

## Session 2. Corpora as an Authentic Resource of Language and Beyond

*Time: 11 am – 12 pm Winnipeg Time (CST)*

*Presented by Yuliana Bagan, English Online Inc., Winnipeg, Manitoba*

December 6<sup>th</sup>, 2014



English Online Inc. Corpora in the ESL  
Classroom by [Yuliana Bagan](#) is licensed under  
a [Creative Commons Attribution-NonCommercial-  
ShareAlike 4.0 International License](#).

Funded by



Citizenship and  
Immigration Canada

Citoyenneté et  
Immigration Canada

# Agenda

*Corpus?*

*Why corpus: applications in the ESL classroom*

*Conclusions*

*More resources*

*Questions*

# Corpus?

What do you think a corpus is? Please type in the chat box.



# Corpus?

- Corpus is a weird word.
- It has a weird plural form.
- It's large?
- It shows how words are connected.
- It shows which words tend to go together.
- It is some kind of a tool that linguists use to research a language.

# Corpus?

- Do you use corpus?
  - If not, then what do you expect to be able to do with corpus?
  - If yes, then how?

# Corpus: definition

corpus

*plural* corpora [countable]

1 *formal* a collection of all the writing of a particular kind or by a particular person

*the entire corpus of Shakespeare's works*

2 *technical* a large collection of written or spoken language, that is used for studying the language:

*a corpus of spoken English*

(Online Longman Dictionary)

# Examples of corpus

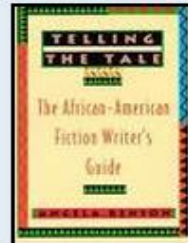
- Conventional
  - [British National Corpus \(BNC\)](#)
  - [Corpus of Contemporary American English \(COCA\)](#)
- Non-conventional
  - [Google](#)
- Corpus based tools
  - [WebCorpLive](#)
  - [Word and Phrase](#)
  - [Just the Word](#)



# COCA: everything covered

## THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH (COCA)

450 MILLION WORDS, 1990-2012



BRIGHAM YOUNG UNIVERSITY

ENTER



# Types of corpus

## 1. **Specialised corpus** – e.g.

- genre: the language of newspapers
- time: 2005 to the present day
- place: just texts published in China

2. **General corpus** – needs to be much larger. E.g. The British National Corpus (BNC) has about 100 million words of spoken and written British English

(Corpus Linguistics Course, Future Learn)

# Types of corpora

3. **Multilingual corpus** – e.g. English and Spanish. Or American English and Indian English.

4. **Parallel corpus** – e.g. English and Spanish – exactly the same texts translated. E.g. the CRATER corpus.

5. **Learner corpus** – language use created by people learning a particular language. E.g. the International Corpus of Learner English.

*The Longman Dictionary of Common Errors is based entirely on a corpus of genuine students' writing - the Longman Learner's Corpus.*

6. **Historical or Diachronic corpus** – e.g. Helsinki corpus – 1.5 million words of texts from 700AD to 1700AD.

7. **Monitor corpus** – continually being added to. e.g. the Bank of English.

(Corpus Linguistics Course, Future Learn)

# **Corpus: applications in the ESL classroom**

*Activities Examples*

# Corpus: from theory to practice

- What *adjectives* collocate with a word Canada?

# Corpus: Let's go to corpus and find out:

Click on the link: <http://corpus.byu.edu/coca/>

1. Type **Canada** in a **Word(S)** string
2. Click on **COLLOCATES**
3. Select **1** on the left side and **0** on the right side
4. In **POS LIST** Choose **Adjectives ALL** from the drop-down list
5. Click **Search**

The screenshot shows the COCA corpus search interface. It has a 'DISPLAY' section with radio buttons for 'LIST', 'CHART', 'KWIC' (which is selected and highlighted in green), and 'COMPARE'. Below this is the 'SEARCH STRING' section. It contains a 'WORD(S)' field with 'Canada' entered. To the left of this field is a yellow box with the number '1'. Below the 'WORD(S)' field is a 'COLLOCATES' section with two input fields: '[j\*]' and '0'. To the left of the first field is a yellow box with the number '2', and to the right of the second field is a yellow box with the number '3'. Below the 'COLLOCATES' section is a 'POS LIST' section with a dropdown menu showing 'adj.ALL'. To the left of this dropdown is a yellow box with the number '4'. At the bottom of the interface are three buttons: 'RANDOM', 'SEARCH', and 'RESET'. To the left of the 'SEARCH' button is a yellow box with the number '5'.

