

# Task-Based Language Teaching (TBLT): From Theory to Practice

## **Using Corpora for Analysis of Discourse**



**Marc Jones** 



## Using Corpora for Analysis of Discourse

### Contents

Introduction	3
Scraping	3
BootCaT	3
SketchEngine	8
Compiling your own corpus (using SketchEngine)	10
Corpus Analysis	11
SketchEngine	12
AntConc	14
BYU	20
Analysing corpora for TBLT	22



### Introduction

Corpora are simply collections of texts, though usually they are sometimes very large, such as the 14-million-word iWeb corpus (available through Brigham Young University's corpus interface, which will be discussed below). Other small, specialized corpora can be around tens of thousands of words.

Corpora are not new. Many people know the word corpus from the Latin, and the name Corpus Christi, or 'the body of Christ'. The term comes from monastic collections of texts on certain subjects. Those monks required a lot of reading and purposeful human searching which can now be accomplished in a matter of seconds, using computer software called a **concordancer**, or in some cases just very advanced text-editing software. While this may make things sound easy, it might not always be. There are times when corpus results may go against one's expectations. However, this is not a negative; one of the bonuses of using corpora is to avoid wholly relying upon intuition and instead draw upon authentic language use when designing tasks.

### Scraping

When using corpora for discourse analysis (or analysis of discourse) in task-based language teaching, there are two main things we can do: gathering discourse -- often by scraping the web -- and analysing. Scraping is a process of automatically pulling text from websites and is one of the most convenient ways to build a corpus. If you want a written corpus this is especially useful. Even if you want a spoken corpus, it might be difficult to get access to real speech in the context you want to analyse discourse in (especially medical contexts), but samples of written discourse can provide the course designer and teachers with lexis and grammar that has a high probability of coming up.

For beginners, you can scrape the web in two ways: one with the **BootCaT** software installed on your computer (PC/Mac/Linux, free), and another with **SketchEngine**, a web-based tool (subscription, free trial available).

#### BootCaT

In this example we are going to scrape the web of cooking-related language.

 Check you have the current version of Java by downloading it from <u>https://www.java.com/en/download/</u>. Install it. It should take about five minutes. Download the BootCaT software from the website, <u>http://bootcat.dipintra.it/</u>, then install it.



#### 2. Start the software.



3. Choose a name for your corpus. I would also recommend giving it a number in case you want to edit it or redo it afterwards.

👮 BootCaT frontend		<u>101</u> 23	×
File Edit Help			
	Project definition		
The	Corpus name FoodNutritionCorpus_written_1		
BootCaT	Language English - United Kingdom		•
Frontend	More options 🧿		
	< Back	Next >	Quit
	< Back	Next >	Quit



4. Click Simple Mode then choose your search engine to be Google. It is not worth the hassle of using Bing.

BootCaT frontend			_		×
File Edit Help	Choose a search engine External Brows	er (Google)			
	< Back		Next	> (	Quit

Choose your seed keywords. If you have multi-word terms these should go into a line by themselves.

👮 BootCaT frontend	-		×
File Edit Help			
The BootCaT Frontend	Insert one seed per line, multi-word seeds go on the same line  Cut Copy Paste food drink recipe cooking nutrition ingredients serving breakfast lunch dinner snack dessert		
	< Back Ne	xt>	Quit

5. Choose your tuples (combinations of seeds). How many seeds should combine and how many tuples do you want? In this case we will just use the default settings.



👮 BootCaT frontend		– 🗆 ×
File Edit Help		
The BootCaT Frontend	The tuples that will	be used as queries
	Generate tuples Edit tuples	Tuple length 3 v N. of tuples (max. 220) 10 v Selected tuples 0
	Click on 'Generate tuples' to proceed	Not > Out

This will then generate the list and you can disable any tuples that you don't want but that is pointless in most cases because BootCaT deletes duplicate data.

BootCaT frontend		-	
	The tuple:	that will be used as queries	
	Inutrition dinner serving		
Ine	🗹 lunch dessert dinner		
BootCaT	cooking recipe nutrition		
Frontend	🗹 nutrition food serving		ľ
Toncend	🗹 dessert snack lunch		
	🗹 recipe lunch breakfast		
	☑ nutrition dessert ingredients		
	🗹 serving recipe snack		
	✓ food dessert cooking		
	dessert recipe cooking		
	Generate tuples	Tuple lengt	dh 3 🌲
	Edit tuples	N. of tuples (max: 220)	30
		Selecte	ed tuples 3
	SBack	Next >	

6. Generate the queries by clicking Open All in Browser, (unless you have an extremely large number of queries, over about 50, otherwise Google might lock your account if you are signed in, or block your IP address temporarily) and saving the Google search results. If Chrome is your default browser you cannot do this from the menu button. Right click on the page and (Save Page As) and choose HTML only. You need to save it into your BootCaT Corpus Directory, which is at My Documents/BootCaT Corpora/Whatever you called your corpus/queries/.



