

Creating Corpus-Based Materials

Sarah Deutchman
Waseda University

Agenda

- Overview of what corpora are
- How the worksheets were created
- Participants get a chance to do corpus searches

What are corpora?

- “A corpus is a large, principled collection of naturally occurring examples of language stored electronically” (Bennett, 2010, p. 2).
- It is possible to create a corpus for anything
 - The English corpora website added a corona virus corpus
- Electronic corpora are easily searchable
- Cannot analyze themselves

How can corpora be used?

- Corpora can help students better understand how to correctly use the language:
 - To check collocates
 - To check for connotations
 - Grammatical patterns
 - Where and when a word is used
 - Infer meaning of a new word from concordance lines
 - Create specific word lists for ESP

Studies Using Corpora as learning tools

- Cobb (1997)
 - Met words through computer game activities either through concordances or definition-based entries
 - concordance learning showed less decay vs. definition-based learners
- Boulton (2009)
 - Learning styles did not significantly affect ability to work with corpora
 - Visual learners slightly more receptive

Studies Using Corpora as learning tools

- Boulton (2010)
 - Learners given paper-based corpus materials and others given dictionary entries.
 - Lower-level students did better with corpora than with standard dictionary entries
- Cobb and Boulton (2015)
 - Meta-analysis showed that corpora can benefit L2 learners receptive and productive skills (e.g., extensive reading and writing tasks) for collocations and idioms

Word Profilers

- Can be used to input text and see the frequency of each word in a text against corpus-based frequency lists.
- Can check how complex your text is
- Vocab Profiler
 - <https://www.lex tutor.ca/vp/>
- AntWordProfiler
 - <https://www.laurenceanthony.net/software/antwordprofiler/>
- Analyze Text
 - <https://www.english-corpora.org/coca/>

Considerations when designing worksheets

- It is necessary for learners to test their hypotheses based on their expectations and refine those hypotheses when presented with evidence (Johns, 1988).
- For there to be benefits of students looking at corpora a long period of practice with scaffolding is necessary (Cobb & Boulton, 2015)
- Purpose of the worksheet
- How advanced are your students' skills

Choosing your corpora

- Sketch Engine for Language Learning
 - Concordance lines easier than Coca
 - Simpler interface
 - Login not required
- Corpus of Contemporary American English
 - More detailed information
 - Can compare with different corpora
 - Better for lower frequency words

COCA Word

See in iWeb Collocates Clusters Topics Dictionary Texts KWIC HELP

leader (NOUN) #539

Media Type	Frequency
BLOG	Low
WEB	Low
TV/M	Very Low
SPOK	Low
FIC	Very Low
MAG	Low
NEWS	High
ACAD	Low

1. a person who rules or guides or inspires others 2. a featured article of merchandise sold at a loss in order to draw customers

D M O C G **E**

YouGlish PlayPhrase Yarn

JA: Google WordRef Reverso Linguee

SYNONYMS (more)

frontrunner **lead**, **leader**, trailblazer **guide** director, guide, guru, leader, mentor, organizer **head** chief, head, kingpin, leader, manager, principal, superior, supervisor

TOPICS (more)

[leadership](#), [minister](#), [senate](#), [prime](#), [democratic](#), [republican](#), [democrat](#), [majority](#), [election](#), [reform](#), [opposition](#), [official](#), [party](#), [meeting](#), [congress](#), [diplomat](#), [peace](#), [minority](#), [vote](#), [military](#)

COLLOCATES (more)

NOUN [majority](#), [senate](#), [party](#), [business](#), [community](#), [republican](#), [minority](#), [church](#)

VERB [elect](#), [urge](#), [gop](#), [assassinate](#), [laden](#), [vow](#), [denounce](#), [overthrow](#)

ADJ [political](#), [religious](#), [democratic](#), [military](#), [local](#), [congressional](#), [civil](#), [palestinian](#)

ADV [democratically](#), [eg](#), [up-country](#), [popularly](#)

[Click here to continue with your search](#)

[简体中文](#) [中文](#) [한국어](#) [Русский](#) [العربية](#)

There is no cost for basic access to English-Corpora.org. But you will occasionally see this message, which asks you to consider upgrading to a "premium" account.