# Corpus: results from COCA

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

COMPARE ▼ ? SIDE BY SIDE ▼

	<input type="checkbox"/>	CONTEXT	FREQ	
1	<input type="checkbox"/>	WESTERN	108	<div></div>
2	<input type="checkbox"/>	EASTERN	100	<div></div>
3	<input type="checkbox"/>	SCHOLASTIC	90	<div></div>
4	<input type="checkbox"/>	NORTHERN	65	<div></div>
5	<input type="checkbox"/>	SOUTHERN	50	<div></div>
6	<input type="checkbox"/>	ENGLISH	26	<div></div>
7	<input type="checkbox"/>	CENTRAL	25	<div></div>
8	<input type="checkbox"/>	UPPER	24	<div></div>
9	<input type="checkbox"/>	ENGLISH-SPEAKING	19	<div></div>
49	<input type="checkbox"/>	ONLY	2	<div></div>
50	<input type="checkbox"/>	SUBURBAN	2	<div></div>
51	<input type="checkbox"/>	TORY	2	<div></div>
52	<input type="checkbox"/>	UNLIMITED	2	<div></div>
53	<input type="checkbox"/>	NON-TERRORIST-PROMOTING	1	<div></div>
54	<input type="checkbox"/>	NON-LOCAL	1	<div></div>
55	<input type="checkbox"/>	NOBLE	1	<div></div>
56	<input type="checkbox"/>	NINETEENTH-CENTURY	1	<div></div>
57	<input type="checkbox"/>	NEUTRAL	1	<div></div>
58	<input type="checkbox"/>	NESTING	1	<div></div>

# #1 Collocations Matter: example

- Use corpora to find “*just the right word*”
  - Search word collocations and use the lists for enriching learners’ vocabulary
  - Brainstorm possible word combinations and check them against corpora

