<u>Sketch Engine 操作マニュアル</u>

東京外国語大学 投野研究室

<目次>

1.		Co	oncordance 画面を使いこなす
	1.	1.	検索の基礎・・・・・・p.2
	1.	2.	CQL 検索
2.		さ	まざまなオプション機能
	2.	1.	Word listp.8
	2.	2.	Word sketch·····p.11
	2.	3.	Thesaurus / Sketch diffp.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.	WebBootCaTp.19

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 character	er 😑 CQL							
Lemma:	PC	oS: unspecified 🔻							
Phrase:									
Word form:	Po	oS: unspecified 🔻 🗖 match case							
Character:									
CQL:	De	efault attribute: Ic = word (lowercase)							
	Tagset summary								
Make Concor	Make Concordance 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)
- ・Lemma…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)。
- ・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。
- ・<u>Character</u>…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
$\langle - \rangle$	

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	r: both ▼ 5 ▼ tokens. Window: both ▼ 5 ▼ tokens.
Lemma(s)	all ▼ of these items. PoS: adjective all ▼ of these items. adverb conjunction determiner noun pour singular

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
- ・<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 V	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 <u>Next</u> <u>Last</u> 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	

↑ファイル情報

 \uparrow 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〇語以内にするか決定。 Minimum Frequency in corpus…<u>共起語</u>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

◆実際に見	てみましょう – 対象コーパス:BNC◆
Sketch	
Concordance Word List Word Sketch Thesaurus Sketch-Diff Corpus Info My jobs (2) Home User guide	Simple query: Make Concordance Query type simple Query type simple Query type simple Lemma: Po5: Phrase: Po5: Word Form: Po5: CQL: Default attribute: Image: Image: Image: Default attribute: Word Form: Default attribute: Image: Image: Image: Image:
	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

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

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

<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"]

※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
 ※ *: O以上n個 / +:1以上n個 / ?:O個または1個 /!:~以外

<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文Engine				▼ Q Send fee
Concordance Word List Word Sketch	Query Page	y turn, V.*, N.*, PRP 620 (5.50 per million)		
Thesaurus	JZW	, 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 unveiled
Corpus into	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
? ?	JSJ	in politics Victorian values have	turned Britain into	a more divided country. Homelessness and
	JNF s	ervices it helps youngsters help themselves by	turning moans into	action. Really we're trying to give young
Home	JSF	way, the intention of these orders er to	turn auditors into	er snoopers or narks er and to do so l
liker guide	JSG	responsibilities and in some way, as I say, to	turn auditors into	snoopers and narks er er and make more
Oser Bride	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, Rib
as subcorpus	J32	, Rib Loc, has developed a technique for t	urning 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
www.c	HRJ	television interviewers, et hoc genus omne . It	turned Blackpool into	a sort of electoral Convention a l'Americaine
Contoneo	HRF	Sheppey. At the age of eighteen, Doris was	turning heads on	the island and was selected as Sheerness
Sentence	HRC	follow him. But then he could have	turned north to	the Tay in safety. With a tired army. But
Sort	HRC	, and especially Scone. Then they should	turn south past	Forteviot and march against us. By that
Left	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. Although
Node	HTP	the self in this characteristic way was to	turn egoism into	altruism, and <corr> aggression </corr> into
References Shuffle	Page	1 of 31 Go <u>Next</u> <u>Last</u>		

でた~~~ 図2

◆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.*[~]]

◇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)

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	0								
Subcorpus:	None (whole corpus)								
Search attribute:	word								
	use n-grams. Value of n: from 2 v to 2 v								
	hide/nest sub-n-grams								
Filter options:									
Filter word list b	y: Regular expression:								
	Minimum frequency: 5								
	Maximum frequency: 0 (0 = no maximum frequency)								
	Whitelist: ファイルを選択 選択されていません Clear								
	Blacklist: ファイルを選択 選択されていません Clear format								
Include non-	words								
Output option	15:								
Frequency figu	res: Hit counts Document counts ARF								
Output ty	/pe: 🖲 Simple								
	Keywords								
	Reference (sub)corpus English Web 2013 (enTenTen13)								
	Prefer: rare words common words 1								
	Ochange output attribute(s)								
	¥ ¥								
	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 Sketch Eng	ine homepag	Q British National Corpus (BNC) / Written_Medium_Book					
Home Concordance	Word lis Corpus: B	st British National Corpus (BNC)					
Word list Word sketch Thesaurus	Subcorpus: Written_Medium_Book Page 1 Go <u>Next ></u>						
Sketch diff Trends	<u>lemma</u> the be	<u>Freq</u> <u>3,246,582</u> <u>2,120,850</u>					
Corpus info My jobs	of and to	<u>1,701,293</u> <u>1,386,210</u> <u>1,352,316</u>					
User guide 🗹	a in have	<u>1,121,858</u> <u>1,026,932</u> <u>660,482</u>					
Save Change options	that it for	<u>553,152</u> <u>521,370</u> <u>420,060</u>					
	ne not I	<u>409,936</u> <u>402,571</u> <u>398,018</u>					
	as with on	<u>385,700</u> <u>343,420</u> <u>338,677</u>					
	you his she	<u>290,158</u> <u>280,464</u> <u>265,410</u> 201					
	by at do	<u>263,321</u> <u>262,123</u> <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):