If you have a [premium account](#) (\$10.00 per month, \$30.00 for one year, \$55.00 for two years, or \$75.00 for three years), then you will not see this message anymore. You will also have increased access to the corpora, more features, and you will help to support English-Corpora.org.

If you are a student, your school or university can also get an [academic license](#), and all of you will have the benefits of a premium account.

Thanks for your support.

How I made the worksheet

- 1) Needs analysis
- 2) Do searches ahead of time
- 3) Take screen shots
- 4) Create specific instructions on how to do the searches

Sketch Engine for Language Learning

<https://skell.sketchengine.eu/#home>

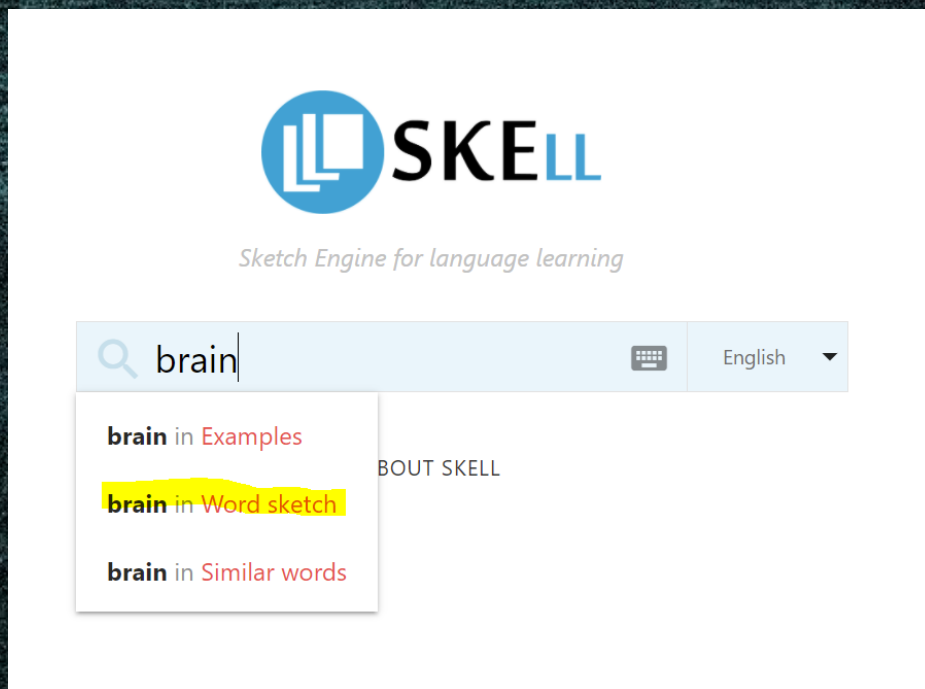
Collocations

- Words neighbors
 - Take (a shower, a bath, medicine, a walk)
- How to search for them
 - Grammar
 - _____ (adj) mind (n.)
 - Walk (v.) _____ (modifier/ adverb)
 - Pretty(adj.) _____ (noun)

adjectives with **mind**

1. sharp mind sharp
2. pure mind is pure
3. busy mind busy
4. full mind full of
5. open mind open
6. active mind active
7. quiet the mind is quiet
8. empty mind is empty
9. free mind free
10. capable mind is capable
11. clear mind clear

How to do a collocates search



- 1) Search for the word and select word sketch
- 2) Analyze the results

Connotations

- Feeling associated with the word
 - Is it positive or negative
- Cause
 - Is it positive or negative?

	object of cause	
1.	<u>damage</u>	damage caused
2.	<u>problem</u>	cause problems
3.	<u>death</u>	cause death
4.	<u>harm</u>	cause harm
5.	<u>loss</u>	losses caused
6.	<u>injury</u>	cause injury
7.	<u>disease</u>	disease caused
8.	<u>pain</u>	cause pain
9.	<u>controversy</u>	caused controversy
10.	<u>change</u>	cause changes

How to search for connotations

Search interface for the word "cause". The search bar contains "cause". Below the search bar are three tabs: "Examples", "Word sketch", and "Similar words". The "Word sketch" tab is currently selected.