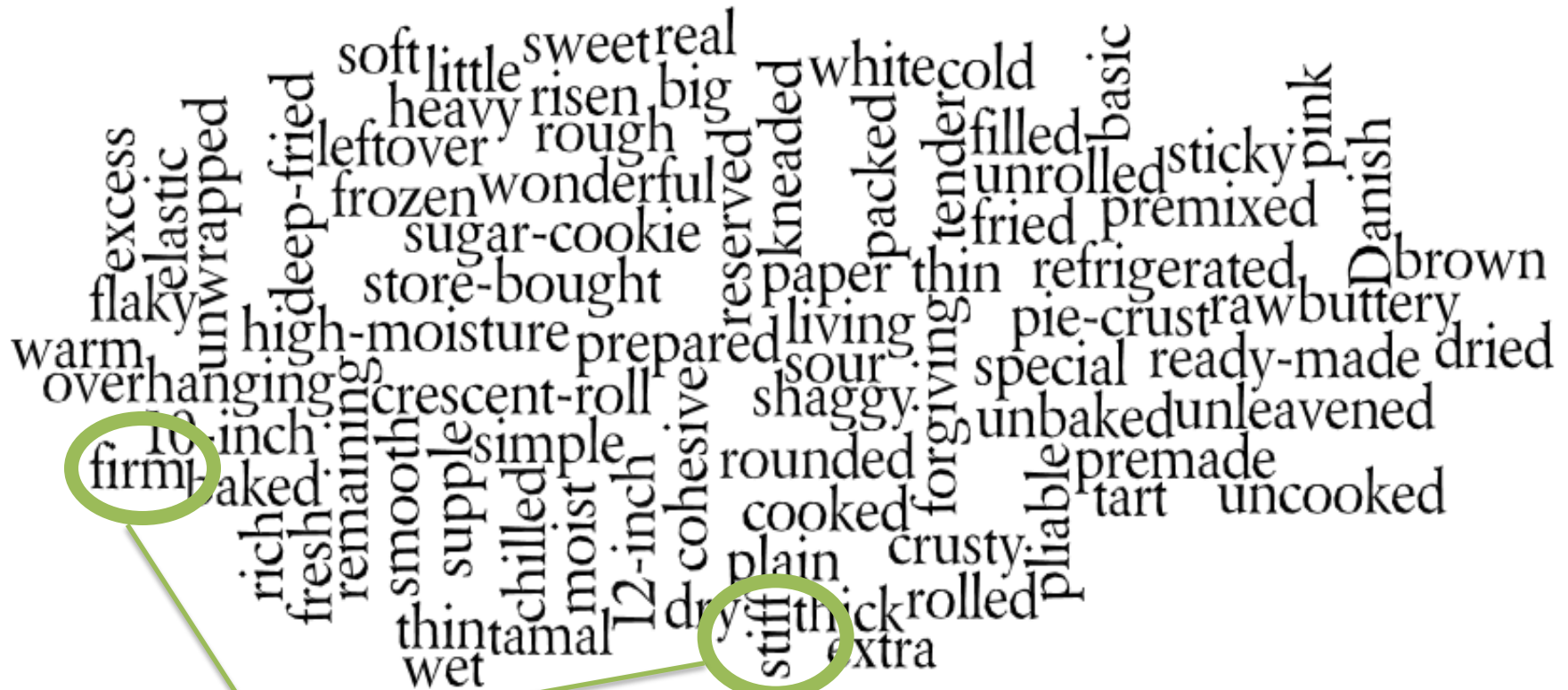


# Collocations matter:

## Example 1

- A learner looking for a word to describe the density of *dough*
  - tough, firm, hard?

# Collocations Matter: dough



# Aha!!!

# And...

# 100 more adjectives to describe *dough*



# Dough can actually be *forgiving*?!



Photo Courtesy of Adrian Scottow CC-BY-SA <http://goo.gl/E7woXw>

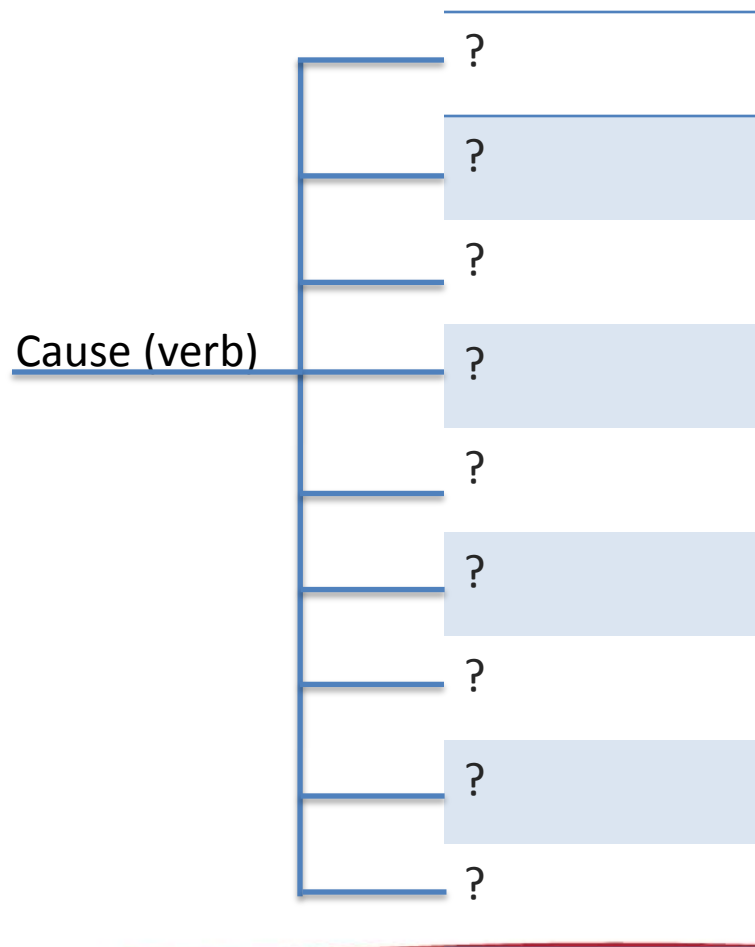
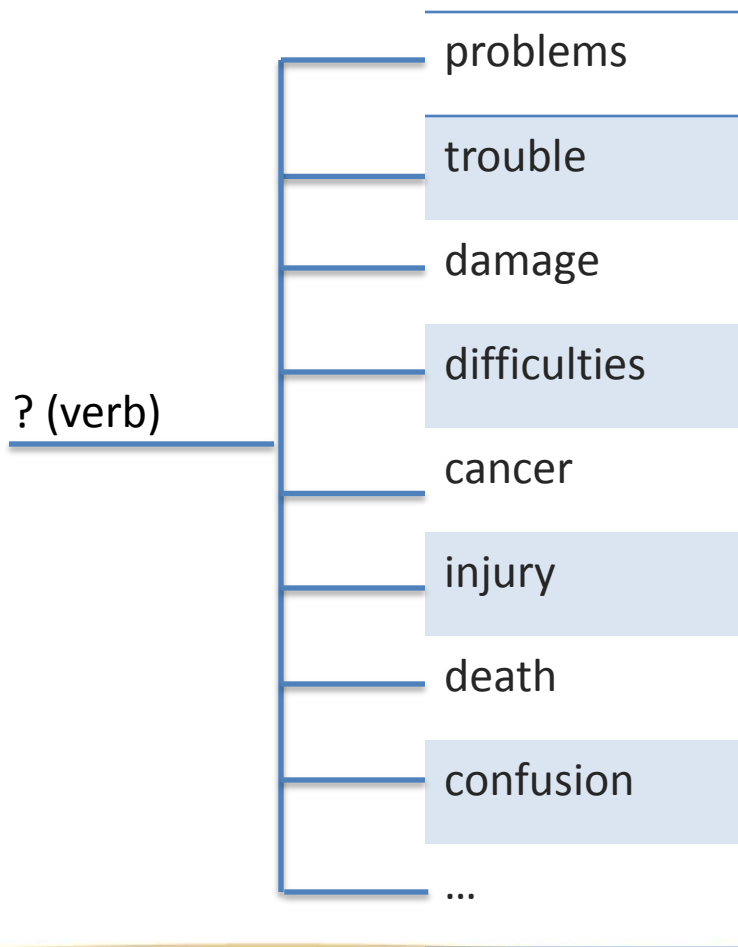
# Collocations Matter

