<u>Sketch Engine</u> 操作マニュアル

東京外国語大学 投野研究室 (2018 年度改訂版)

| 1. | CON | NCORDANCE 画面を使いこなす | 3 |
|----|------------|-------------------------|----------|
| | 1.1. | 検索の基礎 | 3 |
| | 1.2. | COL 検索 | 5 |
| | 1.2.1 | 表層形検索 | 5 |
| | 1.2.2 | レマ検索 | 5 |
| | 1.2.3 | 品詞指定の仕方 | 5 |
| | 1.2.4 | より複雑な検索 | 6 |
| | 1.2.5 | 5. 文境界を指定した検索 | 7 |
| | 1.2.6 | 5. CQL attribute の指定 | 7 |
| | 1.3. | CQL BUILDER | 7 |
| 2. | さま | ざまなオプション機能 | 8 |
| | 2.1 | WORD | 0 |
| | 2.1. | WORD LIST | 5 |
| | 2.2. | THESAURUS / SKETCH DIFF | 5 |
| 3. | 3- | パスの作成10 | 3 |
| | 3.1. | ファイルのアップロード1 | 3 |
| | 3.2. | WEBBOOTCAT | 6 |
| 4. | 練習 | 問題1 | 8 |
| | 4.1 | 其大龄壶1 11 | Q |
| | 4.1. | 至平阪示1 | g |
| | ч.2. 43 | 空午仮示 2 COI | n |
| | 4.4 | COL 检索 2 2 | 1 |
| | 4.5. | 頻度表 | 2 |

<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: | | Make Concordance |
|---------------|--|--|
| | Query types Context Text types 📀 | |
| Query type | \odot simple \bigcirc lemma \bigcirc phrase \bigcirc word \bigcirc charact | ter 💛 CQL |
| Lemma: | | PoS: unspecified • |
| Phrase: | | |
| Word form: | | PoS: unspecified 🔻 🔲 match case |
| Character: | | |
| CQL: | | Default attribute: Ic = word (lowercase) |
| | Tagset summary | |
| Make Concor | dance Clear All | |
| | | |

- <u>Simple query</u>...単語、フレーズ検索(各単語の活用形も含む=lemma 検索)
- <u>Lemma</u>…単語のみ。活用形も含む。品詞指定可。
- <u>Phrase</u>…表層形のみのフレーズ検索(=活用形は含まれない)。
- <u>Word</u>...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)。

| | 単語 | フレーズ | 活用形 | 品詞指定 | 大文字・小文字指定 |
|--------|----|------|-----|------|-----------|
| Simple | 0 | 0 | 0 | | |
| Lemma | 0 | | 0 | 0 | |
| Phrase | | 0 | | | |
| Word | 0 | | | 0 | 0 |

(2) Context

| Simple query: | Make Concordance |
|---------------|---|
| | Query types Context Text types @ |
| Query type | 💿 simple 💿 lemma 💿 phrase 💿 word 💿 character 💿 CQL |
| Lemma: | PoS: unspecified v |
| Phrase: | |
| Word form: | PoS: unspecified 🔻 🗖 match case |
| Character: | |
| CQL: | Default attribute: Ic = word (lowercase) |
| | Tagset summary |
| Context | |
| Lemma filt | er PoS filter |
| Window | : both v 5 v tokens. Window: both v 5 v tokens. |
| Lemma(s) | : all v of these items. PoS: adjective all v of these items. |

- Lemma Filter...前後○語以内の指定の単語(lemma)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] にすると指定した単語/品詞の全てが含まれたもの、[any]に すると指定したもののうち少なくとも 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
- ・<u>Sort</u>(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 | white 7,875 > 301 Left 7,875 > 301 Dicade.id/ 0>0 7,875 > 31011e 7,875 > 301 Left 7,875 (dd |
|-------|---|
| 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'. 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. 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 |
| | s them |

↑ファイル情報

 $L1 \uparrow node \uparrow R1$

- ・<u>Sample</u>(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。
- ・<u>Filter</u>(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。

