ВСТУП ДО ПРИКЛАДНОЇ ЛІНГВІСТИКИ: КОРПУСНА ЛІНГВІСТИКА І ВИВЧЕННЯ МОВ

Повторення матеріалу лекції 3

КОРПУСНА ЛІНГВІСТИКА : ПОНЯТТЯ «КОРПУС»

- **корпусна лінгвістика *(corpus linguistics; corpus linguist)***
- **комп'ютерна лексикографія *(computational lexicography)***

a corpus is a large and principled collection of natural texts.

(Biber, Conrad, & Reppen, 1998, p. 12)

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

(Sinclair, 2005)

A corpus is a collection of (1) machine readable (2) authentic texts (including transcripts of spoken data) which is (3) sampled to be (4) representative of a particular language or language variety.

(McEnery, Xiao, & Tono, 2006, p. 5)

ПРАКТИЧНЕ ЗАНЯТТЯ
КОРПУСНА ЛІНГВІСТИКА І ВИВЧЕННЯ МОВ

ПРОБЛЕМИ створення електронних словників на основі друкованих продуктів:

- Ідеалізація слів і тенденція до науковості визначень (*spider "member of Arachnidae, class Aranea*)
- Ігнорування динаміки лексикону (неологізм *appendicitis* – виключено зі словника Murray (1870)
- Невизначеність щодо власних назв (*Shakespeare, England* vs *Persil, Pepsi*)
- Складність диференціяції значень (*to pour*)
- Ілюзорність меж слів та складність їх розмужування

(*of course, insofar as, fire truck*)

Завдання 1:
Реструктурація словників у корпусний формат

- ***Розгляньте корпуси та визначте:***
- Який матеріал вони містять.
- Що слугувало джерелами матеріалу.
- Яким чином матеріал організований.
- Як можна застосувати корпусні дані для дослідження або вивчення мови.