## Example 2

- Present new vocabulary not as isolated words, but in chunks and collocations, or use it as a warm-up activity
  - Fork activities

# Collocations Matter

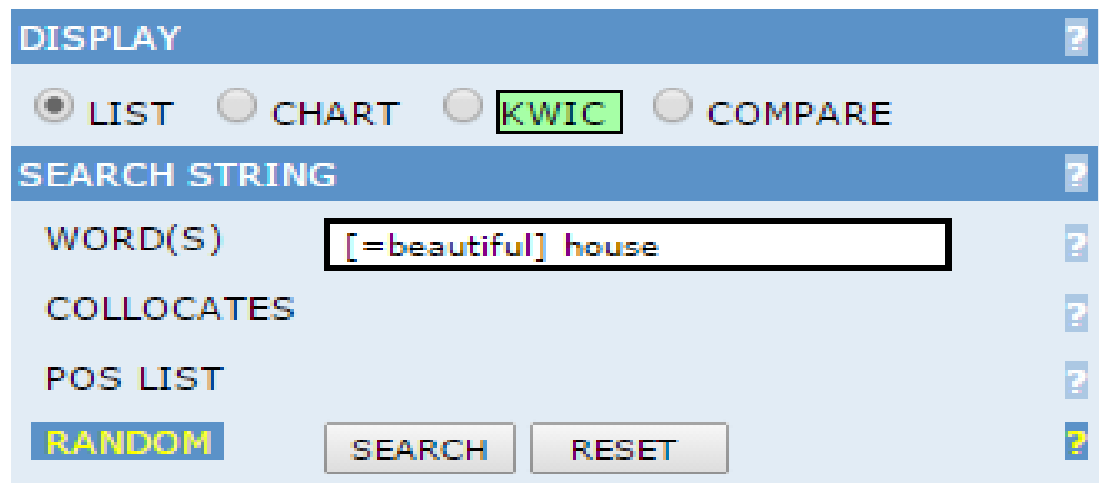
## Fork Activities



# Collocations Matter

## Example 3

- Build on commonly used words learners know and expand their vocabulary by searching for synonyms
  - *a beautiful house*



The screenshot shows a web interface for searching collocations. It features a 'DISPLAY' section with radio buttons for 'LIST', 'CHART', 'KWIC' (which is selected and highlighted with a green box), and 'COMPARE'. Below this is a 'SEARCH STRING' section with a text input field containing '[=beautiful] house'. The input field is also highlighted with a black border. To the right of the input field and below the 'SEARCH STRING' label are question mark icons. At the bottom, there are buttons for 'RANDOM', 'SEARCH', and 'RESET', each with a question mark icon to its right.