		Click to open querie	in a browcor		
_	10	Click to open quene	s in a browser		
The	Open in Browser	snack cooking recipe			
BootCaT	Open in Browser	snack serving breakfa	at		
Frontond	Open in Browser	snack ingredients cool	king		
rivillend	Open in Browser	dessert recipe cooking			
	Open in Browser	recipe lunch breakfast			
	Open in Browser	ingredients recipe des	sert		
	Open in Browser	drink dessert nutrition			
	Open in Browser	cooking recipe nutritic	n		
JAK A	Open in Browser	dinner nutrition break	fast		
	Open in Browser	lunch dessert dinner			
	Open in Browser	lunch food nutrition			
	Open All in Browse	r]	Open Queries Folder	Collect	URL
	Open All in Browse	r ser, save result pages to 'que	Open Queries Folder	Collect	UR
	URLs'				_
	< Back			Next >	C

Sourcer noncent		~
File Edit Help		
The	Hide advanced options <b>O</b> Limit search to the following Internet domain (e.g., edu):	
BootCaT Frontend	www.google.* www.googleadservices.*	gre. ).
	Restrict search to this document type:	
	Exclude these document types:   Excluded document types:	
	Adult filter (filter sexually explicit material)	MODERATE V
	Maximum number of URLs to return for each tuple	10
		Generate Queries
	Click on 'Generate Queries' to generate search engine queries	0
	< Back	Next > Quit

7. Make the corpus by waiting for BootCaT to pull the websites and clean them.



8. You are now ready to choose your concordancer to do the analysis: see the Corpus Analysis section below.



An alternative to BootCaT, which can be temperamental after Java updates, is SketchEngine. Although this requires a paid subscription (after an initial free trial), it's a powerful tool.

#### SketchEngine

- 1. Go to <u>http://app.sketchengine.eu/</u> and register for a free trial.
- 2. You can build your own corpus or upload one. First we'll show you how to use SketchEngine in the same way as we've just demonstrated with BootCaT, i.e. by scraping the web to build a corpus.
- 3. Log in to SketchEngine, select the **My Corpora** tag and click **New Corpus** on the right.



4. Give your corpus a title, select the language and click **next**.

CREATE CORPUS	type to search Q		٢	Ð	?	
1. CREATE CORPUS > 2. ADD	TEXTS > 3. COMPILE					
	Build your own private cor documents.	pus from texts on the web or from your own				
	Name	FoodNutritionCorpus				
	Language	English				
		MULTILINGUAL				
	Description					
	Storage use	ad: 683,391 of 1,000,000 words (68%)				

5. Select **Find texts on the web**. SketchEngine gives you the option of choosing search words or websites. It is advisable to choose words for a wider variety of discourse. Choose your seed keywords. Separate them by commas. If you have



multi-word terms these should be spaced as normal with a comma after the multi-word terms themselves.

Input type	Web search ⑦		
	O URLs ②		
	O Website <sup>(2)</sup>		
	food X nutrition X recipe X cooking X	?	
	ingredients		
	Number of Bing searches: 4		
Folder name ⊘	web1		
	Web search settings 🔹		
	Black list settings 👻		
	White list settings 👻		
	Size restrictions 🔻		
	✓ Compile when finished ⑦		
	CANCEL GO		

6. You can choose how big the files should be. It is probably best not to fiddle with these numbers. SketchEngine works well with PDF files so big files should be included. Remember that a webpage is usually only a matter of kilobytes so the small size is useful. SketchEngine will tell you when it is finished.

	C	Compiled
		complied
	The corpus is	is compiled and ready to use.
	RECOMPILE	E CORPUS DASHBOARD
	Export col	settings • Log •
	Expert ser	settings - Log -
GET TO KI		RPUSI
6= E	TRACT KEYWORDS & TERF	CORPOS DETAILS AND STATISTICS
		ВАСК



### Compiling your own corpus

In this instance we will look only at SketchEngine, which is by far the easiest as it accepts a wide range of document formats and does all the tagging automatically.

The advantage of compiling your own corpus over scraping is that by collecting examples of specific target discourse. e.g. medical abstracts or transcriptions of Scrum demo dialogues, we can analyse language as it is used to complete the target task, rather than the wider language of the domain.

- Collect your examples of target discourse e.g. in .txt, .doc or .pdf format and place them in a folder on your computer. They could be examples of texts written for a specific purpose, or transcriptions of spoken English that you have typed up yourself or used YouTube to create. See the Session 3 Extension for more information on how to do this.
- 2. Follow the above steps on using SketchEngine to scrape a corpus, until **Step 5**. This time, however, select: **I have my own texts**.



3. Select and drag files from the folder your created into the upload box.

- 4. Wait until your texts have been processed. You can add more texts or folders in the meantime.
- 5. Click **Compile** and shortly your corpus should be ready to analyse.

Compiled
The corpus is compiled and ready to use.
ADD MORE TEXTS RECOMPILE CORPUS DASHBOARD
Expert settings 👻 Log 👻
GET TO KNOW YOUR CORPUS!
EXTRACT KEYWORDS & TERMS



### **Corpus Analysis**

The main way language teachers have their first encounter with corpus linguistics is through dictionaries or vocabulary lists in coursebooks. While dictionaries and word lists both have their uses, most coursebooks contain very little in the way of real-life tasks. One way that we can bring authentic language into authentic situations for our learners is by analysing corpora. In this section, we'll use SketchEngine first, then some free software called AntConc made by Laurence Anthony, an academic in Japan who also teaches English for Specific Purposes. Finally, we'll take a quick look at the Brigham Young University corpus interface (BYU Online corpora).

