合、アップロードできます。
- Include 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) で複数回現れる時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):

### 【検索例】

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



Sketch Engine

Concordance  
Word List  
Word Sketch  
Thesaurus  
Sketch-Diff  
Corpus Info  
My jobs  
?

---

Home  
User guide

---

Save  
Change options

**Word list**  
Corpus: British National Corpus

<u>pos</u>	<u>Freq</u>
p	31,231,967
n	25,381,663
v	17,873,969
a	14,696,091
c	7,469,132
j	7,270,424
d	4,350,430
t	1,604,608
x	771,213
u	410,905
i	376,973
o	336,128
e	246,264
z	161,248

② BNCの中のサブコーパス Written\_Domain\_Imaginative で search attribute を lemma、regular expression を wh.\*、minimum frequency を 1、maximum frequency を 0

Sketch文Engine

Concordance  
Word List  
Word Sketch  
Thesaurus  
Sketch-Diff  
Corpus Info  
My jobs  
?

---

Home  
User guide

---

Save  
Change options

### Word list

Corpus: British National Corpus  
Subcorpus: Written\_Domain\_Imaginative

Page   [Next >](#)

<u>lemma</u>	<u>Freq</u>
what	62,804
when	44,413
who	32,451
which	27,625
where	20,374
why	16,900
while	10,599
white	6,539
whole	4,787
whether	3,260
whisper	3,066
whatever	2,716
whose	2,228
whom	1,898
wheel	1,497
whenever	714
whisky	665
whistle	579
whilst	554
whip	533
whoever	520
wherever	410
whirl	342
whereas	290
whore	266
whine	250
whimper	202
wholly	195
whatsoever	174
whereabouts	161

③ BNC 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

Sketch Engine

Concordance  
Word List  
Word Sketch  
Thesaurus  
Sketch-Diff  
Corpus Info  
My jobs  
?  
Home  
User guide  
Save  
Change options

### Word list

Corpus: British National Corpus  
Subcorpus: Written\_Domain\_Informative

Page   [Next >](#)

<u>word</u>	<u>Doc freq</u>
being	2,521
including	2,330
during	2,324
making	2,307
following	2,247
having	2,238
taking	2,186
working	2,182
using	2,120
going	2,098
bring	2,018
something	2,010
giving	2,002
doing	1,943
according	1,939
leading	1,915
nothing	1,914
looking	1,909
running	1,863
building	1,860
coming	1,850
anything	1,847
beginning	1,838
meeting	1,787
thing	1,745
providing	1,743
leaving	1,738
trying	1,733
becoming	1,719
growing	1,714

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

Sketch Engine

Concordance  
Word List  
Word Sketch  
Thesaurus  
Sketch-Diff  
Corpus Info  
My jobs  
?

---

Home  
User guide

---

Save  
Change options

### Word list

Corpus: British National Corpus

Page   [Next >](#)

<a href="#">word (4-grams)</a>	<a href="#">Freq</a>
I do n't know	11,901
the end of the	10,374
at the end of	7,843
I do n't think	6,984
at the same time	4,811
the rest of the	4,712
for the first time	4,711
per cent of the	4,522
as a result of	4,468
one of the most	3,286
is one of the	3,267
do n't want to	3,266
in the case of	3,245
I do n't want	3,239
the Secretary of State	3,221
to be able to	3,167
On the other hand	2,836
in the form of	2,756
on the basis of	2,743
the top of the	2,673
in the middle of	2,641
do n't know what	2,542
by the end of	2,512
as well as the	2,507
on the other hand	2,460
the way in which	2,425
a member of the	2,415
was one of the	2,328
at the time of	2,288
the middle of the	2,219
a great deal of	2,205

⑤ BNC 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

Sketch Engine   [Send feedback](#) corpus: [British National Corpus](#)

Concordance  
 Word List  
 Word Sketch  
 Thesaurus  
 Sketch-Diff  
 Corpus Info  
 My jobs  
 ?