DISPLAY	
<input checked="" type="radio"/> LIST	<input type="radio"/> CHART
<input checked="" type="radio"/> KWIC	<input type="radio"/> COMPARE

SEARCH STRING	
WORD(S)	<input type="text" value="[=beautiful] house"/>
COLLOCATES	
POS LIST	



# Collocations matter

- They will be surprised to see the results:

	<input type="checkbox"/>	CONTEXT	FREQ	
1	<input type="checkbox"/>	BEAUTIFUL HOUSE	43	<div></div>
2	<input type="checkbox"/>	LOVELY HOUSE	31	<div></div>
3	<input type="checkbox"/>	ATTRACTIVE HOUSE	6	<div></div>
4	<input type="checkbox"/>	MAGNIFICENT HOUSE	4	<div></div>
5	<input type="checkbox"/>	HANDSOME HOUSE	4	<div></div>
6	<input type="checkbox"/>	CHARMING HOUSE	3	<div></div>
7	<input type="checkbox"/>	PICTURESQUE HOUSE	3	<div></div>
8	<input type="checkbox"/>	WONDERFUL HOUSE	2	<div></div>
9	<input type="checkbox"/>	SUPERB HOUSE	2	<div></div>
10	<input type="checkbox"/>	DELIGHTFUL HOUSE	2	<div></div>
11	<input type="checkbox"/>	GORGEOUS HOUSE	1	<div></div>
12	<input type="checkbox"/>	EXQUISITE HOUSE	1	<div></div>
		TOTAL	102	

Data by British National Corpus

# Collocations matter

## Let's see examples:

- In the 1780s, when Highgate Hill was so steep and deeply rutted that carriages regularly failed to make the grade, and the drive to town sufficiently dangerous that a wise man went with pistols, a merchant called Thomas Roxborough had constructed **a handsome house** on Hornsey Lane, designed for him by one Henry Holland. (Imajica. Barker, C. Glasgow: HarperCollins, 1992, pp. 7-131. 3030 s-units; BNC)
- The Bishop having departed, Treadwell had walked Theodora down the short drive of his **handsome house**. (Unholy ghosts. Greenwood, D M. London: Headline Book Pub. plc, 1991, pp. 1-142. 3531 s-units; BNC)
- Because although it's a **handsome house**, and the gardens are extensive, they in no way compare to those of the castle which is just up the road. (Love of my heart. Richmond, Emma. Richmond, Surrey: Mills & Boon, 1993, pp. ???. 4267 s-units; BNC)
- The new farm was majestic -- **a handsome house**, a huge acreage, a dairy herd as well as beef cattle, sheep as well as shire horses, and no tractor. (Country Living. London: The National Magazine Company Ltd, 1991, pp. 4-180. 2186 s-units; BNC)

# Collocations matter

- But we can only use handsome when describing a male!?!...

SEARCH STRING		?
WORD(S)	<input type="text" value="handsome"/>	?
COLLOCATES	<input ]<="" td="" type="text" value="[nn*"/> <td>0 ▼ 1 ▼</td>	0 ▼ 1 ▼
POS LIST	<input type="text" value="noun.ALL"/>	?
<b>RANDOM</b>	<input type="button" value="SEARCH"/>	<input type="button" value="RESET"/>

1. Man	81
2. Face	75
3. Features	16
4. Woman	16

# Word and Phrase

## Example 5

- [Word and Phrase](#) - run texts through a corpus based tool to see:
  - the range of vocabulary
  - word lists
  - definitions
  - all within ONE browser tab open

# Word and Phrase: text x-ray

## Frequency Range

SEE LISTS	FREQ RANGE	1-500	501-3000	> 3000	ACAD	HELP
	91 WORDS	73 %	13 %	14 %	11 %	

How do we deal with all this **data** without getting information **overload**? How do we use **data** to **gain** real **insight** into the world? Finding ways to pull interesting information out of **data** can be very **rewarding**, both **personally** and **professionally**. The **managing editor** of Financial Times **observed** on CNN's Your Money: " The people who are able to in a **sophisticated** and **practical** way **analyze** that **data** are going to have **terrific** jobs. " Those who learn how to present **data** in **effective** ways will be **valuable** in every field .