【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 \pounds .*ing. Frequency figures \pounds 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。
- 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													
Concordance	make	(verb)	Alter Natio	native PoS: <u>noun</u> (799) <u>adjec</u> reg - 209	tive (1)	868 97 per million)						
Word list					req - <u>207</u>	,007 (1	,000.77 per miciony						
Word sketch	modifiers of "	make"		objects of "make	<u>2"</u>		subjects of "make"			<u>"make" and/o</u>	<u>r</u>		preposition
Thesaurus	1	22,331	0.40		122,259	5.60		38,330	2.30		2,394	0.10	
Sketch diff	come +	1.5/8		decision +	2,940	9.51	decision +	<u>403</u>	8.24	break +	<u>103</u>	10.14	"make" by
Trends	make sure			sense +	2,530	9.31	decision making	g 5 10	7.50	make or b	reak	0.54	<u>"make" 1n</u>
Corpus info	also 🕇	<u>1,410</u>	8.33	make sense	2 (22	0.24	people +	<u>549</u>	7.58	try +	<u>159</u>	9.50	<u>"make" of</u>
My jobs	also made	7.17	7.00	use +	<u>2,423</u>	9.24	people make	20/	7 40	to try and	make	0.00	"make" for
	only +	<u>141</u>	7.93	make use of	4 (27	0.70	company +	300	7.48	design	, 22	9.20	<u>make to</u>
User guide 🗠	only make	0.40	7.04	mistake +	1,02/	8.73	man 🛨	300	7.34	aesignea	ana maa	1e	make from
	just +	840	7.91	way +	1,900	8.03	government +	<u>301</u>	7.33	sett	00	9.01	make on
Save	just make		7.00	point +	1,000	8.02	person +	179	0.95	making ar	a settin	8	make with
Change options	already +	401	7.83	difference +	1,558	8.61	person making		(00	go	. 92	8.30	<u>"make" at</u>
Cluster	already mo	ide		contribution +	<u>1,410</u>	8.52	friend +	105	6.89	go and mo	ike a		<u>"make" abou</u>
Sort by freq	actually +	326	7.81	effort +	<u>1,384</u>	8.48	my hon. friend	makes		use	36	8.28	"make" unde
Hide gramreis	actually m	ake		statement +	<u>1,308</u>	8.39	woman +	<u>205</u>	6.85	unmake	<u>20</u>	8.08	"make" betw
More data	ever 🕇	<u>304</u>	7.79	make a state	ment		god +	<u>149</u>	6.79	to make o	r unmak	е	"make" into
Less data	ever made			attempt 🕇	<u>1,274</u>	8.37	god made			buy	<u>24</u>	7.82	<u>"make" as</u>
	then 🕇	<u>544</u>	7.78	progress +	<u>1,102</u>	8.17	policy +	<u>133</u>	6.63	make or b	uy		"make" durir
	then made			change 🕇	<u>1,105</u>	8.08	policy making			receive	<u>19</u>	7.73	"make" again
	always 🕇	<u>436</u>	7.61	order 🕇	<u>907</u>	7.84	party 🕇	<u>147</u>	6.59	do	<u>40</u>	7.59	"make" over
	always ma	de		profit +	<u>867</u>	7.81	party made			give	22	7.58	"make" after
	never 🕇	<u>451</u>	7.53	money +	<u>936</u>	7.80	member +	<u>127</u>	6.49	given or n	nade		"make" befor
	never mad	е		make money			minister +	<u>114</u>	6.40	bear	<u>15</u>	7.48	"make" with
	not +	4 873	7 52	love +	816	7 73	minister mode			horn not	made		"make" withi