object of cause	
1. <u>damage</u>	damage caused
2. <u>problem</u>	cause problems
3. <u>death</u>	cause death
4. <u>harm</u>	cause harm
5. <u>loss</u>	losses caused

Search interface for the word "convivial". The search bar contains "convivial". Below the search bar are three tabs: "Examples", "Word sketch", and "Similar words". The "Examples" tab is currently selected.

convivial 0.28 hits per million

1. Some evenings were more **convivial** than others.
2. The food was generous and very **convivial** .
3. The atmosphere is quite busy and **convivial** .
4. People in such communities are **convivial** and welcome interaction.
5. It is a very festive and **convivial** moment.
6. Today its atmosphere is more **convivial** , its aspirations more leisurely.
7. It's all very informal and **convivial** .
8. So lets be **convivial** , not offensive.

Can be done in COCA

CLUSTERS **CONVIVAL** **ADJ** LIMIT: Loose **Medium** Tight Collocates **Clusters** Topics Dictionary Texts KWIC  

12	convivial atmosphere	11	more convivial	2	convivial relations with	4	with a convivial	1	convivial and easy way	2	it was a convivial
7	convivial conversation	9	very convivial	2	convivial place for	3	in a convivial	1	convivial and informative day	1	he was a convivial
5	convivial evening	5	with convivial	1	convivial men who	3	for a convivial	1	convivial and open-minded experience	1	saloon with a convivial
5	convivial group	3	to convivial	1	convivial group at	2	not a convivial	1	convivial and buoyant as	1	overnighter with a convivia
4	convivial place	3	as convivial	1	convivial among men	2	at a convivial	1	convivial and cavalier abandon	1	food at a convivial
4	convivial spirit	3	so convivial	1	convivial party in	2	having a convivial	1	convivial and congratulatory in	1	dinner or a convivial


Frames in COCA

1) Type in your frame and click matching strings

List [Chart](#) [Word](#) [Browse](#) [Collocates](#) [Compare](#) [KWIC](#) -

[POS]?

[Sections](#) [Texts/Virtual](#) [Sort/Limit](#) [Options](#)

 (HIDE HELP) LOGGED IN

LIST display

Find single words like [mysterious](#), all forms of a word like [JUMP](#), words matching patterns like [*break*](#), phrases like [more * than](#) or [rough NOUN](#). You can also search by synonyms (e.g. [gorgeous](#)), and customized wordlists like [clothes](#). In each case, you see each individual matching string.

More information: [basic syntax](#), [part of speech](#), [lemmas \(forms of words\)](#), [synonyms](#), [customized word lists](#), and [combining words](#).

2) Results are based on frequency

ON CLICK: [CONTEXT](#) [TRANSLATE \(JA\)](#) [GOOGLE](#) [IMAGE](#) [PRON/VIDEO](#) [BOOK](#) (HELP)

HELP		SEE FULL LIST (SLOWER; MAY TIME OUT) [?]	FREQ	TOTAL 136,635 UNIQUE 1,375 +
1	<input type="checkbox"/>	AS GOOD AS	11674	
2	<input type="checkbox"/>	AS HIGH AS	5906	
3	<input type="checkbox"/>	AS BAD AS	5646	
4	<input type="checkbox"/>	AS SIMPLE AS	4211	
5	<input type="checkbox"/>	AS BIG AS	3750	
6	<input type="checkbox"/>	AS IMPORTANT AS	3720	
7	<input type="checkbox"/>	AS CLOSE AS	2755	
8	<input type="checkbox"/>	AS LOW AS	2349	

3) Look at the frame in context

CLICK FOR MORE CONTEXT.				<input type="checkbox"/> [?]	SAVE LIST	CHOOSE LIST	Leadership	CREATE NEW LIST	<input type="text"/>	[?]	SHOW DUPLICATES
1	2019	MAG	ESPN	A	B	C	Providence, in which Sam Hauser scored 56 combined points. # " And as good as he is off ball screens, we'll live with him trying to break				
2	2019	MAG	ESPN	A	B	C	the Big East, I didn't play against a guy like that. As good as Gerry McNamara was coming off screens, he didn't do the things that				
3	2017	NEWS	Arizona Daily Star	A	B	C	been plagued by back issues throughout his career, said his body is feeling as good as his game looks. But he didn't have to look far to know				
4	2017	MOV	...DC Super Hero Girls: Brain Drain	A	B	C	are supes gaudy. - How's it going over there? - About as good as you could expect. Oh, that bad, huh? Too scared to				
5	2016	TV	NCIS: New Orleans	A	B	C	' Cause the man's got a heck of a poker face. Almost as good as Pride's. That bill of goods you sold about Mateo's video?				
6	2015	FIC	ParisRev	A	B	C	religious affiliation, et cetera. And even though his suc cess rate was as good as any of the other top dating sites, he went to jail, and				
7	2015	MAG	Essence	A	B	C	colorful print flat iron. # 27 # Created under-water, these candles smell as good as they look! # 28 # It's a great time to try these				