# Word and Phrase: text x-ray

Click to see Academic Vocabulary

SEE LISTS	FREQ RANGE	1-500	501-3000	> 3000	ACAD	HELP
	91 WORDS	73 %	13 %	14 %	11 %	

How do we deal with all this **data** without getting information overload? How do we use **data** to gain real **insight** into the world? Finding ways to pull interesting information out of **data** can be very rewarding, both personally and professionally. The managing editor of Financial Times **observed** on CNN's Your Money: " The people who are able to in a sophisticated and **practical** way **analyze** that **data** are going to have terrific jobs. " Those who learn how to **present data** in **effective** ways will be valuable in every field .

# Word and Phrase: as a dictionary

**OVERLOAD** *n* (RANK 14868, FREQ 814)

	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
CLICK BAR TO LIMIT					
STORED	18	23	52	38	71
MORE	76	106	223	154	255

**DEFINITIONS** (WORDNET) (BAD ENTRY?)

1. an electrical load that exceeds the available electrical power  
2. an excessive burden

**COLLOCATES** (click to see with OVERLOAD)

**adj** sensory, visual, toxic, cognitive, emotional, thermal, psychosomatic, patient, nervous, suffering **noun** information, role, system, conflict, iron, stress, symptom, anxiety, protection, demand **verb** cause, suffer, avoid, result, prevent, experience, occur, refer, deal, link

**CONCORDANCE LINES**

CLICK WORD TO: ☒ **SEARCH AS COLLOCATE** ☐ QUERY THAT WORD [?]

	GENRE	TEXT	WORD	CONTEXT
1	ACAD	P-waves in the mid-precordial leads which <b>denote</b> right atrial	overload	[8] [ ] [ ] # ECG manifestations of straight back syndrome are right
2	FIC	frat , and my smile is plastered on <b>due</b> to information	overload	<b>about</b> lesbians [ ] Two of these guys actually asked for our phone
3	MAG	than any other systems they have produced . <b>Feeling</b> information	overload	<b>already</b> [?] [ ] The following reviews sort through the hype and
4	SPOK	in trouble ? Mr. KAUFMAN : We have a <b>massive</b> debt	overload	<b>among</b> borrowers and secondly , a whole range of financial
5	ACAD	there was no significant <b>correlation</b> between teaching	overload	and <b>adequate</b> studio time during the fall and spring semesters (
6	ACAD	) . Significant Pearson correlations <b>emerged</b> between role	overload	and anxiety (r = .23 , p < .05) and
7	ACAD	and postoperatively . Therefore , <b>with</b> this fluid	overload	and <b>because</b> of dilution factors , the hemoglobin and hematocrit



# Word and Phrase: word lists

- Use word lists for
  - pre-reading activities
  - adapting to different proficiency levels

<b>RANGE 3</b> (COCA LIST > 3000) WORDS
<b>1:</b> analyze, insight, overload, personally, practical, professionally, rewarding, sophisticated, terrific, valuable
<b>RANGE 2</b> (COCA LIST 501-3000) WORDS
<b>5:</b> data <b>1:</b> editor, effective, gain, managing, observed
<b>RANGE 1</b> (COCA LIST 0-500) WORDS
<b>5:</b> to <b>3:</b> how, in, the <b>2:</b> and, are, be, do, information, of, ways, we, who <b>1:</b> a, able, all, both, can, deal, every, field, finding, getting, going, have, interesting, into, jobs, learn, on, out, people, present, pull, real, that, this, those, use, very, way, will, with, without, world
<b>ACADEMIC</b>
<b>5:</b> data <b>1:</b> analyze, effective, insight, observed, practical, present

# Corpus for learning Pragmatics

## Example 6

- Use corpora to teach pragmatic expressions:
  - Thank you, **I am good**. (Politely saying “No” instead of “No, I don’t want it.”)
  - I am **so over it**. (spoken grammar)
  - I **so** want to get it. (Using so followed by a verb)
  - It’s **way** better. (“way better” for “much better”)
  - Are we there **yet**? (yet with Present Simple)

# Other examples

- Spoken corpus for listening activities and spoken grammar
- Corpus of Academic English for EAP/ESP
- Building your own corpus
- And much more...