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。

Home	L (m)
Concordance	Make British National Corpus (BNC) freq = 209,867 (1,868.97 per million)
Word list	modifiers of "make"
Word sketch	22.331 0.40 122.259 5.60 38.330 2.30 2.394 0.10 22.579
Thesaurus	sure + 1,578 10.99 decision + 2,940 9.51 decision + 403 8.24 break + 103 10.14 "make" by 4,267 3.40
Sketch diff	make sure sense + 2,530 9.31 decision making make or break "make" in 3,613 0.80
Trends	also + <u>1,410</u> 8.33 make sense people + <u>549</u> 7.58 try + <u>159</u> 9.56 <u>make" of 2,739</u> 0.30
Corpus info	also made use + 2,423 9.24 people make to try and make
Wy jobs	only + 747 7.93 make use of company + 305 7.48 design 55 9.20
Hor guide 🖉	only make designed and made
oser guide 🗆	naticle site "make" [naticle site "make" with chief [nearginal object of "make"] [nearginal subject of "make"] [why words following "make"]
	articles arter make particles arter make with operating promonia doperation make promonia doperation doperatio
Save	up + 2.761 9.89 up + 3.254 10.57 it + 14.384 10.50 they + 3.498 8.13 which + 111 8.96
Cluster	made up of make up make it they made made which
Sort by freg	off + 242 8.24 out + 1.073 9.28 me + 3.174 9.88 he + 4.919 7.99 whatever 18 8.26
Hide gramrels	made off with make out made me he made to make whatever
More data	out + <u>710</u> 7.81 through <u>36</u> 7.79 them + <u>3.336</u> 9.73 it + <u>5.578</u> 7.97 that <u>60</u> 8.11
Less data	made out of made it through to the make them it makes make sure that
	infinitive objects of "make" -ing objects of "make" adjectives after "make" and noun advectives after "make"
	<u>3,304</u> 0.90 <u>1,106</u> 0.6 <u>18,678</u> 47.00 <u>11,352</u> 3.10
	feel + <u>133</u> 9.69 concern <u>38</u> 9.86 <u>1.792</u> to sure + <u>3.664</u> 12.13
	made to feel made concerning the made it clear that to make sure
	ensure <u>74</u> 8.58 regard <u>24</u> 9.17 easy + <u>1.317</u> 11.01 clear + <u>1.039</u> 10.60
	be made to ensure that made regarding the difficult + 1.293 10.93 made clear made clear
	$\frac{100 \text{ k}}{\text{mode to look}} = \frac{22}{3} \circ .13 \text{Los} = \frac{72}{3} \circ .1 \qquad \text{available } + \frac{1}{1.210} \cdot 10.37 \text{be mode available}$
	more to took De more dump

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)

Concordance Word List Word Sketch Thesaurus Lemma: love	Sketch		• Q Send fe
Sketch-Diff Part of speech: noun * Corpus Info Advanced options My jobs Advanced options Advanced options Advanced options Home Maximum number of items: 60 User guide Minimum score: 0.0 Headword in word cloud Cluster items: Minimum similarity 0.15 Show Similar Words Save Options	Concordance Word List Word Sketch Thesaurus Sketch-Diff Corpus Info My jobs 2 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 between cluster items: Show Similar Words Save Options	



検索結果					
Sketch文Engine				<u>۰</u> ۵	Send feed
Concordance Word List Word Sketch Thesaurus	Love (no Lemma	oun) itish Nat Score	ional Corj Freq	ous freq = <u>14,215</u> (126.71 per million)	
Sketch-Diff	<u>desire</u>	0.195	5,636		
Corpus Info	feeling	0.189	12,135	, belief idea	
My jobs	pleasure	0.185	2,463 5 567	character coirit Joy interest	
?	thing	0.180	74,841	images truth are Spin Cod beauty	
	faith	0.176	5,186	avportion cothing knowlodge	
Hama	<u>fear</u>	0.173	9,006	experience crimig knowledge	
Home User guide	knowledge	0.173	14,381	history Doacurofaithlifo	
User guide	<u>life</u>	0.172	62,484		
	spirit	0.1/1	8,200	identity quality tearDassionteeling	
Clustering	sense	0.168	20,793	thout the sense music more 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	concern pain friendship commitment	
	beauty	0.160	4,294	freedom	
	<u>character</u>	0.159	12,312	enoción	
	relationship	0.159	18,592		
	<u>Joy</u> friendship	0.156	2,302		
	idea	0.154	31,963		
	kind	0.152	26,388		
	image	0.149	10,797		
	sort	0.149	22,262		
	care	0.148	19,561		