Home  
 User guide

Save  
 Change options

### Word list

Corpus: [British National Corpus](#)  
 Subcorpus: [Written\\_Medium\\_Book](#)

Reference corpus: [British National Corpus](#)  
 Reference subcorpus: [Written\\_Medium\\_To-be-spoken](#)  
[Switch focus and reference \(sub\)corpus](#)

Page   [Next >](#)

word	<i>British National Corpus : Written_Medium_Book</i>		<i>British National Corpus : Written_Medium_To-be-spoken</i>		Score
	Freq	Freq/mill <sup>?</sup>	Freq	Freq/mill	
p.	<a href="#">4,628</a>	79.9	0	0.0	80.9
thus	<a href="#">7,757</a>	133.9	<a href="#">1</a>	0.7	79.9
defined	<a href="#">3,812</a>	65.8	0	0.0	66.8
Figure	<a href="#">3,583</a>	61.8	0	0.0	62.8
et	<a href="#">3,348</a>	57.8	0	0.0	58.8
i.e.	<a href="#">3,301</a>	57.0	0	0.0	58.0
Fig.	<a href="#">3,168</a>	54.7	0	0.0	55.7
Thus	<a href="#">7,395</a>	127.6	<a href="#">2</a>	1.4	54.1
glanced	<a href="#">2,597</a>	44.8	0	0.0	45.8
Moreover	<a href="#">2,548</a>	44.0	0	0.0	45.0
realized	<a href="#">2,465</a>	42.5	0	0.0	43.5
q.v.	<a href="#">2,325</a>	40.1	0	0.0	41.1
whispered	<a href="#">2,262</a>	39.0	0	0.0	40.0
organization	<a href="#">3,816</a>	65.8	<a href="#">1</a>	0.7	39.6
assumption	<a href="#">2,228</a>	38.4	0	0.0	39.4
Chapter	<a href="#">5,372</a>	92.7	<a href="#">2</a>	1.4	39.4
latter	<a href="#">5,313</a>	91.7	<a href="#">2</a>	1.4	39.0
v.	<a href="#">2,133</a>	36.8	0	0.0	37.8
Section	<a href="#">2,125</a>	36.7	0	0.0	37.7
categories	<a href="#">2,123</a>	36.6	0	0.0	37.6
smiled	<a href="#">6,603</a>	113.9	<a href="#">3</a>	2.1	37.5
plaintiff	<a href="#">2,098</a>	36.2	0	0.0	37.2
patterns	<a href="#">3,574</a>	61.7	<a href="#">1</a>	0.7	37.1
e.g.	<a href="#">3,518</a>	60.7	<a href="#">1</a>	0.7	36.5
Similarly	<a href="#">2,038</a>	35.2	0	0.0	36.2

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

※この三つだと微妙、最後 word にしたり

Sketch Engine

Concordance  
Word List  
Word Sketch  
Thesaurus  
Sketch-Diff  
Corpus Info  
My jobs  
?

Home  
User guide

Save  
Change options

### Frequency list

Frequency limit:

Page   [Next >](#)

	lemma	pos	lempos	Frequency
P   N	be	v	be-v	82,625
P   N	go	v	go-v	62,850
P   N	something	p	something-p	50,069
P   N	during	p	during-p	43,457
P   N	have	v	have-v	34,176
P   N	thing	n	thing-n	33,850
P   N	nothing	p	nothing-p	32,233
P   N	anything	p	anything-p	27,269
P   N	do	v	do-v	27,010
P   N	make	v	make-v	25,594
P   N	look	v	look-v	25,102
P   N	including	p	including-p	22,894
P   N	use	v	use-v	22,513
P   N	take	v	take-v	21,067
P   N	morning	n	morning-n	19,975
P   N	bring	v	bring-v	19,557
P   N	training	n	training-n	19,100
P   N	work	v	work-v	18,868
P   N	get	v	get-v	18,711
P   N	everything	p	everything-p	17,716
P   N	come	v	come-v	17,682
P   N	try	v	try-v	17,654
P   N	say	v	say-v	17,517
P   N	meeting	n	meeting-n	15,961
P   N	building	n	building-n	15,914
P   N	following	j	following-j	13,200
P   N	evening	n	evening-n	13,172
D   N	talk	v	talk-v	12,220