・<u>Frequency</u>(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。

・<u>Collocation</u>(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。

・<u>Visualize</u>(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

| Concordance Word List Word Sketch Thesaurus Sketch-Diff | Simple query: Make Concordance Query types Context Text types ? Query type Simple Iemma phrase word Character © CQL |
|---|---|
| Corpus Info My jobs ? | Lemma: PoS: unspecified ▼ Phrase: Word Form: PoS: unspecified ▼ match case |
| Home User guide | Character: CQL: Default attribute: word Tagget summary Make Concordance Clear All |
| | |

CQL の一般式: [attribute="value"]

- Attribute: word, lemma, tag, lempos がはいる
- [word="value"]
- [lemma="value"]
- [tag="value"]
- [lempos="value"]
- Value: 検索したい正規表現を含む文字列が入る。文字列は "" (ダブルクオーテション マーク)で挟む。

1.2.1. 表層形検索 <u>turn(表層形)を検索する場合の式</u> [word="turn"] "[tT]hank"

1.2.2. レマ検索 <u>turn (レマ) を検索する場合の式</u> [lemma="turn"]

1.2.3. 品詞指定の仕方 <u>turn の名詞を検索する場合の式</u> [lemma="turn"& tag="N.*"] または [lempos="turn-n"]

※ .(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

※ ※タグの記号に関しては Tagset summary を参照

Common tags:

| POS | tag | lempos |
|------|-----|--------|
| noun | N.* | -n |
| verb | V.* | -V |

| adjective | J.* | -j | |
|-------------|------|----|--|
| adverb | RB.? | -a | |
| conjunction | CC | -c | |
| preposition | IN | -р | |
| pronoun | PP.? | -d | |

