TED ÜNİVERSİTESİ

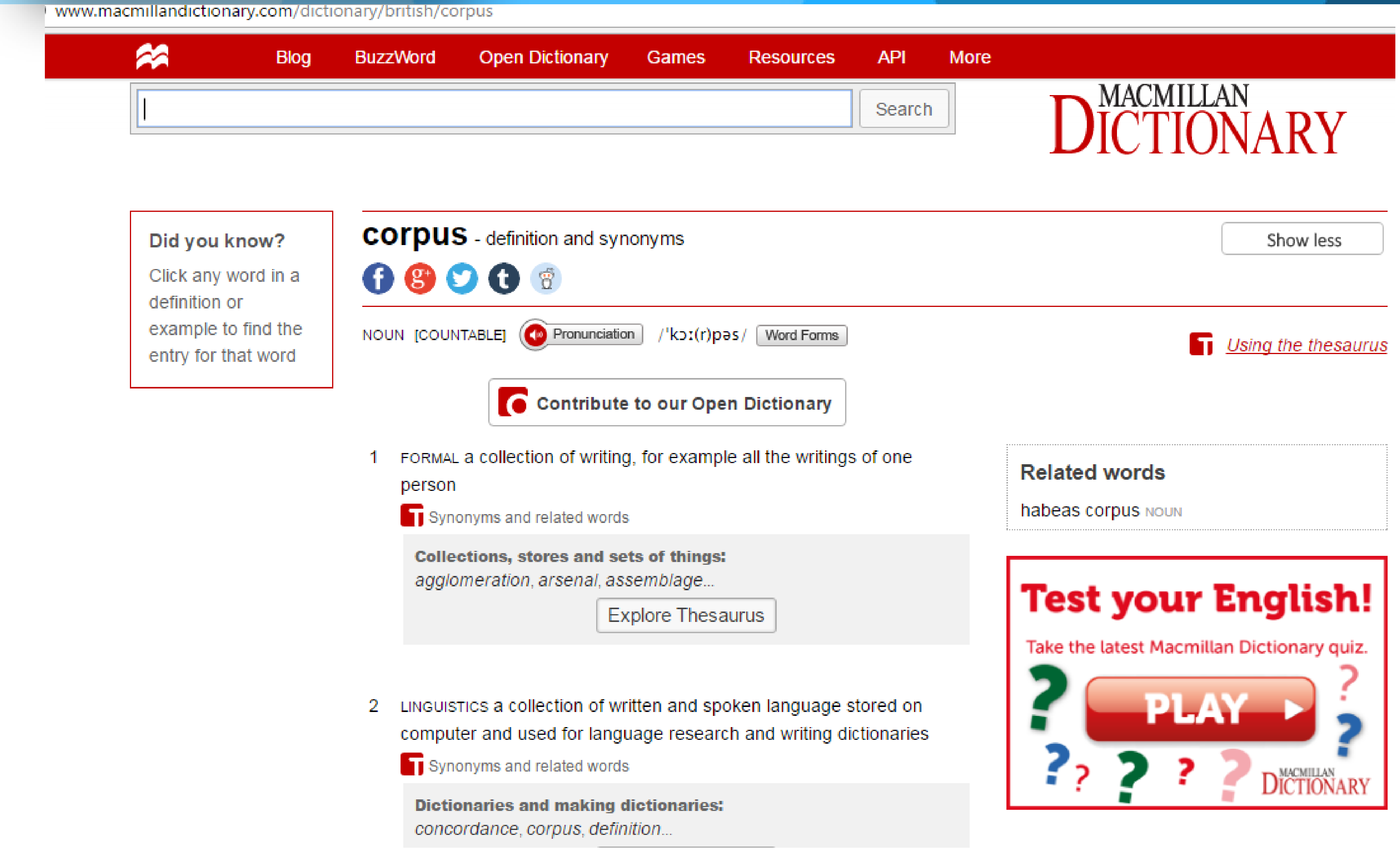
COCA

A Practical and Very Much Helpful Web-based Application for Academic Writing

Aysel Saricaoglu
English Language Education

www.tedu.edu.tr

What does “corpus/corpora” mean?