### Frequency list

Frequency limit:

Page   [Next >](#)

	<a href="#">lemma</a>	<a href="#">pos</a>	<a href="#">lempos</a>	<a href="#">Frequency</a>	
P   N	be	v	be-v	82,625	
P   N	go	v	go-v	62,850	
P   N	something	p	something-p	50,069	
P   N	during	p	during-p	43,457	
P   N	have	v	have-v	34,176	
P   N	thing	n	thing-n	33,850	
P   N	nothing	p	nothing-p	32,233	
P   N	anything	p	anything-p	27,269	
P   N	do	v	do-v	27,010	
P   N	make	v	make-v	25,594	
P   N	look	v	look-v	25,102	
P   N	including	p	including-p	22,894	
P   N	use	v	use-v	22,513	
P   N	take	v	take-v	21,067	
P   N	morning	n	morning-n	19,975	
P   N	bring	v	bring-v	19,557	
P   N	training	n	training-n	19,100	
P   N	work	v	work-v	18,868	
P   N	get	v	get-v	18,711	
P   N	everything	p	everything-p	17,716	
P   N	come	v	come-v	17,682	
P   N	try	v	try-v	17,654	
P   N	say	v	say-v	17,517	
P   N	meeting	n	meeting-n	15,961	
P   N	building	n	building-n	15,914	
P   N	following	j	following-j	13,200	
P   N	evening	n	evening-n	13,172	
P   N	talk	v	talk-v	12,220	