1.2.4. より複雑な検索 <u>tag=""を使った他の複雑な例</u> "confuse.*" [tag="IN" | tag="PP"] "confuse.*" ([tag="IN"] | [tag="PP"]) "confuse.*" [tag="IN|PP"]

```
<u>turn + 名詞 + 前置詞 の検索式</u>
[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 | | Send feed |
|---|--|-----------|
| Concordance Word List Word Sketch | Query turn, V.*, N.*, PRP 620 (5.50 per million) Page 1 of 31 Go Next Last | |
| Thesaurus | J2W , said" the Americans and Europeans have turned whales into a sacred animal, like the Hindu cow If | |
| Sketch-Diff | J2R Waste and Recycling Old incinerator to turn rubbish into energy and money Plans have been un | veiled |
| Corpus Info | JOP landed property into individual property and turning land into a freely saleable commodity like anything | |
| My jobs | JOW Sex appeal? In a way, the fat lady doctor turned sex on to its head and in those prepermissive | |
| ? | J5J in politics 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 | |
| Homo | JSF way, the intention of these orders er to turn auditors into er snoopers or narks er and to do so l | |
| Home User guide | JSG responsibilities and in some way, as I say, to turn auditors into snoopers and narks er er and make more | |
| oser guide | 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. | |
| Save | J32 Waste and Recycling Australian project turns bottles into pipes A company in South Australia, Ril | 2 |
| as subcorpus | J32 , Rib Loc, has developed a technique for turning plastic bottles into pipes. The pipe-making process was first | |
| View options | J18 which travel in groups: hummingbirds would turn trees into individual feeding territories and thus | |
| KWIC | HRJ television interviewers, et hoc genus omne . It turned Blackpool into a sort of electoral Convention a l'Americain | e |
| Sentence | HRF Sheppey. At the age of eighteen, Doris was turning heads on the island and was selected as Sheerness | |
| Sentence | HRC follow him. | |
| Sort | HRC , and especially Scone. Then they should turn south past Forteviot and march against us. By that | |
| Len | HRD Disc Interactive Whatever may be done to turn CD-ROM into a vehicle for multimedia, it is never likely | |
| Right | HRD option making it possible, if desired, to turn CDTV into an overt computer system. | gh |
| Node | HTP the self in this characteristic way was to turn egoism into altruism, and <corr> aggression </corr> into | 2 |
| Shuffle | Page 1 of 31 Go Next Last | |

◆複雑な検索式◆

<u>"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</u> [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)

<u>名詞 + be + -ed 形で終わる動詞 の検索式</u> [tag="N.*"][lemma="be"][tag="V.*"&word=".*ed"]

<u>look/bring + up/down の検索式</u>

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

<u>OR の意味のバー(|)の例</u> [tag="JJ.*"] [tag="N.*"] "and|or" [tag="N.*"]

1.2.5. 文境界を指定した検索
 ◆within を使った検索式◆
 <u>文境界を指定した式</u>
 [word="confus.*"][tag!="V.*"]*[word="by"]within<s/>

<u>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</u> [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)

1.2.6. CQL attribute の指定

CQL attribute を指定することで、少し検索式が短くなります。が、簡単になるのかは不明。エラーが起きたときにどこが間違っているのか特定しにくくなりそうな印象

| default attribute | 通常の CQL の検索式 | attribute を指定した場合の検索式の例 |
|-------------------|-----------------|-------------------------|
| word | [word= "refil"] | "refill" |
| tag | | "work" |
| lempos | | |
| lempos-lc | | |
| lemma | | |
| lc=word | | |
| lemma-lc | | |

1.3. CQL builder

CQL builder という機能が追加されていますが、CQL を知っていないと活用できないので基礎的なことはしっかり抑えること!

1.4. 新しいインターフェイス

2018 年より新しいインターフェイスが導入されているが、右下に表示される[Back to the original interface]をクリックすることで従来のインターフェイスに戻ることができる。