Some output we are likely to want from the corpora we are investigating:

- Concordances samples of your corpus, with your search term in the middle. Your concordancer usually shows between 10-20 lines of examples. You won't usually get full sentences but you can click to get the full context. The concordance functions are extremely similar in SketchEngine, BYU Online Corpus and AntConc. You can search for single words, phrases and you can also search for parts of words using wild card functions like \* and ?. The wildcard functions let you leave a gap so you can search for grammatical patterns in AntConc. Additionally, if you have programming experience, you can also search using Regex, which can save you time, but is not essential for our purposes.
- Word list the words in the corpus ordered by frequency;
- **Keyword list** the words in the corpus that occur at a statistically higher frequency than those in a reference corpus. This is often referred to as 'keyness' which is positive for higher frequency words, and negative for words that appear statistically more frequently in the reference corpus.
- **N-Grams** these are essentially chunks of language ordered by frequency. It's usually a good idea to check the frequency of N-Grams against high frequency words because some of your N-Grams are likely to be more high frequency than your key words. The name comes from the length being N words long, or the user chooses how long they are. Mostly you'll choose up to five words long.
- **Collocates** these are words that appear within a certain distance of your search term. The default settings in most tools are 5L 5R, which simply means within five words to the left and five words to the right.
- **Clusters** these are lists of words that are frequently to the left or to the right of your search term.



#### SketchEngine

In the sections above, you made a corpus either by scraping the web or uploading your own folder of texts. You're ready to analyse that corpus now. It's a good idea to generate your word list first (the most frequently occuring words in your corpus). This will give you an idea of how frequent your search terms are (in the case of a scraped corpus) and if any other specific terms are high in the list. It might mean going back a stage and remaking your corpus (and possibly deleting the one you just made due to lack of storage space).

ø	WORD	LIST FoodNutrition	nCorpus Q (j
	BASIC	ADVANCED	
		<u>^</u>	
	find	words	all
$\odot$		lemmas	starting with
0		nouns	ending with
•-		verbs	containing
=		adjectives	
•••		adverbs	
=•= =•=		(C)	
ţ≡			GO
NE			
≣گ			
1			

Your **keyword list** is also incredibly useful. It is a list of the most probable words to occur in your corpus as compared to a reference corpus. The reference corpus is just another corpus, usually much larger, that you compare your own corpus to. The default reference corpus in SketchEngine is their own EnTenTen15 web corpus (which is around a billion words in size). You can change this to other corpora such as BNC15 (British National Corpus, 2015) and COCA (Corpus of Contemporary American English). For specialised corpora, such as healthcare, it gives you an idea of how much jargon you might expect in typical written communication, and therefore how much of this would need to be known for spoken communication.

Alongside the keyword list, the new version of SketchEngine will also generate a list of key **N-Grams** or **terms**. There's more on this below. It's worth noting that all lists on SketchEngine can be downloaded in PDF, CSV, XLS and XML formats.



You can also generate a separate **N-Gram list**. If you choose your N-Gram to be 3 or 4 words long (the default) you get three-to-four word chunks ranked by frequency in the corpus. You can also specify the minimum range (how many texts each N-Gram should appear in). Don't be afraid to fiddle with these settings until you get something you can work with. It's really useful for finding idioms and discourse markers, and especially for pragmatic functions like hedging.



Most frequent N-Grams in spoken academic advice sessions: from R. C. Simpson, S. L. Briggs, J. Ovens, and J. M. Swales. (1999) <u>The Michigan Corpus of Academic Spoken</u> <u>English</u>. Ann Arbor, MI: The Regents of the University of Michigan

Cooperativa de

Serveis Lingüístics de Barcelona



Finally, the **Word Sketch** function is the major advantage of SketchEngine. You can type in a lemma (root form of a word. e.g. 'ride' is a lemma, 'rides', 'riding' and 'rider' are not lemmas). It will give you a comprehensive display of how the word collocates with different words and different parts of speech. It's a good idea to run the Word Sketch on at least the top 20 or so keywords. This should give you an idea of any obvious 'marked' (somewhat unusual) language use, such as nominal modification (nouns used to modify nouns as opposed to adjectival modification), which is often a feature of technical/scientific language.

<b>S</b>	WORD SKET	CH FoodNutritionCor	rpus Q (j		٢	eð (?) 🖪 🚦
	diet as noun 57× ···				ঽ	生 💿 🌼 🛈 🕁
	.≓	.≓	.≓ ™ ¤ ×	.≓ ≣•≣ ⊠ ×	+7 ₩ Ø ×	+→ 1•1 10 ×
	modifiers of "diet"	nouns modified by "diet"	verbs with "diet" as object	verbs with "diet" as subject	"diet" and/or	prepositional phrases
<ul> <li>○</li> </ul>	paleo ···	plan •••	revamp	consist ···	moment	on "diet" •••
õ	the Paleo diet	diet plans	revamp your diet	The raw food diet consists	diet diet	of "diet" •••
•≡	Low-Carb Diet	shake ···· diet shake	adopt ···· adopting a plant-rich diet	Diet Live	fun •••	to "diet" ••••
•	mediterranean ···· The Mediterranean Diet	requirement ····	tell told the Paleo diet	fit ···· diet fits	low-carb	in "diet" •••
	balanced	requirements	process	follow	low-carb , high-protein diet	than "diet"
	balanced diet	philosophy ···	processed diet	diet follows	whole-food	from "diet"
	raw	philosophy	build	do	whole-foods , plant-based diet	"diet" for •••
NE	The raw food diet consists	food	diet is built	Diets do	need	for "diot"
δ≣	special ····	Diet Anti-inflammatory foods	follow	be ··· Paleo diet is	diet and specific nutritional needs	"diet" with •••
10	food ····		base ····	contain	requirement ····	about "diet" ••• Back to the original interface

#### AntConc

This is another way to analyse corpora you have compiled yourself, and is free. However, it involves more work on your part.