КОРПУСИ АНГЛІЙСЬКОЇ МОВИ

COCA Corpus of Contemporary American English

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

The Student-Transcribed Corpus of Spoken American English

https://www.spokencorpus.org

Та інші https://www.english-corpora.org

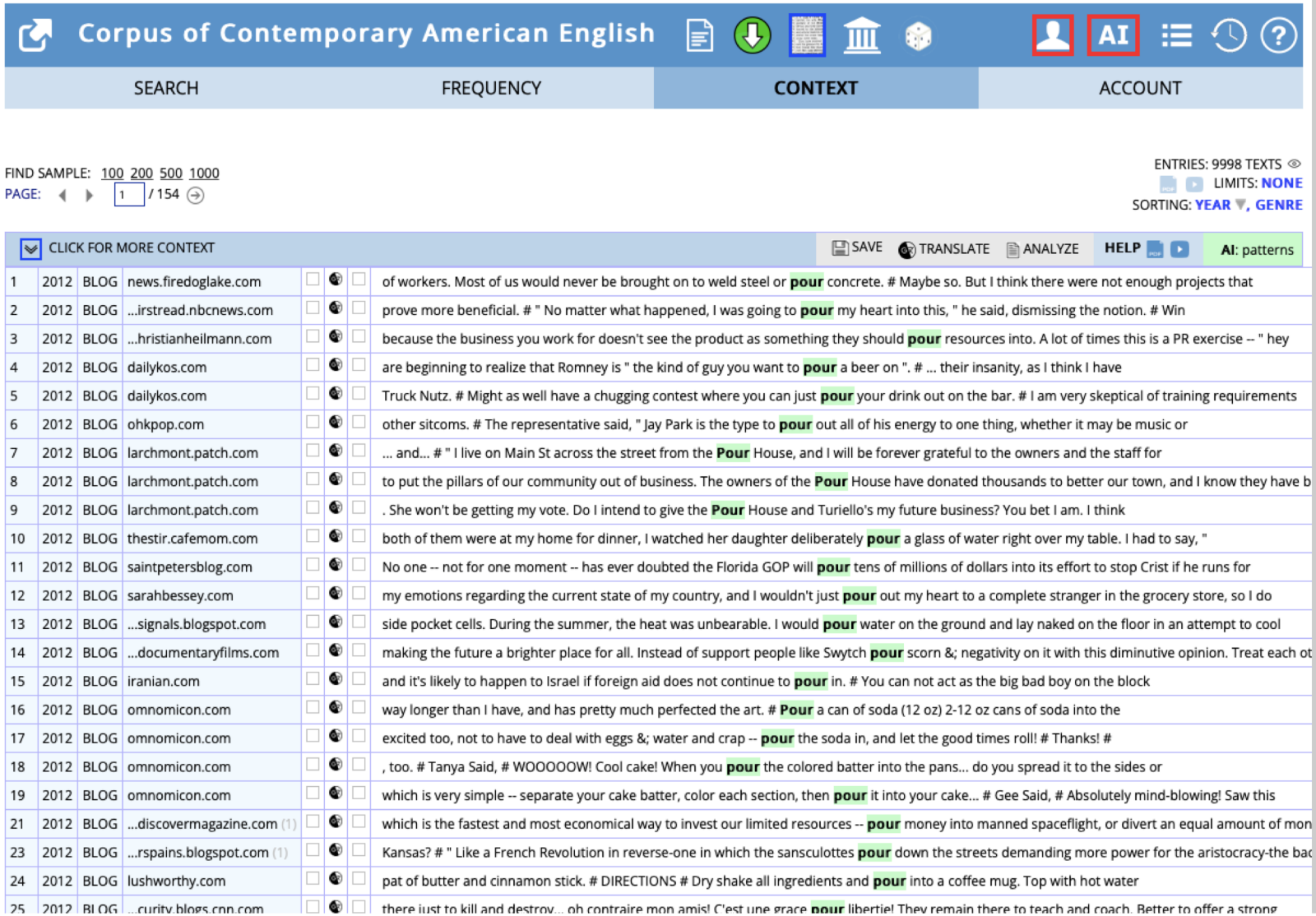
Corpus	Download	# words	Dialect	Time period	Genre(s)
News on the Web (NOW)		23.1 **billion+**	20 countries	2010-**yesterday**	Web: News
iWeb: The Intelligent Web-based Corpus		14 **billion**	6 countries	2017	Web
Global Web-Based English (GloWbE)		1.9 **billion**	20 countries	2012-2013	Web (incl blogs)
Wikipedia Corpus		1.9 **billion**	(Various)	2014	Wikipedia
Coronavirus Corpus		1.5 **billion**	20 countries	2020-2023	Web: News
Corpus of Contemporary American English (COCA)		1.0 **billion**	American	1990-2019	Balanced
Corpus of Historical American En		475 million	American	1820-2019	Balanced
The TV Corpus		325 million	6 countries	1950-2018	TV shows
The Movie Corpus		200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas		100 million	American	2001-2012	TV shows
Hansard Corpus		1.6 **billion**	British	1803-2005	Parliament
Early English Books Online (EEBO)		755 million	British	1470s-1690s	(Various)
Corpus of US Supreme Court Opinions		130 million	American	1790s-2017	Legal opinions
TIME Magazine Corpus		100 million	American	1923-2006	Magazine
British National Corpus (BNC) *		100 million	British	1980s-1993	Balanced
Strathy Corpus (Canada)		50 million	Canadian	1970s-2000s	Balanced
CORE Corpus		50 million	6 countries	2014	Web
From Google Books n-grams (compare)					
American English		155 billion	American	1500s-2000s	(Various)
British English		34 billion	British	1500s-2000	(Various)

The only large, recent, genre-balanced corpus of English, as well as the most widely-used online corpus

Download spreadsheet for all 485,179 texts (with summary by year, genre, and sub-genre)

The corpus is composed of more than **one billion words** in 485,202 texts, including 24-25 million words each year from 1990-2019. For each year (and therefore overall, as well), the corpus is evenly divided between the genres of TV and Movies subtitles, spoken, fiction, popular magazines, newspapers, and academic journals. This is important, because if you want to compare different years, you need to be comparing "apples" to "apples" (i.e. same genre balance in the different periods).

YEAR	BLOG	WEB	TV / MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	TOTAL
TOTAL	125,496,215	129,899,426	128,013,334	127,396,916	119,505,292	127,352,014	122,959,393	120,988,348	1,001,610,938
1990			3,207,900	4,374,469	4,162,242	4,101,447	4,082,931	3,983,143	23,912,132
1991			3,379,151	4,316,898	4,192,646	4,209,838	4,104,806	4,051,046	24,254,385
1992			3,183,858	4,523,054	3,893,956	4,288,694	4,092,031	4,028,147	24,009,740
1993			3,785,924	4,487,978	3,921,244	4,254,351	4,153,070	4,150,671	24,753,238
1994			4,375,338	4,457,726	3,870,757	4,310,375	4,147,947	4,047,115	25,209,258
1995			5,006,966	4,548,602	3,846,412	4,314,737	4,122,703	4,016,371	25,855,791
1996			4,384,976	4,095,266	3,758,787	4,338,766	4,099,305	4,110,209	24,787,309
1997			4,380,670	3,904,996	3,617,741	4,368,917	4,153,906	4,420,786	24,847,016
1998			4,390,197	4,446,217	3,779,801	4,393,835	4,122,295	4,111,453	25,243,798
1999			4,381,144	4,445,564	4,154,537	4,391,146	4,107,423	4,023,282	25,503,096