Did you know?

Click any word in a definition or example to find the entry for that word

corpus - definition and synonyms

NOUN [COUNTABLE] /ˈkɔː(r)pəs/

1 FORMAL a collection of writing, for example all the writings of one person

Synonyms and related words

Collections, stores and sets of things:
agglomeration, arsenal, assemblage...

2 LINGUISTICS a collection of written and spoken language stored on computer and used for language research and writing dictionaries

Synonyms and related words

Dictionaries and making dictionaries:
concordance, corpus, definition...

Related words

habeas corpus NOUN

Corpus/Corpora

In Latin: “body” (same root as “corpse”)

Linguistically: a large (sampled to be representative) and usually electronic (machine-/computer-readable) collection of authentic texts (naturally occurring / written by native speakers)

Self-compilable

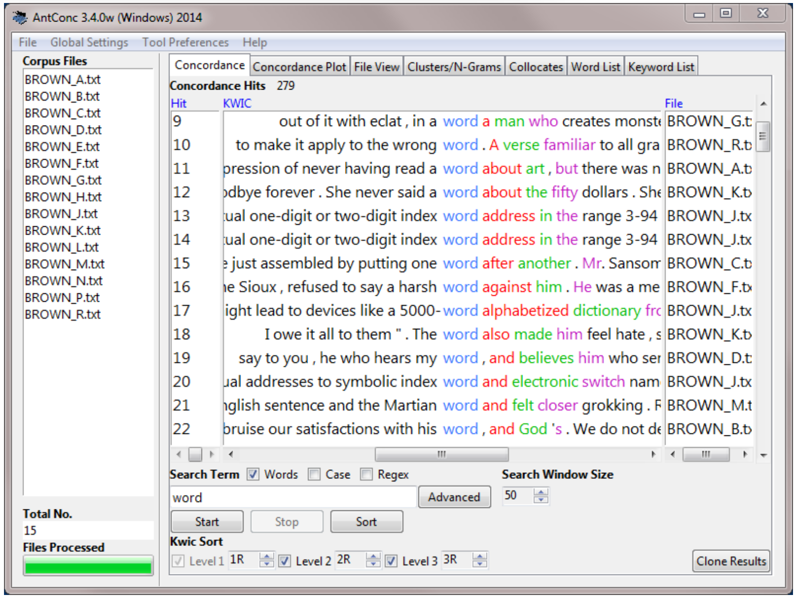
E.g.,

- a corpus of 90s songs
- a corpus of civil engineering research articles
- a corpus of newspapers
- a corpus of telephone conversations
- a corpus of air traffic control conversations

Corpus linguistics: the use of corpora for language study

TED ÜNİVERSİTESİ

You don’t have to be a computational linguist or a computer engineer to use corpora: you can use *concordancers*



Corpus Files

BROWN_A.txt
BROWN_B.txt
BROWN_C.txt
BROWN_D.txt
BROWN_E.txt
BROWN_F.txt
BROWN_G.txt
BROWN_H.txt
BROWN_J.txt
BROWN_K.txt
BROWN_L.txt
BROWN_M.txt
BROWN_N.txt
BROWN_P.txt
BROWN_R.txt

Concordance Hits 279

Hit	KWIC	File
9	out of it with eclat , in a word a man who creates monste	BROWN_G.t:
10	to make it apply to the wrong word . A verse familiar to all gra	BROWN_R.t:
11	pression of never having read a word about art , but there was n	BROWN_A.t:
12	odbye forever . She never said a word about the fifty dollars . She	BROWN_K.t:
13	ual one-digit or two-digit index word address in the range 3-94	BROWN_J.tx
14	ual one-digit or two-digit index word address in the range 3-94	BROWN_J.tx
15	e just assembled by putting one word after another . Mr. Sanson	BROWN_C.t:
16	he Sioux , refused to say a harsh word against him . He was a me	BROWN_F.t:
17	ight lead to devices like a 5000-word alphabetized dictionary fro	BROWN_J.tx
18	I owe it all to them " . The word also made him feel hate , s	BROWN_K.t:
19	say to you , he who hears my word , and believes him who ser	BROWN_D.t:
20	ual addresses to symbolic index word and electronic switch nam	BROWN_J.tx
21	nglish sentence and the Martian word and felt closer grokking . F	BROWN_M.t
22	bruise our satisfactions with his word , and God 's . We do not de	BROWN_B.t:

Search Term ☑ Words □ Case □ Regex

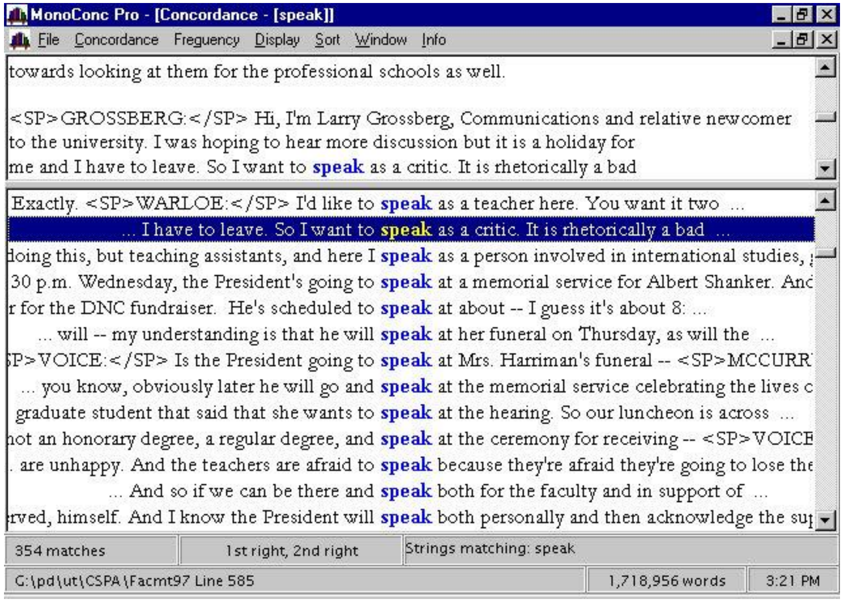
word

Search Window Size 50

Total No. 15

Files Processed

Kwic Sort ☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R



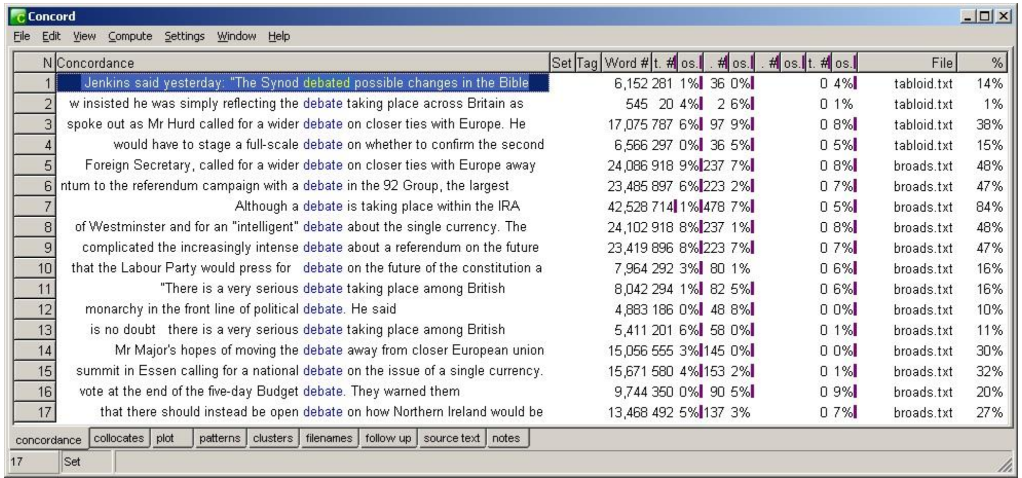
towards looking at them for the professional schools as well.