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: No | ne (whole corpus) Info |
| Search attribute: WO | vrd 🔹 |
| | use n-grams. Value of n: from 2 V to 2 V @ |
| | hide/nest sub-n-grams |
| Filter options: | |
| Filter word list by: R | tegular expression: |
| N | Ainimum frequency: 5 |
| N | Aaximum frequency: 0 (0 = no maximum frequency) |
| V | Vhitelist: ファイルを選択 選択されていません Clear |
| В | Aacklist: ファイルを選択 選択されていません Clear format |
| Include non-word | ls |
| Output options: | |
| Frequency figures: | Hit counts Document counts ARF |
| Output type: | • Simple |
| | ○ Keywords |
| | Reference (sub)corpus English Web 2013 (enTenTen13) |
| | Prefer: rare words common words 1 |
| | O Change output attribute(s) |
| | Y Y |
| | You can select one or more output attributes. Please note that this option can be time- consuming. |
| Make word list | |

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。

| Sketch Engine | ▼ Q ■ British National Corpus (BNC) / Written_Medium_Book |
|----------------------------------|---|
| Home Concordance Word list | Word list Corpus: British National Corpus (BNC) |
| Word sketch Thesaurus | Subcorpus: Written_Medium_Book Page 1 Go Next > <u>lemma Freq</u> |
| Trends Corpus info | the 3.246.582 be 2.120.850 of 1.701.293 and 1.386.310 |
| My jobs User guide ☑ | to <u>1,352,316</u> a <u>1,121,858</u> in <u>1,026,932</u> |
| Save Change options | have 660,482 that 553,152 it 521,370 for 420,060 |
| | he 409.936 not 402.571 I 398.018 as 385.700 |
| | with 343,420 on 338,677 you 290,158 |
| | his 280,464 she 265,410 by 263,321 at 262,123 |
| | do <u>253,091</u> |

(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):

2.2. 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)

| Sketch | | ▼ Q Send fe |
|--|--|-------------|
| Concordance Word List Word Sketch Thesaurus Sketch-Diff Corpus Info My jobs ? Home User guide | Thesaurus Entry Form Lemma: love Part of speech: noun Advanced options Advanced options Maximum number of items: 60 Minimum score: 0.0 Headword in word cloud Cluster items Minimum similarity Detween cluster items: Show Similar Words Save Options | |

図1

| Sketch | | | • Q Send feed |
|--|--------------------|--|--|
| Concordance Word List Word Sketch Thesaurus | LOVE (no Bri | oun) tish National Corp Score Freq | us freq = <u>14,215</u> (126.71 per million) |
| Sketch-Diff | desire | 0.195 5,636 | |
| Corpus Info | feeling | 0.189 12,135 | helief idea |
| My jobs | passion | 0.185 2,463 | character iov interest |
| 0 | pleasure | 0.183 5,567 | happiness the utb anger SDIFIC vision L SOFt |
| U. | thing | 0.180 74,841 | image UUUI God Deauty |
| | faith | 0.1/6 5,186 | experience[N]Ngknowledge |
| Home | <u>tear</u> | 0.173 9,006 | tespert story |
| User guide | life | 0.173 14,301 | |
| - | spirit | 0.171 8.200 | kind C peace D D C C I O D C |
| | experience | 0.169 21.983 | identity quality TearDdSSIUI Feeling |
| Clustering | sense | 0.168 20,793 | thou tothe und sense I music money 5 |
| Save | understanding | 0.166 7,615 | |
| | truth | 0.166 8,391 | |
| | thought | 0.164 13,383 | strength power understanding marriage |
| | God | 0.161 19,332 | friendship commitment |
| | <u>beauty</u> | 0.160 4,294 | freedom |
| | <u>character</u> | 0.159 12,312 | emotion |
| | relationship | 0.159 18,592 | |
| | iov | 0.158 2,582 | |
| | triendship idea | 0.156 2,325 | |
| | kind | 0.154 31,763 | |
| | image | 0 149 10 797 | |
| | sort | 0.149 22.262 | |
| | care | 0.148 19,561 | |

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

| Sketch | | | | | | | | | | | | | | • Q | Send feed |