Corpus of Contemporary American English
AI
SEARCH
FREQUENCY
CONTEXT
ACCOUNT
FIND SAMPLE: 100 200 500 1000
PAGE: 1 / 154
ENTRIES: 9998 TEXTS
LIMITS: NONE
SORTING: YEAR, GENRE
CLICK FOR MORE CONTEXT
SAVE
TRANSLATE
ANALYZE
HELP
AI: patterns
1 2012 BLOG news.firedoglake.com of workers. Most of us would never be brought on to weld steel or pour concrete. # Maybe so. But I think there were not enough projects that
2 2012 BLOG ...irstread.nbcnews.com prove more beneficial. # " No matter what happened, I was going to pour my heart into this, " he said, dismissing the notion. # Win
3 2012 BLOG ...hristianheilmann.com because the business you work for doesn't see the product as something they should pour resources into. A lot of times this is a PR exercise -- " hey
4 2012 BLOG dailykos.com are beginning to realize that Romney is " the kind of guy you want to pour a beer on ". # ... their insanity, as I think I have
5 2012 BLOG dailykos.com Truck Nutz. # Might as well have a chugging contest where you can just pour your drink out on the bar. # I am very skeptical of training requirements
6 2012 BLOG ohkpop.com other sitcoms. # The representative said, " Jay Park is the type to pour out all of his energy to one thing, whether it may be music or
7 2012 BLOG larchmont.patch.com ... and... # " I live on Main St across the street from the Pour House, and I will be forever grateful to the owners and the staff for
8 2012 BLOG larchmont.patch.com to put the pillars of our community out of business. The owners of the Pour House have donated thousands to better our town, and I know they have b
9 2012 BLOG larchmont.patch.com . She won't be getting my vote. Do I intend to give the Pour House and Turiello's my future business? You bet I am. I think
10 2012 BLOG thestir.cafemom.com both of them were at my home for dinner, I watched her daughter deliberately pour a glass of water right over my table. I had to say, "
11 2012 BLOG saintpetersblog.com No one -- not for one moment -- has ever doubted the Florida GOP will pour tens of millions of dollars into its effort to stop Crist if he runs for
12 2012 BLOG sarahbessey.com my emotions regarding the current state of my country, and I wouldn't just pour out my heart to a complete stranger in the grocery store, so I do
13 2012 BLOG ...signals.blogspot.com side pocket cells. During the summer, the heat was unbearable. I would pour water on the ground and lay naked on the floor in an attempt to cool
14 2012 BLOG ...documentaryfilms.com making the future a brighter place for all. Instead of support people like Swytch pour scorn &; negativity on it with this diminutive opinion. Treat each ot
15 2012 BLOG iranian.com and it's likely to happen to Israel if foreign aid does not continue to pour in. # You can not act as the big bad boy on the block
16 2012 BLOG omnomicon.com way longer than I have, and has pretty much perfected the art. # Pour a can of soda (12 oz) 2-12 oz cans of soda into the
17 2012 BLOG omnomicon.com excited too, not to have to deal with eggs &; water and crap -- pour the soda in, and let the good times roll! # Thanks! #
18 2012 BLOG omnomicon.com , too. # Tanya Said, # WOOOOOW! Cool cake! When you pour the colored batter into the pans... do you spread it to the sides or
19 2012 BLOG omnomicon.com which is very simple -- separate your cake batter, color each section, then pour it into your cake... # Gee Said, # Absolutely mind-blowing! Saw this
21 2012 BLOG ...discovermagazine.com (1) which is the fastest and most economical way to invest our limited resources -- pour money into manned spaceflight, or divert an equal amount of mon
23 2012 BLOG ...rspains.blogspot.com (1) Kansas? # " Like a French Revolution in reverse-one in which the sansculottes pour down the streets demanding more power for the aristocracy-the bac
24 2012 BLOG lushworthy.com pat of butter and cinnamon stick. # DIRECTIONS # Dry shake all ingredients and pour into a coffee mug. Top with hot water
25 2012 BLOG ...curity.blogs.cnn.com there just to kill and destroy... oh contraire mon amis! C'est une grace pour liberte! They remain there to teach and coach. Better to offer a strong

