



*Using data-driven learning to  
discover patterns of grammar  
and collocations*

---

Sarah Deutchman

*Waseda University*



# *Agenda*

Explanation of corpora and DDL

How to do basic searches using COCA

How to create classroom materials

# *What are corpora?*

- “Essentially a corpus is a collection of naturally occurring, computer-readable texts, often comprising many millions of words, which is considered more or less representative of a particular domain of language use” (Hyland, 2013, p. 248).

# *What is Data-driven learning(DDL)?*

- DDL can be defined as an inductive, process that allows discovery-oriented learning (Johns, 1990).
- DDL allows learners to look at naturally occurring language and find patterns on their own (Boulton, 2009).
- DDL worked well for learning vocabulary items, basic grammar items, and verb phrases (Mizumoto & Chujo, 2015)

# What are collocations and colligations?

- Collocations can be defined as what words co-occur together or word associations (Hoey, 2014)

*Strong coffee not powerful coffee*

- Colligations can be defined as the environment in which a word would be used in, what words are connected to it, and what the grammatical patterns are (Sinclair 1996, 1999, 2004; Hoey 1997a, 1997b, as cited in Hoey & O'Donnell, 2008).

*Despite +det+ adj + noun [ Despite a serious injury; Despite a pristine record; Despite a decent cast]*

- Both collocations and colligations can be difficult for L2 learners as they cannot always be directly translated (e.g., wear vs. 'kiru', 'kaburu', 'haku', 'hameru', 'kakeru', 'shiteru', 'chakuyou shiteiru').

# *What is the COCA?*

- Corpus of Contemporary American English (COCA) is a large corpus that is separated in different genres: news, spoken, academic, etc. (Davies, n.d.)
- Contains over 1 billion words (Davies, n.d.)
- Can be used to check collocations, colligations, historical trends, genres, etc.

## *Keyword in context searches*

- Shows how the words are connected.

List Chart Word Browse Collocates Compare **KWIC** -

analyze [POS] ?

L - - - - - - R \*

**Keyword in Context (KWIC)** Reset

Sections Texts/Virtual Sort/Limit Options

# KWIC 200

ad never heard Bush "	analyze	a complex issue , parse opposing positions ,
if you would , Brit , to	analyze	a couple of those points that I just made in p
sed . For example , we	analyze	a deletion in the A2M gene at the 5 ? splice s
o the room every day ,	analyze	a handful of schedules that we deemed wort
collect , organize , and	analyze	a large amount of data to produce original co
needed to create and	analyze	a model of a human bone as it grew . For thi
ther it was possible to	analyze	a National Basketball Association (NBA ) bas
han Salam are here to	analyze	a packed week of news . Plus : uncovering th
,000 search words , to	analyze	a passage and compares texts to norms crea
sign , administer , and	analyze	a student survey that would give voice to mic
, p < .01 (d= 0.26 ) . To	analyze	acquisition of partial knowledge for syntactic
al is to get students to	analyze	all sides of an issue . Mr-TOM-TREECE-1Te : Y

# Patterns with analyze

- Analyze + determiner (a) + adjective + noun
- Analyze + determiner (a) + noun + preposition



List Chart Word Browse Collocates Compare **KWIC** -

due to the fact that

[POS]?

L - - - - 1 2 3 R \*

Find collocates

Reset

Sections Texts/Virtual Sort/Limit Options

- 1
- IGNORE
  - 
  - TV/MOVIES
  - BLOG
  - WEB-GENL
  - SPOKEN
  - FICTION
  - MAGAZINE
  - NEWSPAPER
  - ACADEMIC

- 2
- IGNORE
  - 
  - TV/MOVIES
  - BLOG
  - WEB-GENL
  - SPOKEN
  - FICTION
  - MAGAZINE
  - NEWSPAPER
  - ACADEMIC

*Can do KWIC  
with strings*



quality of the residuals is apparently

due to the fact that

1973-74 is the largest misprediction of

tions from taxable income . This is

due to the fact that

26 U.S.C. 170(c) does not require that

current health state . This may be

due to the fact that

a majority of the participants were fe

of assessment products . This was

due to the fact that

a majority of units indicated use of Li

n , but there are also fluctuations

due to the fact that

a small rise or fall in the Earth

y , 1987 ; Sodowsky et al. , 1991 ) .