 Go to Laurence Anthony's webpage <u>http://www.laurenceanthony.net/software.html</u>. Download **AntConc** if you haven't already. You might also download **TagAnt** and the tag list, too. You should also download some useful word lists, such as the BNC British English word list and BNC American English word list in AntConc format from Paul Baker's website: <u>https://www.lancaster.ac.uk/linguistics/about-us/people/paul-baker</u>. You'll need to scroll down guite far. It's probably more useful to have these word lists on your

desktop than in your downloads folder so move them before you start.

2. If you are using a corpus you made in BootCaT, it is not tagged yet. If you are using a corpus that you know is tagged, skip to step 3.

Open TagAnt. Your corpus you made is in a directory that is usually at My Documents/BootCaT Corpora/Name of your corpus project/ so go to File > Open Directory and choose the directory your corpus is in. Load and start. It shouldn't take much longer than a couple of minutes.



Relieven Koloo Finatox	lie.	and thin iteration and iteration		łaz	Second What Hater
Dividitory		🐥 TagAnt 1.2.0			- 🗆 ×
		File Help			
Open Target Directory	×	<ul> <li>Input Text</li> </ul>	Clear	Results	0%
→ · ↑ ■ · PC · デスクトップ · ひ デスクトップの検索	م				
整理 マ 新しいフォルダー					
ConeDrive         A         Gat         Marc Jones         Dr         Dr         Dr	更新日時 2017/09/26 6:01 2019/01/31 5:37 2018/03/14 5:59 2019/01/30 13:53 2018/07/28 0:15 2017/08/18 9:15	0 input files: 0 Load	Clear		
フォルダー:    フォルダーの選択	キャンセル	language English V St	art Stor	Horizontal O Vertical	Clear

- 3. AntConc runs as an executable file which means you can run it from a USB flash drive or your computer. Open the file and tell your computer that you really do want to run the file.
- 4. Go to File > Open Directory and choose the directory your corpus is in.
- 5. Next, open Global Settings. We'll start with tags hidden. Click Apply after you change this setting.





6. Go to **Word List**. Click start. You have your word list and you can export it as a text file by clicking File > Save Output As.

Open File(s)	Ctrl+F	pus Files	Conc	ordance Co	ncordance Plot File View Clusters/N-Gran	ns Collocates Word List Keyword List	
Open Dir	Ctrl+D	tagged.txt	Word	Types: 752	2 Word Tokens: 98797 See	arch Hits: 0	
Close Selected File(s) Close All Files		tagged.txt	Rank 1	5233	and L	emma Word Form(s)	í
lear Tool		_tagged.txt	2	2861	the		
lear All Tools		_tagged.txt	3	2509	of		
lear All Tools and Files		_tagged.txt	4	2473	to		
ave Output to Text File	Ctrl+S	_tagged.txt	5	2303	a		
nport Settings from File.		_tagged.txt	6	1449	in		
port Settings To File		_tagged.txt	7	1243	you		
estore Default Settings		_tagged.txt	8	1215	is		
iit		_tagged.txt	9	1183	for		
-	0	16_tagged.txt	10	946	it		
O Press t	he r	7_tagged.txt	11	945	that		
3 toko o c		8_tagged.txt	12	927	are		
lake a s	o	20_tagged.txt	13	883	or		
brietty.	0	21_tagged.txt	14	843	vour	1	
Your	scree 0	23_tagged.txt	Search	Term 🖂 V	Vords Case Regex	Hit Location	
is mo	st col 0	24_tagged.txt			Advanced	Search Only 0 💠	
<ul> <li>If you</li> </ul>	IT SCRE TO	tal No.	S	tart	Stop Sort	Lemma List Loaded	
Ctrl	⊞ W 10	19	Sort b	v 🗌 Invert	Order		

7. To generate your keywords list you need to load a word list or a reference corpus (see part 1 of this section), then click apply. You probably don't have a whole reference corpus but you have word lists. Go to **Tool Preferences**, choose your word list, load it and then click **Apply**.

Tool Preferences	— X	-	
Category	Keyword List Preferences		
Concordance Cluster/N-Grans Collocates Word List Keyword List	Display Options            Rank          Frequency         Keyness         Keynoss         Other Options         Tretal all data as lowercase         Tretal all data as lowercase         Remose Values         Keynovard Generation Method         Log-Likelihood         Xell         Xell	ate: Word List Keyword Lis 0 ard Form(s)	A
	ândiv Carrel	a List Loaded	> •



Tool Preferences	- [	×
Tool Preferences Category Concordance Clusters/N-Grams Collocates Word List Keyword List	Keyword List Preferences         Display Options         Stank       Frequency         Image: Stank       Frequency         Trest as in sort         Keyword Generation Method       Cog-Likelihood         Threshold Value       All Values         Show negative keywords (using highlight color)         Reference Corpus       Oue word lists)         Loade       Load         Total No. 1         And Directory       Add Files       Swap with Target Files	x x - C x