|--|---------------|-------------------|------------|------|------------------|-----------------------|-----------------|------------------------|---------------------|------------|----------------|-----------|----------|------|-----------|
| Concordance Word List Word Sketch Thesaurus | love/c | des 4.0 | ire | Br | itish Na -2.0 | tional Corpus -4.0 | freqs : -6.0 | = <u>14,2</u> desir | <u>15 5,</u> e | <u>636</u> | | | | | |
| Sketch-Diff | and/or | 2,275 | 1,080 | 1.50 | 1.20 | subject_of | 716 | 353 | 1.10 | 1.00 | adj_subject_of | 190 | 65 | 1.40 | 0.80 |
| Corpus Info My jobs | hate | 18 | 0 | 7.7 | | hurt | <u>25</u> | 0 | 7.5 | | patient | 5 | 0 | 6.8 | |
| 2 | compassion | <u>16</u> | 0 | 7.4 | | triumph | <u>5</u> | 0 | 7.3 | | stronger | Z | 0 | 6.6 | |
| | joy | <u>29</u> | 0 | 7.3 | | conquer | Z | 0 | 7.2 | | wonderful | <u>11</u> | 0 | 6.3 | |
| Hama | forgiveness | <u>12</u> | 0 | 7.1 | | exist | <u>10</u> | 0 | 4.9 | | blind | <u>5</u> | 0 | 6.0 | |
| Home User guide | respect | <u>35</u> | 0 | 7.1 | | surround | <u>8</u> | 0 | 4.6 | | strange | 5 | 0 | 4.7 | |
| Oser guide | tenderness | <u>11</u> | 0 | 7.0 | | die | <u>10</u> | 0 | 4.1 | | satisfied | 0 | <u>6</u> | | 6.1 |
| | warmth | <u>20</u> | 0 | 6.9 | | grow | <u>11</u> | 0 | 4.1 | | | | | | |
| Change options | devotion | <u>12</u> | 0 | 6.8 | | love | 5 | 0 | 3.6 | | | | | | |
| | lust | <u>10</u> | 0 | 6.7 | | fail | <u>6</u> | 0 | 3.2 | | | | | | |
| | friendship | <u>19</u> | 0 | 6.6 | | turn | <u>11</u> | 0 | 2.8 | | | | | | |
| | peace | <u>41</u> | 0 | 6.6 | | lose | <u>8</u> | 0 | 2.7 | | | | | | |
| | companionship | 7 | 0 | 6.4 | | break | 5 | 0 | 2.7 | | | | | | |
| | romance | <u>10</u> | 0 | 6.4 | | change | <u>6</u> | 0 | 2.6 | | | | | | |
| | loyalty | <u>15</u> | 0 | 6.4 | | mean | <u>9</u> | 0 | 2.3 | | | | | | |
| | grace | <u>12</u> | 0 | 6.3 | | remain | <u>5</u> | 0 | 2.2 | | | | | | |
| | jealousy | <u>8</u> | 0 | 6.3 | | come | <u>19</u> | 0 | 2.2 | | | | | | |
| | kindness | <u>8</u> | 0 | 6.3 | | begin | <u>8</u> | 0 | 2.2 | | | | | | |
| | happiness | <u>11</u> | 0 | 6.2 | | inspire | <u>10</u> | <u>5</u> | 6.0 | 5.1 | | | | | |
| | affection | <u>50</u> | <u>5</u> | 8.4 | 5.5 | bring | <u>15</u> | 2 | 2.8 | 2.1 | | | | | |
| | passion | <u>19</u> | <u>10</u> | 6.6 | 6.0 | drive | 0 | 10 | | 4.1 | | | | | |
| | fantasy | 0 | <u>10</u> | | 6.4 | dominate | 0 | 5 | | 4.1 | | | | | |
| | hunger | 0 | Z | | 6.4 | influence | 0 | 2 | | 4.7 | | | | | |

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

【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

| Corpora + Create corpus | Corpora: | Recent | My own | Featured | Parallel | All | | |
|------------------------------|-------------------------------|---------------------|-------------------|--------------|----------|------|-------|--|
| + WebBootCaT + Upload TMX | Language No corpora found | Name | | | | Work | rds 🌲 | |
| Compare corpora My jobs | + <u>Create new</u> れかクリック | corpus 🕂 <u>V</u> | <u>VebBootCaT</u> | ➡ Upload TMX | | | | |

(ii) コーパスの名前と扱う言語の設定

| Sketch | | About Home Settings Change password Log out Search in Help |
|---|--|--|
| user: Prof. Yukio Tono used word | ls: 20 % / 10,000,000 days left: 267 Searc | in Kaken RLD |
| Corpora Create corpus WebBootCaT Upload TMX Parallel corpora Compare corpora Configuration templates Sketch grammars Subcorpus definitions User groups | Corpus name Language English | Cancel Create |
| Support Help index Penort an error | | |

(iii)ファイルの選択

| Corpora | English: Add new file: Step 1 Supported file types: .doc, .docx, .htm, .html, .pdf, .ps, .tar.bz2, .tar.gz, .tgz, .tmx, | | | | | | |
|---|--|--|--|--|--|--|--|