Due to the fact that

adolescence is a critical period of dev

ing supply of historical material is

due to the fact that

Africa has changed dramatically in th

disciplined each other 's behavior ,

due to the fact that

agency relationships involved repeate

to place restrictions . This may be

due to the fact that

an individual diagnosed with SMI ma

cate a non-zero dollar wage value

due to the fact that

any total wages paid to the employe

l undergraduate diversity may be

due to the fact that

at NYU Irish Studies is one of six

oable . In part this may have been

due to the fact that

baseball as a sport/cultural enterp

*verb + 'due to the fact that' + determiner +  
noun*

# Collocates Search

List Chart Word Browse **Collocates** Compare KWIC -

Word/phrase [POS] ?

Collocates

+	4	3	2	1	0	0	1	2	3	4	+
---	---	---	---	---	---	---	---	---	---	---	---

Sections Texts/Virtual Sort/Limit Options

HELP			FREQ	
1	<input type="checkbox"/>	POP	734	
2	<input type="checkbox"/>	CHANGE	445	
3	<input type="checkbox"/>	UNDERSTAND	227	
4	<input type="checkbox"/>	CREATE	219	
5	<input type="checkbox"/>	CHANGING	176	
6	<input type="checkbox"/>	CREATING	120	
7	<input type="checkbox"/>	THINK	109	
8	<input type="checkbox"/>	CREATED	99	
9	<input type="checkbox"/>	CHANGED	88	
10	<input type="checkbox"/>	KNOW	83	
11	<input type="checkbox"/>	PROMOTE	82	
12	<input type="checkbox"/>	PRESERVE	81	
13	<input type="checkbox"/>	SEE	72	
14	<input type="checkbox"/>	BUILD	70	
15	<input type="checkbox"/>	BUILDING	62	
16	<input type="checkbox"/>	EXPERIENCE	62	

# *Verbs associated with culture*



*Word search function- can see bundles, collocations, genre, synonyms, KWIC*

List Chart **Word** Browse +

indicate [POS] ?

See detailed info for word Reset

# indicate

(VERB)  


#915 



1. be a signal for or a symptom of 2. indicate a place, direction, person, or thing 3. to state or express briefly

D M O C G **E**

  YouGlish PlayPhrase Yarn

 JA: Google WordRef Reverso Linguee

## SYNONYMS

NEW: DEFIN +SPEC +GENL

**denote** imply, indicate, reveal, show, signify, suggest **point to**  
indicate, show, signpost **signal** indicate, signal, wink

## CLUSTERS (more)

indicate • indicated by • indicated in • indicated they • indicated he • indicated to • indicate how • indicate to • indicated on

## TOPICS (more)

variance, eg, finding, questionnaire, variable, statistically, correlate, participant, respondent, significantly, sample, ie, score, perceived, scale, selected, positively, assess, correlation, student

## COLLOCATES (more)

NOUN result, study, research, data, finding, report, evidence, analysis

VERB univariate

ADJ significant, recent, positive, previous, negative, present, preliminary, relative

ADV clearly, otherwise, above, significantly, strongly, statistically, overall, respectively

## RELATED WORDS

indicator, indication, indicative, indicated, contraindicate

# Designing classroom materials

- Questions to ask:

*What vocabulary, grammatical patterns do students need to complete the assignment?*

*What mistakes with collocations, grammar, unusual wording have students made before?*

*What level are the students?*

*How much scaffolding will they need?*

*How many lessons should be dedicated to this activity?*

*Might need some lessons to introduce corpora and their usefulness, have students create an account, show basic searches (can be done through pictures on a worksheet), give students time to debrief*

*What media should be used for the activity ( Word, Google Docs)*

# *Types of activities*

- Can use a variety of activities depending on the level of students.

*Multiple choice activities/ Matching activities*

*Gap-fill activities*

*Infer what a word means through context*

*Look at colligations ( grammatical patterns)*

*Use coded feedback and have students look at corpus data to fix their own errors  
(Tono, Satake, & Miura, 2014)*



# Examples of activities

---

**Find the meaning of the idioms. Look at the concordance lines. What do you think the idioms mean?**

- |                       |   |
|-----------------------|---|
| 1) In light of        | a) someone or something that has the power to make something happen |
| 2) Driving force      | b) something done if nothing else works                             |
| 3) Along the lines of | c) because of, considering  |
| 4) Last resort        | d) similar, alike   |

# Examples of activities

WORD 1 (W1): MUCH (1.07)

	WORD	W1	W2	W1/W2	SCORE
1	FASTER	2435	0	4,870.0	4,554.6
2	ELSE	1976	0	3,952.0	3,696.0
3	EASIER	6788	2	3,394.0	3,174.2
4	LONGER	7730	3	2,576.7	2,409.8
5	DEEPER	1190	0	2,380.0	2,225.9
6	EFFORT	1173	0	2,346.0	2,194.1
7	EVERYTHING	872	0	1,744.0	1,631.0
8	FARTHER	864	0	1,728.0	1,616.1
9	WIDER	767	0	1,534.0	1,434.6
10	HARDER	3011	2	1,505.5	1,408.0