8. Generate your keyword list. Save your output again.

Corpus Files		Conco	rdance Co	ncordance Plot Fi	le View Clusters/N-Grams Collocates Word List Keyword List	
000_tagged.txt 001_tagged.txt		Types B Bank	lefore Cut:	7522 Typ	es After Cut: 5206 Search Hits: 0	,
002_tagged.txt		1	510	2376.567	diet	- ii
04 tagged bt			401	2120.200	fande	
05 tagged.txt		2	481	2138.398	loous	
06_tagged.txt		3	440	1803.111	protein	
07_tagged.txt		4	401	1698.082	fat	
08_tagged.txt		5	843	1589 944	vour	
09_tagged.txt		6	245	1551 252	maal	
10_tagged.txt		0	545	1331.332	Illeal	
12 tagged by		7	497	1528.918	tood	
13 tagged.txt		8	399	1484.886	weight	
14_tagged.txt		9	319	1466.371	calories	
15_tagged.txt		10	249	1202 269	oat	
16_tagged.txt		10	340	1392.200	cal	
17_tagged.txt		11	330	1363.078	healthy	
18_tagged.txt		12	245	1093.524	fiber	
20 tagged by		13	359	1052,907	a	
21 tagged.txt		14	217	1005 577	nutrition	
22_tagged.txt		<	> < >	<	> <	> .
23_tagged.txt		Search	Term 🔽 V	Vords Case	Regex Hit Location	
024_tagged.txt	~				Advanced Search Only 0	
otal No.		Sta	art	Stop	ort Reference Corpus V Loaded	
09		Carthe		Orden		

If you are using a corpus which is just one text file (and some corpora are set up like this) set the minimum range (number of text files in the corpus) to 1 in the following steps, otherwise you will have no output.

9. You can also generate N-Grams and clusters. N-Grams are easiest. Choose the minimum and maximum size of N-Gram and minimum range you want each one to occur in. Again, do not be afraid to fiddle with things here. When you are ready you can save your output again.

AntConc 3.4.4w (V e Global Settings	Windo Too	ws) 2014 Preferer	ices Help						_	>
orpus Files		Conco	rdance Co	ncordance	Plot File View	Clusters/N-Grams	Collocates	Word List	Keyword List	
00_tagged.txt 01_tagged.txt	^	Total N Rank	o. of N-Gra	am Types Range	19568 N-gram	Total No. of N-Gran	n Tokens	81269	Reyword List	
02_tagged.txt		1	253	71	of the					
04 tagged by			240	0	- A					
05 tagged.txt		2	248	60	nı					
06_tagged.txt		3	240	68	in the					
07_tagged.txt		4	213	53	it s					
08_tagged.txt		5	210	59	if you					
09_tagged.txt		6	100	24	in you	-				
10_tagged.txt		0	188	34	weight los	S				
12 tagged bt		7	173	54	is a					
13 tagged.txt		8	145	55	on the					
14_tagged.txt		9	143	44	to eat					
15_tagged.txt		10	127	14	moal repla	comont				
16_tagged.txt		10	157	14	mearrepla	cement				
17_tagged.txt		11	136	47	such as					
18_tagged.txt		12	134	51	you can					
19_tagged.txt		13	130	30	high in					
20_lagged.txt		<	> < 3	> < 3	< C					>
22 tagged.txt		Search	Term 🖂 🛛	Words	Case Rege	x 🔽 N-Grams	N-Gram Siz	e		
23_tagged.txt						Advanced	Min 2	A Max 5		
24_tagged.txt	~	1				Autoriced	-	• •	•	
		Sta	art	Stop	Sort		Min. Freq.	Min. Rang	e	
otal No.		Sort by	Inver	t Order	Search Term Po	osition	2	2		
Jy Inc Processed		Sort by	Freq	~	🗹 On Left [	On Right				

10. Clusters are groups of words with your chosen word at either the left of the cluster or the right. Check the cluster box, set your size and range (quite low is usually better). Type a word then press Start. You can save these to text files in exactly the same way. The same applies regarding minimum range here.

le Global Settings	Тоо	l Preferenc	es Help	,			~
Corpus Files		Concorr	dance Co	ncordance	Plot File View Clusters/N-Grams Collocates Word List Keyword List		
000_tagged.txt 001_tagged.txt	^	Total No Rank	of Cluste Freq	er Types Range	1380 Total No. of Cluster Tokens 2023 Cluster		^
003 tagged.txt		1	64	31	your diet		
004_tagged.txt		2	33	18	healthy diet		
005_tagged.txt		2	20	17	a healthy dist		
006_tagged.txt		3	29	17	a healthy diet		
07_tagged.txt		4	23	6	of diet		
008_tagged.txt		5	21	18	a diet		
10 tagged.txt		6	19	12	balanced diet		
11 tagged.txt		7	10	14			
12 tagged.txt		1	19	14	in your diet		
13_tagged.txt		8	17	12	to your diet		
)14_tagged.txt		9	16	1	healing diet		
)15_tagged.txt		10	14	3	food diet		
16_tagged.txt		1.0	1.4	10			
1/_tagged.txt		11	14	12	the diet		
18_lagged.txt		12	13	1	loss diet		
20 tagged.txt		13	13	2	raw food diet		
21 tagged.txt		< >	< :	× >	<	>	
22_tagged.txt		Search T	erm 🗹 🕯	Nords 🗌	Case Regex N-Grams Cluster Size		
23_tagged.txt		diet			Advanced Min. 2 + Max. 5 +		
24_tagged.txt	~			<b>C</b> 1			
otal No.		Star	τ	Stop	Sort Min. Freq. Min. Kange		
09		Sort by		Order	Search Term Position		
les Processed		Sort by F	req	~	On Left		