# Corpus is not ...

- Replacing dictionaries, grammar books, course books or Google
- Substituting teacher in the classroom
- Denying everything we know about the language

# Corpus is ...

- Showing how language exists in reality
- Confirming or denying our intuitions about the language
- Revealing interesting facts about language
- Promoting teacher and learner autonomy
- Proving that language is rich, creative, allowing and forgiving

# Use corpus to ...

- Promote exploratory learning and learner independency
- Trigger imagination, creativity and curiosity about language



Picture Courtesy of Public Domain CC0 <http://goo.gl/VAixHK>

# Got interested?

More free resources here:

- [Corpus Linguistics Community on Google +](#)
- Corpus Linguistics Massive Open Online Course at [Future Learn](#) (coming in 2015)
- [EFL Notes by Mura Nava](#)
- [Classroom Games From Corpora](#) by Ken Lackman
- [A lesson plan to introduce COCA activities in class](#)
- [Professor Geoffrey Leech at Lancaster University](#)



# Corpora list

1. British National Corpus 100 million words, 1980s-1993  
<http://corpus.byu.edu/bnc/>
2. Corpus of Canadian English (Strathy) 50 million words, 1920s-2000s  
<http://corpus.byu.edu/can/>
3. Corpus of Contemporary American English 450 million words, 1990-2012  
<http://corpus.byu.edu/coca/>
4. Word and Phrase (on the basis of COCA) <http://www.wordandphrase.info/>
5. Word and Phrase (on the basis of academic texts from COCA)  
<http://www.wordandphrase.info/academic/analyzeText.asp>
6. Michigan Corpus of Academic Spoken English  
<http://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple>

good enough " And " **thank you** for the present but what's wrong with it? " :  
 MELA: Mrs. Jervis, **than** your kindness. Heaven reward you :  
 Pleasant Hall. PAMELA: **I thank you** /r. MR. WILLIAMS: to MRS. JEWKE  
 formation so I can't **thank you** for it. JACKY: " Ods blood! PAMELA: I a  
 GUIL: (Wry, gentle) **Thank you**; we'll let you know. (The PLAYER has me  
 1. Smiles.) **Thank you**. (The PLAYER turns and goes. ROS has bent for th  
 no confidence in England. **Thank you**. (Thinks about this.) And even if it'  
 the ability of life to **thank you**: He never gave commandment for their d  
 irl Guides or church fet **Thank you** s. # **THANK you** to the great n  
 ence' kettle'means wat **Thank you** I all the other rhymesters, inclu  
 behalf of them, we **thank you**. # ~~KEEP IT COMING~~ FOLKS # TWO million  
 Services department at the council **Thank you**. THANK you, too -- fo  
 artment at the council? Thank you. # **THANK you**, too -- for giving me a  
 youngsters cuddled up yesterday for - - - - - ' **Thank you** ' to the man wh  
 ngering kisses. The princess beam **Thank you**. ' Psychiatric nurse Hub  
 id suggested she sit th **Thank you**, ' said Anne stiffly. ' We think yo  
 328. # A song to say **Thank You** ' # MARJE PROOPS # MR K. Bailey, of  
 n she used a consultant's ' **thank you** ' letter to back her cheeky scam --  
 and gold crown. ' Father, **thank you**, thank you, ' the newly-weds scream  
 . ' Father, thank you, **THANK you** newly-weds screamed. At least 50 l  
 g text of the leaked n **THANK you** for taking time out of your busy sc

# Have questions ?



- Email me at [ybagan@myenglishonline.ca](mailto:ybagan@myenglishonline.ca)
- Skype me at yuliana\_myenglishonline
- Phone me at (204) 946-5140 ext. 204

## Stay tuned for upcoming events:

**What:** [REALIZE National Realize Forum for EAL/ESL Professionals](#)

**When:** January 23 & 24, 2015

**Where:** Online at English Online Inc.

A promotional banner for the REALIZE! forum. It has an orange background with the word 'REALIZE!' in large, bold, white letters with a red outline and a red maple leaf to its right. To the right of the word, the text 'THE NATIONAL ONLINE FORUM FOR EAL / ESL PROFESSIONALS' is written in white. Below this, the text 'Save the date: January 23 & 24, 2015' is written in a dark grey font. At the bottom, a dark grey bar contains the text 'A FREE ONLINE EAL/ESL PROFESSIONAL DEVELOPMENT FORUM' in orange.

**REALIZE!** THE NATIONAL  
ONLINE FORUM  
FOR EAL / ESL  
PROFESSIONALS

Save the date: January 23 & 24, 2015

A FREE ONLINE EAL/ESL PROFESSIONAL DEVELOPMENT FORUM