WORD 2 (W2): MANY (0.94)

	WORD	W2	W1	W2/W1	SCORE
1	WAYS	14244	0	28,488.0	30,460.8
2	YEARS	26522	2	13,261.0	14,179.3
3	CASES	8223	1	8,223.0	8,792.4
4	REASONS	2927	0	5,854.0	6,259.4
5	AREAS	2167	0	4,334.0	4,634.1
6	INSTANCES	1492	0	2,984.0	3,190.6
7	THOUSANDS	1423	0	2,846.0	3,043.1
8	TIMES	28347	10	2,834.7	3,031.0
9	FORMS	1413	0	2,826.0	3,021.7
10	LEVELS	1366	0	2,732.0	2,921.2

- 1) This activity is (much / many) harder than I thought.
- 2) Surveillance should not be allowed for (much / many) reasons.
- 3) It took (much/many) times to get it right.
- 4) The police put (much/ many) effort into catching the criminal.
- 5) What do you notice about the words used with much and the words used with

# Examples of activities

general , most participants earned between \$9 and \$20 for their	approximate	1.5 hours of participation in the experiment . Overview of
helped to inform future coding . The initial 125 code and	approximate	150 subcode categories were narrowed and finalized into
her husband , she 'd been a parent figure to the	approximate	180 Mormon missionaries in the field - their surrogate mother
placed on each side of the background and turned to an	approximate	30 degrees angle . The background colors are 24 " x 28
of all-cause mortality , CHD , and stroke . With an	approximate	35-year life span after surgery , one additional death would be
to Observe , Recond , and Analyze . The earth is	approximate	400 Billion Years Old , yet the Scientists promoting Man-Cause
I 'm not a strong in math , but from the	approximate	60- 32% and 40% ca n't both be 19 . And 19+12+19=50
The magnitude of SEM in a typical population is likely to	approximate	7 WRCM ( range = 5 to 7 ) when the reliability
, and to substitute for their ignorance they want her to	approximate	a general truth : Athenian , Byzantine , Florentine , Parisian .
to contact the agency you are working with directly . #	Approximate	Adoption Costs # Foster Care Adoptions \$0-\$2500 # Licensed
Americas by boat around 15,000 years ago , which is the	approximate	age of Monte Verde , the oldest known site in the Americas
his theory , point out that carbon dating only indicates the	approximate	age of the item , not its calendar year . So if

- What part of speech usually comes after approximate?

- Numbers
- Verb
- Preposition
- Noun


He	is	my	friend	and	he	is	m
he	is	a	friend	and	he	kind	
He	is	my	friend	and	he	will	
He	is	my	friend	and	I	like	th
he	is	my	friend	and	I	will	ha
he	is	my	friend	and	I	've	go

I	take	my	medicine	every	day
I	get	my	medicine	here	.
I	need	my	medicine	now	. Yes
I	need	my	medicine	soon	. If
I	take	my	medicine	to	make
I	took	my	medicine	today	.
I	realized	my	medicine	was	n't in
I	take	my	medicine	when	it 's

## *Examples of activities*

- KWIC can be used
- I \* my medicine.
- He is \* friend.

# *Conclusion*

- Collocations and colligations can usually be problematic for students
  - Using DDL can help students become more autonomous and realize their own errors
  - Creating materials helps make using the COCA more accessible
- 

# References

- Boulton, A. (2009). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL*, 21(1), 37-54.
- Davies, M. (n.d.). Corpus of Contemporary American English. <https://www.english-corpora.org/coca/>
- Hoey, M. (2014). Words and Their Neighbours. In J.R. Taylor (Ed.), *The Oxford handbook of the word*.  
<https://doi.org/10.1093/oxfordhb/9780199641604.001.0001>
- Hoey, M., & O'Donnell, M. B. (2008). Lexicography, grammar, and textual position. *International Journal of Lexicography*, 21(3), 293-309. <https://doi.org/10.1093/ijl/ecn025>
- Hyland, K. (2013). Corpora and innovation in English language education. In Hyland, K & Wong, L. (Eds.) *Innovation and change in language education* (pp 248-262). Routledge.
- Johns, T. (1990). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, 10, 14-34.
- Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies*, 22, 1-18.
- Tono, Y., Satake, Y., & Miura, A. (2014). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL*, 26(2), 147-162. doi:10.1017/S095834401400007X