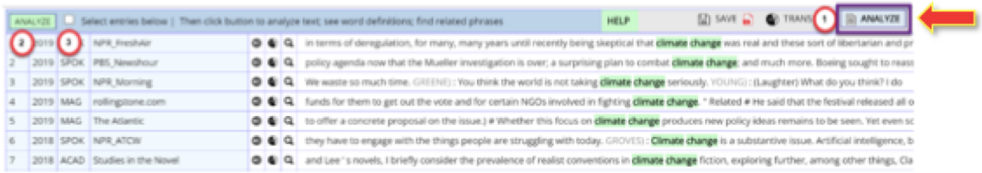
FINDING KEYWORDS FOR A WORD OR PHRASE WITH KWIC ENTRIES

The corpora can help you to quickly and easily find words related to a given word or phrase. For example, for single words (such as *enzyme, lighthouse, environmental*, or *sew*) you can find collocates (nearby words) and topics (words that co-occur anywhere in the text). And using Virtual Corpora, you can quickly and easily create a "sub-corpus" of texts that contain a given word or phrase (such as *New York, Harry Potter, investment*, or *refugee*).

Another easy way to find words related to a particular word or phrase is via the "Analyze (Text)" function from the Keyword in Context page (KWIC; concordances). The following are the steps to do this.

1. Do a LIST search for a word or phrase, and then click on SAMPLE in the LIST display.

This will give you randomly-selected KWIC lines. Click on Analyze [1], which will show the Analyze function [2]. For the best results, click the checkbox [3] to select all of the KWIC lines.



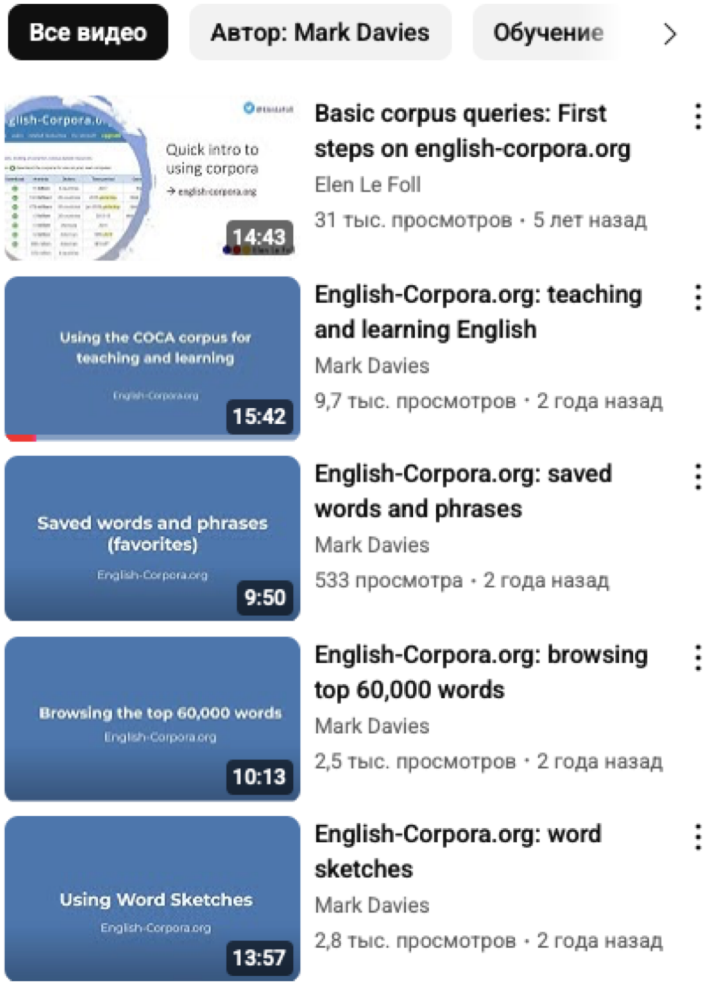
This then creates a "mini-text" of all of the KWIC lines. If there are about 30 words per line and 100 lines of text, this is a text of about 3000 words, and you can analyze this as you would any other text. You can click on any word in the "text" [1] or click on one of the words in the keyword list [2].

For any word that you click on (such as *habitat*, above), you can see a detailed word sketch, which shows (among other things) frequency information [1], definitions [2], links to external pronunciation, videos, and images [3], synonyms and semantically-related words [4], related topics [5], collocates [6], and Keyword in Context entries (not shown here)

And all of this can be done is just 5-10 seconds, via the "Analyze" function in the Keyword in Context (KWIC) display.

Корисні відео: як працювати з COCA

https://www.youtube.com/watch?v=__G2PG46180&t=1s



Корисні відео: як працювати з COCA

https://www.youtube.com/watch?v=OmDsGCJZMx8