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。

Sketch														• Q	Sen
Concordance Vord List	love/	les	ire	Br	itich Na	tional Corpur	from	- 14 2	15 5	626					
ord Sketch		40	2.0	0	-2.0	-4.0	-6.0	- <u>14,2</u>		030					
hesaurus		-10	2.0		2.0	4.0	0.0	dean	-						
ketch-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
orpus Into	hate	<u>18</u>	0	7.7		hurt	<u>25</u>	0	7.5		patient	5	0	6.8	-
)	compassion	<u>16</u>	0	7.4		triumph	5	0	7,3		stronger	Z	0	6.6	-
	joy	<u>29</u>	0	7.3		conquer	Z	0	7.2		wonderful	11	0	6.3	
	forgiveness	<u>12</u>	0	7.1		exist	<u>10</u>	0	4.9		blind	5	0	6.0	
ome	respect	<u>35</u>	0	7.1		surround	<u>8</u>	0	4.6		strange	5	0	4.7	-
ser 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						
hange 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	Z	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	9	0	2.3						
	grace	<u>12</u>	0	6.3		remain	5	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	19	<u>10</u>	6.6	6.0	drive	0	<u>10</u>		4.1					
	fantasy	0	<u>10</u>		6.4	dominate	0	5		4.1					
	hunger	0	Z		6.4	influence	0	2		4.7					

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(<u>http://www.w3schools.com/xml/xml_tree.asp</u>) \downarrow $\vartheta >$

【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

Corpora + Create corpus	Corpora: Recent My own Featured Parallel	All
+ WebBootCaT	Language 🔶 Name	🔶 Words 🔶
Parallel corpora	No corpora found	
Compare corpora		
My jobs		
<u>ل</u> اب	ヂれかクリック	

(ii) コーパスの名前と扱う言語の設定

Sketch				About	Home	Settings Search	Change password	Log out
user: Prof. Yukio Tono used word	ls: 20 % / 10,000,000 days left:	267	Search			in Ka	ken RLD	
Corpora Create corpus WebBootCaT Upload TMX Parallel corpora Compare corpora Configuration templates Sketch grammars Subcorpus definitions User groups	Create new corpo	US English		Cano	cel	• Create		
Support Help index Penort an error								

(iii)ファイルの選択

Corpora 🐥 Create corpus	English: Add new file: Step 1						
💠 WebBootCaT 🔶 Upload TMX	Supported file types: .doc, .docx, .htm, .html, .pdf, .ps, .tar.bz2, .tar.gz, .tgz, .tmx, .txt, .vert, .xml, .zip.						
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 other intellectual property issues of the uploaded content.						
👝 Corpus page							
💠 Add new file	Upload from disk Oファイルを選択 選択されていません						
4 Add web data (BootCaT)							
Compile corpus	Download from location 💿 http://						
Search corpus							
Extract keywords & terms	Use file or directory on 🕖						
Change sketch grammar	show files in subdirectories						
Set subcorpus definitions	FTP to the sketchengine.co.uk at port 10021 to upload						
Expert mode	files. Use the same user name and password as for logging						
Download corpus	into this web interface.						
Access privileges	Paste text						
O View logs							

→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

Sketch tengine About Home Settings Change password Search								
user: Prof. Yukio Tono used word	ls: 34 %	days left: 252		Search		i	n ELT-CourseBoo	ok-02
Corpora Create corpus WebBootCaT Upload TMX	EL kake	T-CourseBook-01 en_rld Add new file / & Add data from web using	WebBootCaT /	Compile	e corpus / 🕷 S	クリッ earch corpus	ク	
Parallel corpora	#	Original file	Plain text	Vertical	Tokens 🗟	Owner		
Configuration templates	1	001 A1 Internatiistening.xml.txt	st.	1	12,255	Prof. Yukio Tono	o 🔍 🧶 🌾	
Sketch grammars	2	002 A2 Internati2 Access.xml.txt	I	1	18,559	Prof. Yukio Tono	s 🔍 🖉 🇶	
Subcorpus definitions	3	003 B1 Internatiistening.xml.txt	I	1	23,818	Prof. Yukio Tono	o 🔍 🖉 🗶	
User groups	4	004 B2 Internatior Listening.txt	1	1	36,756	Prof. Yukio Tono	. 🛡 🖋 🕷	
	5	005 C1 Internati5 Expert.xml.txt	1	1	47,526	Prof. Yukio Tono	o 🔍 🧶 🗶	
Corpus	6	006 C2 Internati Mastery.xml.txt	1	1	34,859	Prof. Yukio Tono	o 🔍 🍠 🌋	
	7	007 B1 Close-Up.xml.txt	1	1	16,776	Prof. Yukio Tono	o 🔍 🍠 🗶	
Add new file	8	008 A1 English Explorer 1.xml.txt	-	1	8,699	Prof. Yukio Tono	o 🔍 🧶 🗶	
👍 Add web data (BootCaT)	9	009 B1 English Explorer 4.xml.txt	1	4	17,214	Prof. Yukio Tono	o 🔍 🍠 🗶	
Compile corpus	10	010 A1 Holiday Explorer 1.xml.txt	al and a second	1	3,660	Prof. Yukio Tono	o 🔍 🖉 🍂 🕱	
Search corpus	11	011 B1 Just Righrmediate.xml.txt	1	1	10,587	Prof. Yukio Tono	o 🔍 🖉 🗶	
Extract keywords & terms	12	012 B1 Just Righo Script.xml.txt	I	1	4,642	Prof. Yukio Tono	o 🔍 🖉 🗶	