Creating ESP Lists with COCA

<https://www.english-corpora.org/coca/>

1) Put in your search term in the list

List Chart Word Browse +

leader [POS]?

Find matching strings Reset

Sections Texts/Virtual Sort/Limit Options

1 NEWS:Sports
NEWS:Editorial

ACAD:Education
ACAD:History
ACAD:Geog/SocSci
ACAD:Law/PolSci
ACAD:Humanities
ACAD:Phil/Rel
ACAD:Sci/Tech

2 IGNORE

TV/MOVIES
BLOG
WEB-GENL
SPOKEN
FICTION
MAGAZINE
NEWSPAPER
ACADEMIC

2) Click on the word this search is for academic words only

ON CLICK: [CONTEXT](#) [TRANSLATE \(JA\)](#) [GOOGLE](#) [IMAGE](#) [PRON/VIDEO](#) [BOOK](#) (HELP)

HELP			ALL	BLOG	WEB-GENL	TV/MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2014	2015-2019
1	<input type="checkbox"/>	LEADER	8486								8486	1603	1576	1630	1554	1263	860

3) Save your list

PAGE: << < 1/85 > >>

CLICK FOR MORE CONTEXT				<input checked="" type="checkbox"/> [?]	SAVE LIST	CHOOSE LIST	-----	CREATE NEW LIST	Leadership	[?]	SHOW DUPLICATES
1	2019	ACAD	Business and Economic Horizons	A	B	C	the case of Gabon in 1995, and the country appears to be the productivity leader in the region. Very high values of productivity are noted also in				
2	2019	ACAD	Business and Economic Horizons	A	B	C	Namibia (number 13). Large economies, such as South Africa (regional leader in terms of economic development and industrialization) and Nige				
3	2019	ACAD	Business and Economic Horizons	A	B	C	in the region is lesser - even in the case of Gabon that was the leader in the previous categories. Apart from the South Africa, the highest value o				
4	2019	ACAD	Global Education Review	A	B	C	and socially constructed identities. # Sinead Harmey is lecturer in literacy education and national leader for reading recovery at the Internationa				
5	2019	ACAD	...rve Bank of New Zealand Bulletin	A	B	C	5822 # She was a mission leader , serving as an area director for East Asia, supervising Southern Baptist mission work				
6	2019	ACAD	Baptist History and Heritage	A	B	C	than Lottie Moon. She was primarily an evangelist but also was a respected mission leader , looked to for guidance by her colleagues and FMB le				
7	2019	ACAD	Baptist History and Heritage	A	B	C	Area Director for East Asia # When Sam James was elected as the new regional leader for Europe in 1992, Pearson performed the work of the ar				
8	2019	ACAD	European Research Studies	A	B	C	leadership, and different attitudes and behaviors of employees. Some are trust in the leader (Zeinabadi and Rastegarpour, 2010), citizen behavio				
9	2019	ACAD	European Research Studies	A	B	C	latter is generated precisely by the perception of injustice. The opposite occurs when the leader understands and offers a personalized treatme				
10	2019	ACAD	European Research Studies	A	B	C	of organizational justice represents the psychological mechanism by means of which the behavior of the leader contributes to reduce the emplo				
11	2019	ACAD	European Research Studies	A	B	C	297-313. # Fleishman, E.A. and Salter, J.A. 1963. Relation between the leader's behaviour and his empathy towards subordinates. Journal of Indu				
12	2019	ACAD	European Research Studies	A	B	C	among five ASEAN countries found co-movement among them. Hence, Malaysia has become the leader because it creates a spillover effect in a				