# SketchEngine 演習 4. Word Sketch

ゼミ合宿 2 日目

外国語学部 フィリピン語専攻 4 年

安宅優希

## ◎ 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: 黄色の円): 左から

1. 文法関係

2. 用例全体における頻度

3. 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言える。

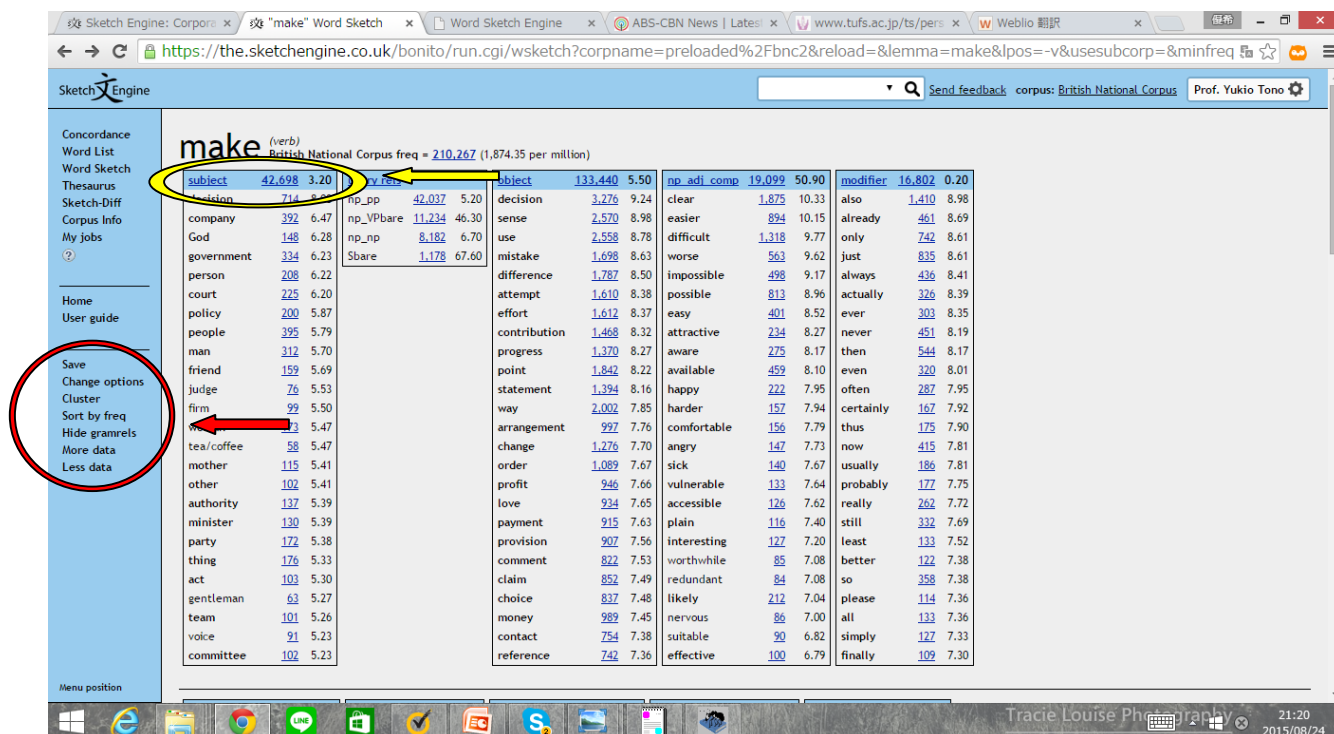


図 1: make の検索結果

4. その他のオプション(図 1: 赤の円): 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 British National Corpus を選択。
2. Lemma に”make”、Part of speech で”verb”を選択。
3. 検索結果は図 1 参照。
4. “np adj comp”のスコア値(50.90)が最も高い(図 2 参照)  
→make は”make+O+C”の構文が特徴的と言える。

The screenshot shows the Sketch Engine interface for the word 'make'. The URL is <https://the.sketchengine.co.uk/bonito/run.cgi/wsketch?corpname=preloaded%2Fbnc2&lemma=make&lpos=&usesubcorp=;;minfreq=6;minscore=1>. The main content area displays the word 'make' with its alternative part of speech (noun) and frequency (210,267). Below this, there are several tables of grammatical categories and their scores. The 'np adj comp' category is circled in red, indicating its high score of 50.90. Other categories include 'subject', 'decision', 'unary reIs', 'np\_VPbare', 'Sbare', 'object', 'decision', 'clear', 'part\_intrans', 'part\_in-p', 'part\_up-a\_obj', 'part\_of-p', 'part\_for-p', 'part\_to-p', 'and/or', 'part\_from-p', 'part\_on-p', 'part\_with-p', 'part\_at-p', 'part\_out-a\_obj', 'part\_under-p', 'part\_into-p', 'part\_against-p', 'part\_without-p', and 'part\_as-p'.

Category	Score 1	Score 2
subject	42,698	3.20
decision	714	8.08
unary reIs	11,234	46.30
np_VPbare	1,178	67.60
Sbare	1,178	67.60
object	133,440	5.00
decision	3,276	10.33
clear	1,875	10.33
part_intrans	3,640	1.50
part_in-p	3,231	0.50
part_up-a_obj	2,706	9.90
part_of-p	2,533	0.20
part_for-p	2,384	0.80
part_to-p	2,059	0.60
and/or	1,802	0.10
part_from-p	1,354	0.90
part_on-p	1,236	0.50
part_with-p	867	0.40
part_at-p	759	0.50
part_out-a_obj	377	2.00
part_under-p	361	1.90
part_into-p	292	0.40
part_against-p	165	0.80
part_without-p	133	1.50
part_as-p	114	0.20