| Parallel corpora Compare corpora Configuration templates Sketch grammars Subcorpus definitions User groups | 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). | | | | | | |
| Corpus | You are responsible for the copyright and otkerUntæll&ctual property issues of the uploaded content. | | | | | | |
| ➢ Corpus page ☆ Add new file ☆ Add web data (BootCaT) | Upload from disk 🔵 ファイルを選択 選択されていません | | | | | | |
| Compile corpus | Download from location 💿 http:// | | | | | | |
| Extract keywords & terms Configure corpus | Use file or directory on the server the server subdirectories | | | | | | |
| Change sketch grammar Set subcorpus definitions Expert mode | FTP to <u>the.sketchengine.co.uk</u> at port 10021 to upload files. Use the same user name and password as for logging | | | | | | |
| Download corpus | into this web interface. Paste text | | | | | | |
| ○ View Inst →ライブラリからファイル | ーを選択し、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

| Sketch tengine Settings Change password Search | | | | | | | | | |
|--|--------|-------------------------------------|--|--|----------|-----------------|----------|----------|-------|
| user: Prof. Yukio Tono used word | ds: 34 | % / 10,000,000 days left: 252 | | Search | | | in ELT-C | ourseBoo | ok-02 |
| Corpora Corpora Create corpus WebBootCaT Create corpus WebBootCaT Create corpus WebBootCaT Create corpus Add new file / ◆ Add data from web using WebBootCaT / ③ Compile corpus / ④ Search corpus | | | | | | | | | |
| Parallel corpora | # | Original file | Plain text | Vertical | Tokens 모 | Owner | | | |
| Configuration templates | 1 | 001 A1 Internatiistening.xml.txt | Ś | 1 | 12,255 | Prof. Yukio Tor | IO 🔍 | 🖉 🗙 | |
| Sketch grammars | 2 | 002 A2 Internati2 Access.xml.txt | I. | 1 | 18,559 | Prof. Yukio Tor | io 🔍 | 🖉 🗶 | |
| Subcorpus definitions | 3 | 003 B1 Internatiistening.xml.txt | s and a second s | s de la constanción de la constancición de la constanción de la constanción de la co | 23,818 | Prof. Yukio Tor | io 🔛 | Ø 🗙 | |
| User groups | 4 | 004 B2 Internatior Listening.txt | - | « | 36,756 | Prof. Yukio Tor | IO 🔍 | 1 X | |
| | 5 | 005 C1 Internati5 Expert.xml.txt | 1 | ~ | 47,526 | Prof. Yukio Tor | ю 👳 | 1 X | |
| Corpus | 6 | 006 C2 Internati Mastery.xml.txt | - | ~ | 34,859 | Prof. Yukio Tor | ю 🔍 | 1 X | |
| | 7 | 007 B1 Close-Up.xml.txt | 1 | 1 | 16,776 | Prof. Yukio Tor | ю 🔍 | 1 X | |
| + Add new file | 8 | 008 A1 English Explorer 1.xml.txt | al and a second | « | 8,699 | Prof. Yukio Tor | ю 📃 | 1 X | |
| 💠 Add web data (BootCaT) | 9 | 009 B1 English Explorer 4.xml.txt | 1 | 1 | 17,214 | Prof. Yukio Tor | io 📃 | 1 × | |
| 🕒 Compile corpus | 1 | 0 010 A1 Holiday Explorer 1.xml.txt | al and a second | « | 3,660 | Prof. Yukio Tor | io 📃 | ð 🗙 | |
| Search corpus | 1 | 1 011 B1 Just Righrmediate.xml.txt | 1 | 1 | 10,587 | Prof. Yukio Tor | io 🔍 | 1 X | |
| Extract keywords & terms | 1 | 2 012 B1 Just Righo Script.xml.txt | s. | « | 4,642 | Prof. Yukio Tor | ю 🛡 | Ø 🗙 | |

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 | WebBootCaT: Cre | eate corpus 🕜 | 図 1 |
|--|---------------------------------|--|-----|
| Parallel corpora Compare corpora My jobs | Corpus name Language | ····································· | |
| Advanced features | Input type | be automatically tokenised. | |
| Corpus templates Sketch grammars Subcorpus definitions GDEX configurations User groups | input type | Seed words URLs Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs. | |
| Subscription overview Support | Seed words | | |
| User guide Feedback | | 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 > | |

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)。

| | | Cancel Next> 図3:URLの指定 |