4) Choose the words you want to keep

RETURN TO KWIC PAGE: [SAME](#) [NEXT](#) [\[?\]](#) USAGE: 24 HOURS STORED

Leadership [\[HELP\]](#)

<input type="checkbox"/> [?] # <input type="text"/> - <input type="text"/> [?]	<input type="radio"/> DELETE ENTRIES	<input type="radio"/> MOVE ENTRIES	<input type="radio"/> EXPAND ENTRIES [?]
1 <input type="checkbox"/>	COCA:2019:ACAD Business and Economic Horizons	the case of Gabon in 1995, and the country appears to be the productivity leader in the region. Very high values of productivity are noted also in the c	
2 <input type="checkbox"/>	COCA:2019:ACAD Business and Economic Horizons	Namibia (number 13). Large economies, such as South Africa (regional leader in terms of economic development and industrialization) and Nigeria (la	
3 <input type="checkbox"/>	COCA:2019:ACAD Business and Economic Horizons	in the region is lesser - even in the case of Gabon that was the leader in the previous categories. Apart from the South Africa, the highest value of	
4 <input type="checkbox"/>	COCA:2019:ACAD Global Education Review	and socially constructed identities. # Sinead Harmey is lecturer in literacy education and national leader for reading recovery at the International Liter	
5 <input type="checkbox"/>	COCA:2019:ACAD ...rve Bank of New Zealand Bulletin	5822 # She was a mission leader , serving as an area director for East Asia, supervising Southern Baptist mission work	
6 <input type="checkbox"/>	COCA:2019:ACAD Baptist History and Heritage	than Lottie Moon. She was primarily an evangelist but also was a respected mission leader , looked to for guidance by her colleagues and FMB leaders	

Different View

You can add or remove texts below, and then either save these texts as a new virtual corpus, or else add them to an existing virtual corpus.

SAVE AS: leadership  OR ADD TO: --SELECT--

HELP	<input type="checkbox"/> 100	YEAR	GENRE	SOURCE	TITLE
1	<input checked="" type="checkbox"/>	1990	ACAD	AfricaToday	Angola--Prospects for Peace Seem Brighter....
2	<input checked="" type="checkbox"/>	1990	ACAD	AfricaToday	Apartheid, the law and reform in South Africa....
3	<input checked="" type="checkbox"/>	1990	ACAD	AfricaToday	The impending demise of Nigeria's forthcoming third republic....
4	<input checked="" type="checkbox"/>	1990	ACAD	AfricaToday	The right to food...as a weapon?...
5	<input checked="" type="checkbox"/>	1990	ACAD	AfricaToday	Why is participation a dirty word in South African politics?...
6	<input checked="" type="checkbox"/>	1990	ACAD	AmerEthnicHis	A stage in the emergence of the Americanized synagogue among East European Jews: 1890-1910....
7	<input checked="" type="checkbox"/>	1990	ACAD	AmerEthnicHis	Unintentional immigrants: Chicago's Filipino foreign students become settlers, 1900-1941....
8	<input checked="" type="checkbox"/>	1990	ACAD	AmerIndianQ	`He was going along': Motion in the novels of James Welch....
9	<input checked="" type="checkbox"/>	1990	ACAD	AmerIndianQ	In search of recognition: Federal Indian policy and the landless tribes of Western Washington....

4) Your virtual corpus has been created

MY VIRTUAL CORPORA



Detailed overview (NEW: Aug 2020)

HELP		↑	↓	LIST NAME ↑	# TEXTS ↓	# WORDS ↓	FIND KEYWORDS <input checked="" type="radio"/> SPECIFIC <input type="radio"/> FREQ	CREATED ↓
1					8	29,362	NOUN VERB ADJ ADV N+N ADJ+N	108 d
2				CLIMATE CHANGE	100	1,463,917	NOUN VERB ADJ ADV N+N ADJ+N	121 d
3				GLOBAL	41	568,414	NOUN VERB ADJ ADV N+N ADJ+N	121 d
4				LEADERSHIP	3631	22,161,893	NOUN VERB ADJ ADV N+N ADJ+N	0 h
5				SURVEILLANCE	38	253,492	NOUN VERB ADJ ADV N+N ADJ+N	90 d