図 2: make の検索結果

\*一覧を見やすくするために“Less Data”で表示データ数を減らしている。

## 5 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）

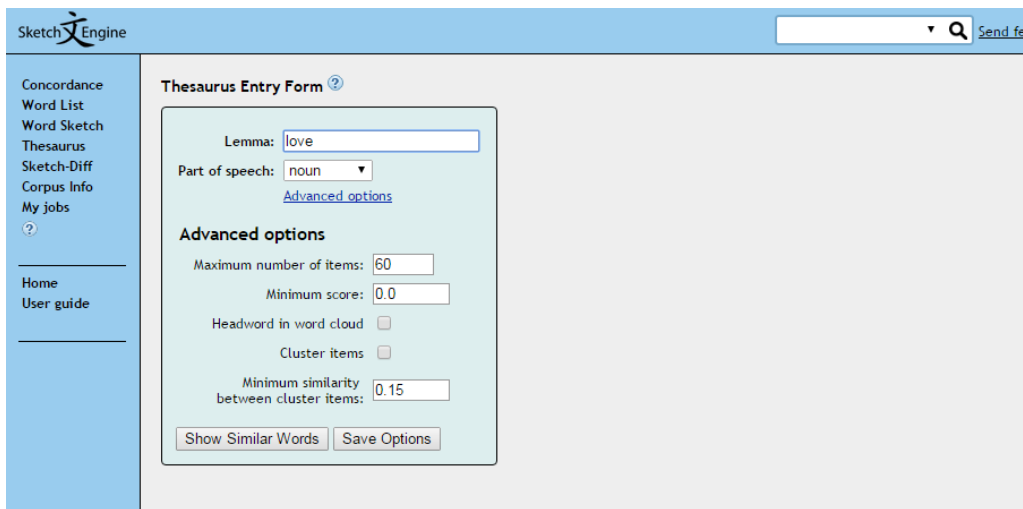


図 1

検索結果

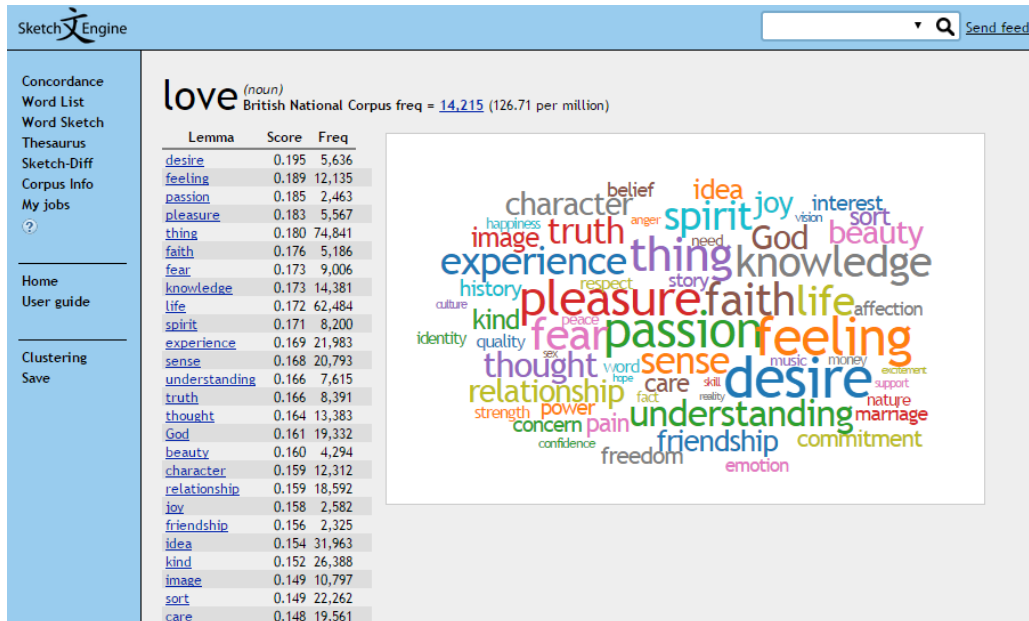


図 2（下に結果が続く）

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・・・)