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: Create corpus 🚱							
 WebBootCaT Upload TMX 	Get seed words from Wikipedia							
Parallel corpora Compare corpora	Corpus name							
My jobs	Language							
Advanced features	WebBootCaT is unavailable for languages which cannot be automatically tokenised.							
Corpus templates Sketch grammars Subcorpus definitions	Input type Seed words URLs							
GDEX configurations User groups	Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs.							
Subscription overview	Seed words							
Support								
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 >							

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 $(3 \sim 20 \text{ } \text{b})$ を入力。
- 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。

Home + Create corpus	WebBootCaT: Create corpus 😮
 WebBootCaT Upload TMX Parallel corpora Compare corpora My jobs 	Select URLs to download Query: create corpus education Mathematical Action Act
Advanced features Corpus templates Sketch grammars Subcorpus definitions GDEX configurations User groups Subscription overview Support User guide Feedback	 http://www.caller.com/news/education/ http://www.jagoinvestor.com/forum/creating- corpus-for-childs-education-and-marriage https://www.linkedin.com/jobs/training- education-jobs-corpus-christi-tx https://www.linkedin.com/jobs/education- jobs-corpus-christi-tx http://www.careerbuilder.com/jobs-education- in-corpus-christi,tx http://www.careerbuilder.com/jobs-education- in-corpus-christi,tx http://www.facebook.com/CCEducationFoundation unselect all invert selection
	Cancel Next >

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)。

Home + Create corpus	WebBootCaT: Create corpus 😮							
+ WebBootCaT	Get seed words from Wikipedia							
+ Upload TMX Parallel corpora Compare corpora	Corpus name Sample							
My jobs	Language English 🔻							
Advanced features	WebBootCaT is unavailable for languages which cannot be automatically tokenised.							
Sketch grammars	URLs							
Subcorpus definitions GDEX configurations User groups Subscription overview	Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs. URLs <u>http://time.com/</u>]							
Support								
User guide Feedback								
	List of URLs separated with whitespace.							
	Compile corpus when finished Automatically compile corpus when WebBootCaT processing is finished.							
	Show advanced options							
	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	Character encoding UTF-8 (all languages)	
My jobs	If some characters are not displayed correctly in the preview, changing this may	
	help.	
Advanced features	Canada Uladata ami'am	
Corpus templates	Cancel Opdate preview Prinsh	
Sketch grammars		
Subcorpus definitions	File preview (plain text)	
GDEX configurations	I≤ ≤≤ Showing bytes 11500 / 2887 ≥≥ ≥1	
User groups Subscription overview		
subscription overview	<	
file	Alberta man John Ridsdel by Islamist extremists as a "heinous act." Ridsdel was one of four hostages abducted in the Philippines by Abu Sayyaf militants in	
File	September	
Edit file	plans of American athletes as they prepare for the most important moments of	
Delete file	Control of the sound of th	
III View vertical	Northeastern states holding primaries on Tuesday, but there's plenty of drama down ballot that could affect Congress next year.	
Download original	A federal appeals court has ruled that New England Patriots Tom Brady must serve a four-wave "Deflaterate" suspension imposed by the NFL, overturning a	
🛓 Download plain text	lower judge and siding with the league in a battle with the players union	
Download vertical	608-page document, which takes apart the official account of what happened to	
Edit metadata	the 43 students who went missing in Iguala in 2014	
	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 toucher stance on climate	
Support	change (/p) (c) Delice in rist way deployed in force in Cairo ² s central districts	
User guide	cordoning off locations where dozens of activists h	
Feedback		

図4:完成したコーパス