This could be used as a vocab list

LEADERSHIP [22,161,893 WORDS, 3631 TEXTS] (2.2% OF TOTAL) NOUN VERB **ADJ** ADV N+N ADJ+N [\[ALL CORPORA\]](#) [SAVE LIST](#)

HELP	WORD (CLICK FOR CONTEXT)	FREQ	# TEXTS	SPECIFIC		ENTIRE CORPUS	EXPECTED
				FREQ	TEXTS		
1	DESTINED	176	155	96.2	82	1.8	
2	MULTI-PARTY	339	130	46.7	325	7.3	
3	SOCIOPOLITICAL	310	160	13.7	1,013	22.6	
4	NATIONALIST	1313	405	13.0	4,532	101.1	
5	POST-COLD	325	143	11.7	1,248	27.9	
6	POSTCOLONIAL	338	140	11.6	1,304	29.1	
7	HEGEMONIC	354	172	11.5	1,377	30.7	
8	AGRARIAN	406	117	10.7	1,703	38.0	
9	MULTILATERAL	561	214	9.9	2,548	56.9	
10	AUTHORITARIAN	1160	413	9.8	5,299	118.3	

Your corpus is also searchable

[List](#) [Chart](#) [Word](#) [Browse](#) +

[POS]?

[Find matching strings](#) [Reset](#)

Sections **Texts/Virtual** [Sort/Limit](#) [Options](#)

MY CORPORA ▲


climate change
global
leadership
Surveillance ▼

[Create corpus](#)

[Edit corpora](#)

[Find keywords](#)

[Refresh list](#)


 (HIDE HELP) LOGGED IN

VIRTUAL CORPORA

Create a "virtual corpus" -- essentially your own personalized corpus within COCA. You can create the corpus either by **keywords in the texts** (e.g. texts with the words *investments*, *basketball*, or *biology*), or **information about the texts** (e.g. date, title, or source), or a combination of keyword and text information.

You can then **edit** your virtual corpora, **search** within a particular virtual corpus, **compare** the frequency of a word, phrase or grammatical construction in your different virtual corpora, and also create "**keyword lists**" based on the texts in your virtual corpus.

Click on any of the links above for more information.

 [Detailed overview](#) (NEW: Aug 2020)

Your Turn

In your groups first try working with the corpus searches yourself and then try to think of your own examples.

Answers

nouns modified by **few**

1. month a few months
2. year a few years
3. day a few days
4. week a few weeks
5. minute a few minutes
6. hour a few hours
7. people few people

Answers

- Castigate means to reprimand someone severely
- It has a negative connotation

Conclusion

- Corpora can be useful for beginners
- Allows learners to produce their own hypotheses on word usage and lets learners test them
- Targeted focus on words to teach by frequency

References

- Báisa, V., & Suchomel, V., (2014) SkELL: Web interface for English language learning. In Horák, A. & Rychlý, P. (ed.), Proceedings of Recent Advances in Slavonic Natural Language Processing. Karlova Studánka, Czech Republic, 5-7 December, 63-70.
- Bennett, G.R. (2010). Using Corpora in the language learning classroom: Corpus linguistics for teachers. University of Michigan Press
- Boulton, A. 2009. Corpora for all? Learning styles and data-driven learning. In M. Mahlberg, V. González-Díaz and C. Smith (eds.), *Proceedings of 5th Corpus Linguistics Conference*. http://ucrel.lancs.ac.uk/publications/cl2009/150_FullPaper.doc
- Boulton, A. (2010) Data-driven learning: Taking the computer out of the equation. *Language Learning*, 60(3): 534-572.
- Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System*, 25(3): 301-315.
- Cobb, T. & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (eds), *Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, p. 478-497. doi: 10.1017/CBO9781139764377.027
- Davies, M. (2020). English-Corpora.org: a guided tour. English Corpora.org. <https://www.english-corpora.org/pdf/english-corpora.pdf>
- Johns, T. (1988). Whence and whither *classroom concordancing*? In T. Bongaerts, P. de Haan, S. Lobbe and H. Wekker (eds.), *Computer applications in language learning*, 9-27. Dordrecht: Foris.

Contact Information

sarah@aoni.waseda.jp