#### 参考資料

投野由紀夫 (2010)「第 35 回英語コーパス学会ワークショップ Web コーパス検索  
ツール Sketch Engine の基本操作と活用」

# SketchEngine 演習 7. WebBootCaT

ゼミ合宿 2 日目  
外国語学部 フィリピン語専攻 4 年  
安宅優希

## ◎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(\*後半で詳しく解説)。

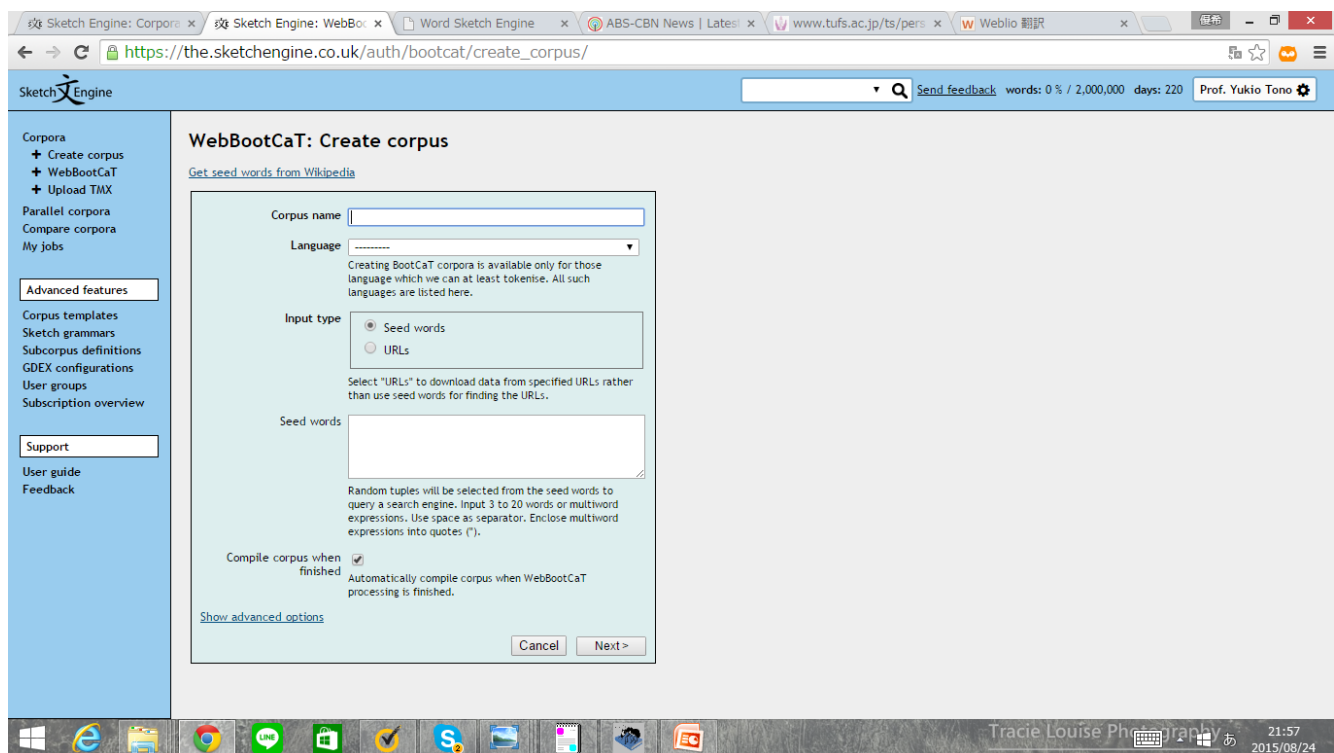


図 1: WebBootCaT の基本画面

## ◎WebBootCaT の具体的な使用例(Seed words/URLs)

★” Seed words”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。