- 11. **Collocates** in AntConc generates words that appear within different proximities to your search term. The default search is 5L,5R, or five words to the left and to the right of your search term. You get to see if the collocating words occur to the right, the left and their frequencies. You can also save this as a text file.
- © Marc Jones for Serveis Lingüístics de Barcelona (SLB) SCCL, 2019

Cooperativa de

Serveis Lingüístics de Barcelona



orpus Files		Cancar	dance Ce	ncordonce	Diet File	View Chusters	(N Grame Collocates Ward List Kenward List	
00_tagged.txt	^	Total No		rate Types	1222	Total No	of Collocate Tokens: 5066	
01_tagged.txt		Rank	Freq	Freq(I)	Freq(R)	Stat	Collocate	
02_tagged.txt		1	255	101	6A	4 4 2 2 0 0		1
03_tagged.txt			235	191	04	4.42200	a	
04_tagged.txt		2	157	88	69	3.62040	to	
05_lagged.txt		3	155	93	62	3.39165	the	
07 tagged.txt		4	133	81	52	3,36021	of	
08_tagged.txt		2	114	40	CC	2 77200		
09_tagged.txt		S	114	48	00	2.77206	and	
10_tagged.txt		6	95	79	16	4.44829	your	
11_tagged.txt		7	95	24	71	3.92094	is	
12_tagged.txt		0	07	6	01	7 20020	soda	
13_tagged.txt		•	0/	0	01	1.29950	soua	
14_tagged.txt		9	86	47	39	3.52325	in	
15_tagged.txt		10	74	64	10	4.62240	on	
15_tagged.txt		11	72	21	52	3 50804	VOU	
18 tagged txt			15	21	52	3.30004	you	
19 tagged.txt		12	64	30	34	3.38959	for	
20 tagged.txt		13	50	20	30	3.35751	that	
21_tagged.txt		14	50	42	8	4 87536	healthy	
22_tagged.txt		< 3	< :	) ( )	< >	< >	<	×
23_tagged.txt		Search 1	erm 🗹 ۱	Nords 🗌	Case 🗌	Regex	Window Span Same	
24_tagged.txt	~	diet				Adv	anced From 5L 🗘 To 5R 🜲	
otal No.		( Sta	+	Stop	Sor	+	Min Collocate Frequency	
09		L		oup	501	<u>.</u>	The second care requercy	

12. To search for grammar patterns, you can go to Global Settings and choose show tags. Cluster, collocates and concordance tools are the most useful for syntax (word order and pattern) investigations. You might want to see if a given noun is more usually preceded by adjectives or nouns. You can click to see examples of these. Tree Tagger tags (the tags used in TagAnt) for simple adjectives are '\_jj' and for singular nouns '\_nn'. To search for [infinite verb + noun] patterns, search "# vvp # n\*". # and \* are wildcard characters. # means 'any one word' and \* means 'one or more characters'. This means we search for ['any one word' 'vvp' 'any one word' 'any noun tag'] because all noun tags start with n and are two or three characters long. It may be useful to search this in the cluster tool on the left as well as the concordance tool.

AntConc 3.4.4w (W	Vindo	ows) 2014					×
Corpus Files 000_tagged.txt 001_tagged.txt	^	Concore Total No Rank	ance Co of Cluste	ncordance r Types 9 Range	Plot File View Clusters/N-Grams Collocates Word List Keyword List Total No. of Cluster Tokens 20 Cluster		~
002_tagged.bt 003_tagged.bt 005_tagged.bt 006_tagged.bt 006_tagged.bt 006_tagged.bt 009_tagged.bt 010_tagged.bt 010_tagged.bt 011_tagged.bt 011_tagged.bt 013_tagged.bt 013_tagged.bt 014_tagged.bt 016_tagged.bt 016_tagged.bt 016_tagged.bt		1 2 3 4 5 6 7 8 9	4 2 2 2 2 2 2 2 2 2 2	4 2 2 2 2 2 2 2 2 2 2	contain_vvp caffeine_nn contain_vvp caffeine_nnsent contain_vvp lots_nns contain_vvp lots_nns of contain_vvp lots_nns of_in contain_vvp sugar_nn include_vvp soy_nn offer_vvp protein_nn provide_vvp nutrients_nns		
020_tagged.txt 021_tagged.txt 022_tagged.txt 023_tagged.txt 024_tagged.txt Total No. 109 Files Processed	>	Search T # vvp # Star Sort by Sort by F	erm V n* t Invert	Vords  Vords  Vorder	Case Regex N-Grams Cluster Size Advanced Min. 4 Max. 7  Sort Min. Freq. Min. Range Search Term Position 2  On Left On Right Cle	one Re	sults



#### **BYU Online Corpora**

There are several different corpora available at https://corpus.byu.edu/. These are handy for more general applications, as well as for getting used to how corpora work. If you need to analyse 'general' written English, the iWeb or GloWBe corpora may be useful. If you have certain localities in mind, like the US, COCA could be useful and for Canada, there is the Strathy corpus. The Corpus of American Soap Operas could also be handy for generalised spoken language.