|---|---|------------------------|
| Home | Sample: page: Edit file | |
| + Create corpus + WebBootCaT + Upload TMX | File type [HTML | |
| Parallel corpora Compare corpora My jobs | Character encoding UTF-8 (all languages) If some characters are not displayed correctly in the preview, changing this may help. | |
| Advanced features | Cancel Update preview Finish | |
| Sketch grammars Subcorpus definitions GDEX configurations User groups Subscription overview File / Edit file Delete file View plain text III View vertical Download plain text Download plain text Download plain text Download vertical Edit metadata Support Usersenide | File preview (plain text) Is set Showing bytes 11500 / 2887 25 21 (>) The Briet (#)> (>) The Briet (#)> (>) Chards's Prime Minister Justin Trudeau has condemed the beheading of Alberta wan John Ridsdel by Islanist extremists as a "heinous act." Ridsdel was one of four hostages abducted in the Philippines by Abu Sayat militants in september (#)> (>) The threat of 21ka has thrown an unexpected and unwelcome winkle in the plans of American athletes as they prepare for the most important moments of their lives this sourcer at the 2016 Rio Olympics (#)> (>) Frontrumers Hillary Clinton and Donaid Trume are likely to sweep the five down ballot that could aftect Congress next year. (#)> (>) A federal appeals court has ruled that New England Patriots Ton Brady most serve a four-game "Deflategate" suspension imposed by the NFL, overturning a lower Judge and siding with the league in a battle with the players union (*)> (>) Bore answers are needed, says an independent panel of experts in a new 800-bage document, which takes arough that if he loses the Democratic nomination, he will help Hillary Clinton in the fight against Republicans. But he expects her to back his Medicare plan, and for her to take a tougher stance on climate her back his Medicare plan, and for her to take a tougher stance on climate her back his Medicare plan, and for her to take a tougher stance on climate her back his Medicare plan, and for her to take a tougher stance on climate her back his Medicare plan, and for her to take a tougher stance on climate her back his Medicare plan, and for her to take a tougher stance on climate her back his Medicare plan, and for her to take a tougher stance on climate her back his Medicare plan, and for her to take a tougher stance on climate her her her her standers of her her her take tougher stance on climate her her her her her her her her her her | |

- 4. 練習問題
- 4.1. 基本検索1

練習問題

~基本検索画面編~

以下の1~8に挙げる用例を抽出したいときの基本検索を用いた検索方法を考えてみましょう!

1. talk (大文字・小文字・活用形を問わない)を含む用例を抽出したい。 >> [] 2. 名詞としての talk の用例を抽出したい。 1 >> [3. 名詞としての talk のうち、単数形の用例だけ抽出したい。 >> [1 4. 動詞としての Talk (大文字 T で始まる talk)を抽出したい。 >> [1 5. talk over (大文字・小文字・活用形を問わない)の用例を抽出したい。 >> [1 6. talking over の用例を抽出したい。 >> [1 7. trans-(接辞)を含む用例を抽出したい >> [1 8. talk (大文字・小文字・活用形を問わない)のあとに、with, talk, about のいずれも続かない用例を 抽出したい。 >> [1 9. 直前に形容詞で修飾されている名詞の talks (複数形に限る)の用例を抽出したい。 >> [1

4.2. 基本検索 2

練習問題

~基本検索画面編~

以下の1~11に挙げる用例を抽出したいときの基本検索を用いた検索方法を考えてみましょう!

4.3. CQL 検索 1-基本検索と関連付けて-

4.4. CQL 検索 2

4.5. 頻度表

Concordance 画面から frequency 機能を使って、検索結果の頻度内訳を見てみましょう!

疑問1) talk (活用形・大文字・小文字問わない)の検索結果のうち、最も用例数が多い表層形はなんでしょうか?

疑問2) talk の名詞用法のうち、単数形と複数形のどちらがより多く使われるのでしょうか?

疑問3) talk は名詞用法、動詞用法のどちらがより多く使われるのでしょうか?

疑問4) talk の直後に来る前置詞のうち、最も頻度の高いものはなんでしょうか?

1.5. ワードリスト

【実際に検索してみよう】

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

(2)BNC 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

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

- (5)BNC 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