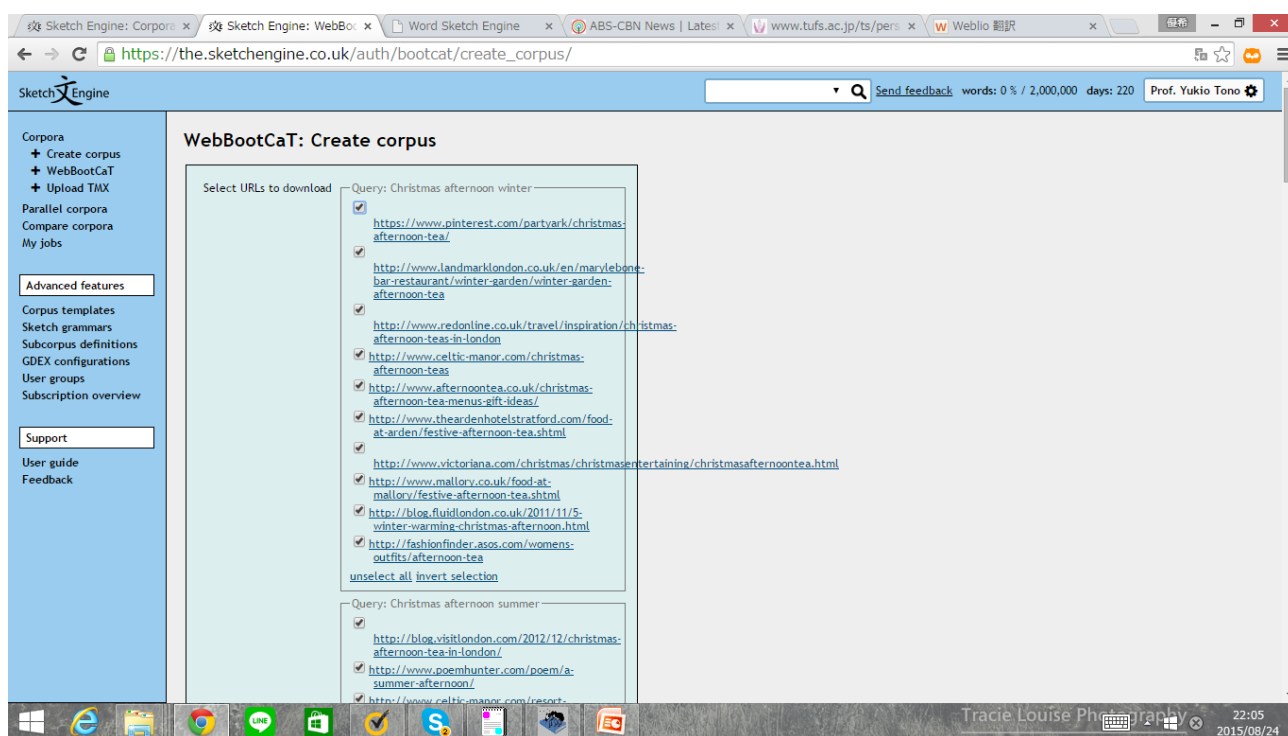


図 2: URL の一覧

★”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)。



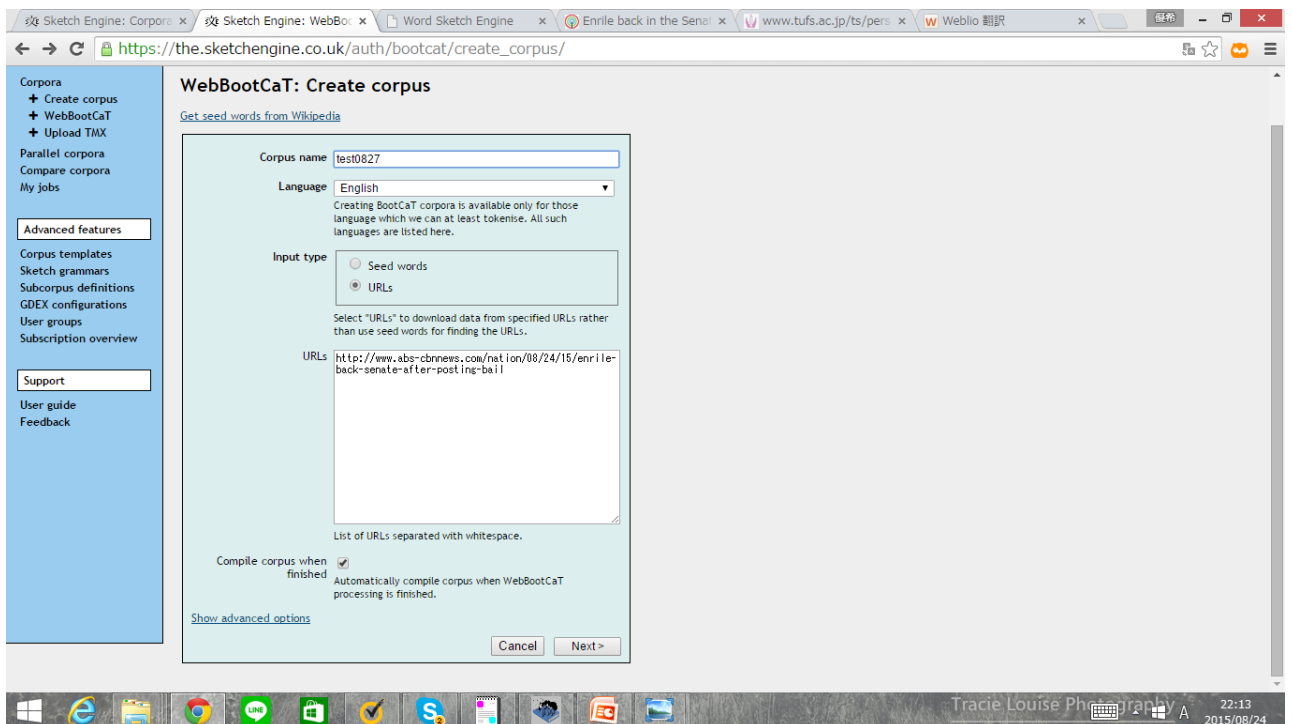


図 3: URL の指定

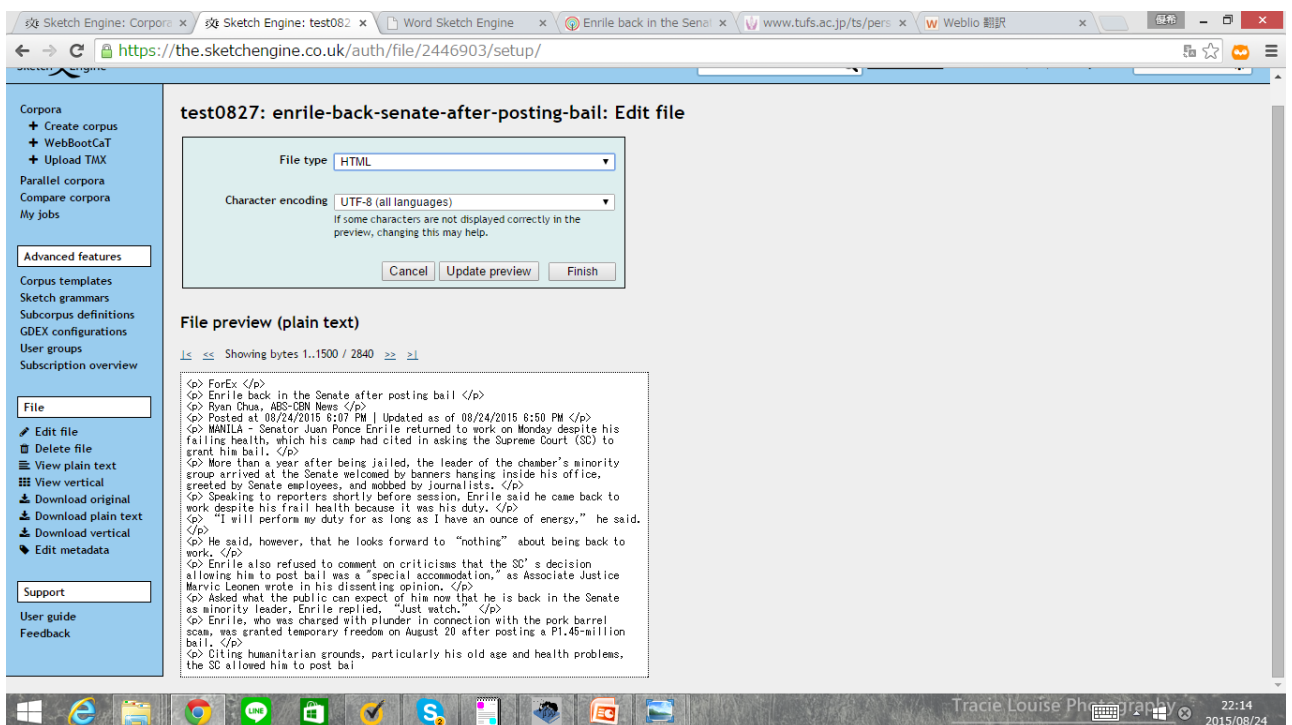


図 4: 完成したコーパス