	corpus.byu.edu								
home corpora	users related	resources my accou	nt upgrade he	lp					
reated by Mark Davies, BYU (Google Scholar). <b>Overview</b> , so 'he most widely used online corpora – more than 130,000 <b>English</b>	earch types, looking at distinct researchers, te <b># words</b>	variation, corpus-based r eachers, and students each	esources. n month. time period	compare					
iWeb: The Intelligent Web-based Corpus	14 billion	US/CA/UK/IE/AU/NZ	2017	Info(中文)					
News on the Web (NOW)	1.5 billion+	20 countries / Web	2010-last month						
Global Web-Based English (GloWbE)	1.9 billion	20 countries / Web	2012-13						
Wikipedia Corpus	1.9 billion	English	2014	Info					
Hansard Corpus	1.6 billion	British (parliament)	1803-2005	Info					
Early English Books Online	755 million	British	1470s-1690s						
Corpus of Contemporary American English (COCA)	560 million	American	1990-2017	****					
Corpus of Historical American English (COHA)	400 million	American	1810-2009	**					
Corpus of US Supreme Court Opinions	130 million	American (law)	1790s-present						
TIME Magazine Corpus	100 million	American	1923-2006						
Corpus of American Soap Operas	100 million	American	2001-2012	*					
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993	* *					
Strathy Corpus (Canada)	50 million	Canadian	19705-20005						

¡You can search for words and collocates, much like in SketchEngine or AntConc above. Unfortunately, there is no N-Gram search in the BYU Online corpus interface. You may want to look at the KWIC concordances.

i	Web: The 14 I	Bill	lic	on	Word Web Corpus  📴 🔕	③ ■ (	<b>⅔</b> ?	🖪 ★ 🗉 🕚				
	SEARCH				FREQUENCY		CONTEXT	OVERVIEW				
FIND	) SAMPLE: <u>100 200 500 101</u> E: << < 1/468 > >>	00						ŕ				
CLIC	CK FOR MORE CONTEXT					[7]		SHOW DUPLICATES				
1	scrum.org	Α	в	с	Scrum Product Owner II Prove Your Advanced Knowledge of	Product Ownership	# The Professional Scrum Produ	ict OwnerG level II (PSPO II) assessment is available t				
2	scrum.org	A	в	с	the PSPO I assessment and wishes to demonstrate his or he	ability to apply the	Scrum framework to solving con	nplex problems regarding product ownership- in the				
3	scrum.org	Α	в	c	recognized PSPO II Certification as an indication of their advanced knowledge and abilities pertaining to Scrum and the role of the Scrum Product Owner. # Anyone a							
4	scrum.org	A	в	с	indication of their advanced knowledge and abilities pertaining to Scrum and the role of the Scrum Product Owner. # Anyone attempting the PSPO II should have in-							
5	scrum.org	Α	в	с	of the Scrum Product Owner. # Anyone attempting the PSPO II should have in-depth Scrum and Product Owner experience and have taken the-Professional Scrum F							
6	scrum.org	A	в	с	PSPO II should have in-depth Scrum and Product Owner experience and have taken the-Professional Scrum Product Owner course- prior to taking this assessment.							
7	scrum.org	Α	в	c	Prepare for the Assessment # The PSPO- assessments are- grounded in the- Professional Scrum Product Owner- subject areas. The assessment is exhaustive and ri							
8	scrum.org	A	в	с	rigorous. Many questions ask you- to think about or interpret the meaning from the- Scrum Guide, - apply content from the PSPO- subject areas, and in some cases							
9	scrum.org	А	в	с	additional information to help you prepare for the PSPO- Assessment and ways to learn about Scrum Certification # If you pass the PSPO II assessment you will rec							
10	scrum.org	A	в	с	In addition, your name will be listed on Scrum.org # Unlike other Scrum certifications that require only class attendance, Scrum.org certification requires a minimu							
11	scrum.org	A	в	с	, and they are very quickly losing respect for the PO. # As a Scrum Master, I'm struggling to rein him in. I've mentioned their tendency							
12	scrum.org	A	в	с	Technical Product Owner' which might be a role that can be incorporated in SCRUM alongside the Business Product Owner. Having technical AND business knowled							
13	scrum.org	Α	в	с	then it's time to find out what his agenda really is # In Scrum PO ought to do a mental trade-off: less predictable, but release quality outcomes							
14	scrum.org	A	в	с	not complex (it might be complicated, though), therefore benefits coming from Scrum implementation are limited (is PO able to formulate a Sprint Goal?) -							
15	scrum.org	A	в	с	their job properly. # Try focusing more on the actual structu	e and rigor of Scrur	n Events. For example: - a Scrum	Master should ensure that only the Development				

You can, however, gain an excellent overview on the main page of the iWeb corpus with some main clusters and collocates.

© Marc Jones for Serveis Lingüístics de Barcelona (SLB) SCCL, 2019



iWe	b: 1	he 14 Billio	n Wo	rd V	/eb	Corpus 🕞		<b>(i)</b>		C ?				<u> </u>	=	9
		SEARCH				COLLOCATES				CONTEXT				OVERVIEW		
COLLOCATES SCRUM NOUN Advanced options 🛞 Collocates Clusters Topics Dictionary Websites KWIC 📀 🛓																
+ NOU	N	NEW WORD	?	+ ADJ		NEW WORD	?	+ VE	RB	NEW WORD	?	+ A	DV	NEW WORD	?	
5332	4.40	team		3191	9.53	agile		316	3.23	implement		67	4.09	half		
4588	6.53	master		1050	6.40	certified		217	3.34	adopt		29	3.58	forwards		
1640	3.30	project		995	4.55	daily		187				26	2.51	backwards		
1448	2.50	product		184	4.81	dominant		165			1	16	2.54	front		
1021	3.04	development		155	4.74	lean		165				12	6.10	offside	E	
923	3.82	owner		140	6.80	attacking		153			1	5	2.71	superbly		
923	5.89	framework	E	134	2.91	extreme		124	4.71	collapse		4	2.52	purposefully		
889	7.18	methodology	E	122	2.64	Irish		118	2.68	award		4	2.59	optimally		
877	4.10	meeting	E	110	4.21	resulting		110	2.92	opt		4	3.04	comprehensively	E	
848	3.24	role		91	8.56	uncontested	E	103	3.76	certify	圓	4	3.05	deceptively		
770	6.16	alliance	E	84	7.27	iterative		102	3.72	lean	8	3	2.58	interchangeably	E	
768	5.50	penalty	E	72	6.03	ensuing		96	3.01	coach	8	3	2.68	hotly	E	
758	4.76	half		72	8.97	kanban		87	3.02	practice	E	3	2.72	inexplicably	E	
666	3.67	guide		71	2.69	defensive		68	3.12	master		3	4.15	uncharacteristically	E	
656	2.73	media		69	5.78	halfway		61	4.54	concede	140	3	6.47	diversely	E	
547	7.01	sprint		59	5.64	retrospective		56	2.65	reset	and a	2	2.67	ruthlessly	1	
539	3.80	coach	E	58	4.81	empirical	E	50	2.69	collaborate		2	2.86	tactically		
402	2.97	ball	E	53	4.30	Welsh		35	4.50	ensue	E	2	3.09	headfirst	E	

Veb: The 14	Billion Wor	d Web Corpus		) (ì [	C 🖓				
SEARCH WORD					CONTEXT	OVERVIEW			
SCRUM (NOU	JN) 😭 🚯	#10333	TOPICS	more)	Collocates Clusters Topic	s Dictionary Websites KWIC 🞍 HELP			
1. (rugby football) th forwards of each tea M O C G	ne method of beginning am crouch side by side	g play in which the with locked arms D	agile, team, try, penalty, bail, sprint, half, score, project, rugby, software, development, backlog, defence, match, kick, developer, kick, minute, master						
PlayPhras	e YouGlish Yarn		COLLOCATES (more)						
Translate:	choose language		NOUN team, master, project, product, development, framework, owner, methodology						
			VERB	mplement, ad	opt, scale, facilitate, sprint, do	ominate, collapse, award			
SYNONYMS (more)			ADJ	DJ agile, certified, daily, dominant, lean, attacking, extreme, irish					
crowd struggle, fi jostle	ray, scrimmage, free-	for-all, tussle, scrum,	ADV	half, forwards, backwards, front, offside, superbly, purposefully, optimally					
NOUN + NOUN (mo	re)								
scrum NOUN	scrum master • se	rum team « scrum teams «	scrum alliar	ce 🛛 scrum ma	sters • scrum half •				
NOUN scrum	media scrum -								



### Analysing corpora for TBLT

Before you start analysing the corpus, you need to bear in mind the task for which you are searching for related discourse. In this example, let's take the corpus we made using BootCaT and think of a task related to English for nutrition students, *Give general advice regarding diet.* 

First it would be useful to select some words to search in the collocates and clusters tools in AntConc or in SketchEngine. These words can be based upon our own intuitions regarding the keywords, e.g. diet, eat, healthy, you\* (equals you/your). Among other items were:

for a healthy	plans	fewer
of a healthy	keeps	smaller
(help) maintain a	(may/can/to) help you	foodists
healthy	good for you	choosing
as part of a healthy	boost your	regularly
healthy diet	increase your	what you eat
healthy eating	you lose weight	foods to eat
healthy fats	your metabolism	foods you eat
healthy weight	your diet	that people who eat
heart	your body	feel free to eat
choices	you can	the foods you eat
living	you to	eat fewer calories
maintaining	you should	eat plenty of

We then can look at the words and patterns that are found with our search terms. We then have more to look at and can search our corpus concordances for these and they will give use some ideas of written discourse which we shall have to evaluate as to whether it is appropriate for spoken output in our prototypical task.

What comes back to add to the above is:

Eat plenty of fruits and vegetables Eat plenty of fruit, vegetables and wholegrains Eat plenty of soy and bean products As part of a healthy diet Help you maintain Maintain a healthy weight



Maintain muscle The (comparative adjective) food you eat, the (comparative adjective) To boost your metabolism Foods in your diet People eat fewer calories (preposition) You can also/easily/get

We now have a large amount of authentic language to help create a prototype task featuring these main patterns. The task could look something like the below, although bear in mind we would need to take other factors into consideration apart from the key vocabulary and collocations, i.e. the modality of the task (spoken or written), the audience, and the typical structure (e.g. turns, moves etc.). See the Session 6 presentation for more on this.

#### Prototype Task: Give general advice regarding diet.

As part of a healthy diet, you should eat less fewer salty foods and drink less caffeine. The more salt and caffeine you consume, the faster your heart beats and in our current environment, a lot of people have racing hearts. It's also important to eat plenty of fruit, vegetables and wholegrains to maintain energy throughout the day while also reducing the amount of heavier carbohydrates you eat, such as bread, rice, pasta and potatoes. The more vegetables you eat, the more that these count toward your carbohydrate intake. This will help you lose weight if necessary, while also maintaining enough calories to fuel you for the day. You shouldn't cut out fats and heavy carbohydrates completely: feel free to eat these in moderation. The important point is to reduce the foods that are unhealthy when eaten too much and to increase foods that will boost your metabolism and maintain your health.