<SP>GROSSBERG:</SP> Hi, I'm Larry Grossberg, Communications and relative newcomer to the university. I was hoping to hear more discussion but it is a holiday for me and I have to leave. So I want to **speak** as a critic. It is rhetorically a bad

```
Exactly. <SP>WARLOE:</SP> I'd like to speak as a teacher here. You want it two ...
        ... I have to leave. So I want to speak as a critic. It is rhetorically a bad ...
doing this, but teaching assistants, and here I speak as a person involved in international studies, ;
30 p.m. Wednesday, the President's going to speak at a memorial service for Albert Shanker. Anc
r for the DNC fundraiser.  He's scheduled to speak at about -- I guess it's about 8: ...
  ... will -- my understanding is that he will speak at her funeral on Thursday, as will the ...
SP>VOICE:</SP> Is the President going to speak at Mrs. Harriman's funeral -- <SP>MCCURR
 ... you know, obviously later he will go and speak at the memorial service celebrating the lives c
 graduate student that said that she wants to speak at the hearing. So our luncheon is across ...
hot an honorary degree, a regular degree, and speak at the ceremony for receiving -- <SP>VOICE
. are unhappy. And the teachers are afraid to speak because they're afraid they're going to lose the
        ... And so if we can be there and speak both for the faculty and in support of ...
rved, himself. And I know the President will speak both personally and then acknowledge the suj
```

354 matches | 1st right, 2nd right | Strings matching: speak

G:\pd\ut\CSPA\Facmt97 Line 585 | 1,718,956 words | 3:21 PM



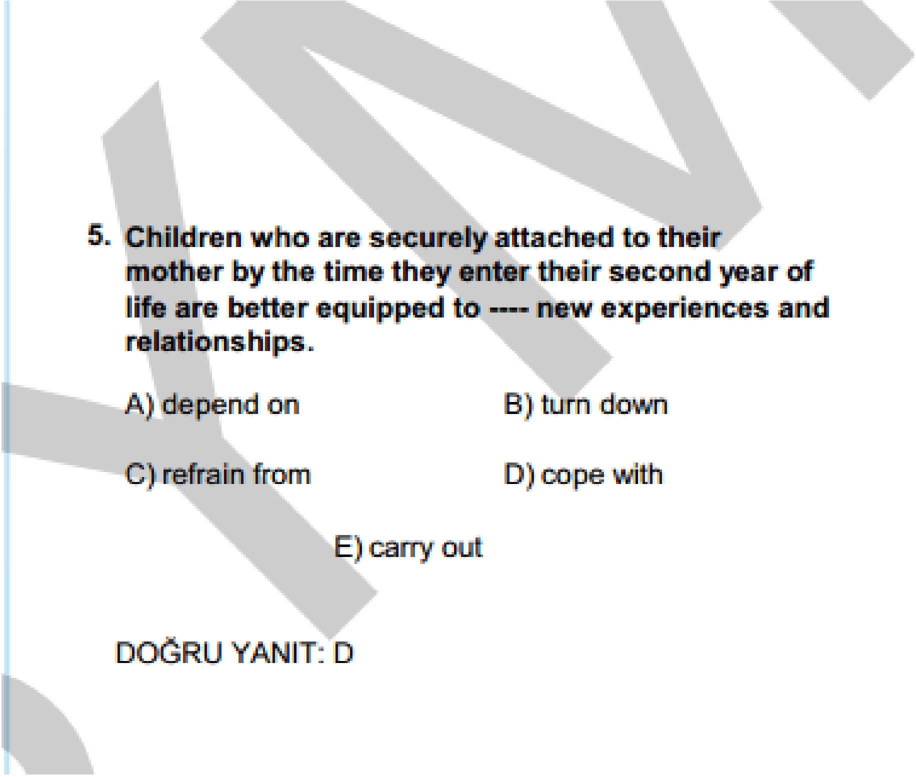
N	Concordance	Set	Tag	Word #	t. #	os.	. #	os.	. #	os.	t. #	os.	File	%
1	Jenkins said yesterday: "The Synod debated possible changes in the Bible			6,152	281	1%	36	0%			0	4%	tabloid.txt	14%
2	w insisted he was simply reflecting the debate taking place across Britain as			545	20	4%	2	6%			0	1%	tabloid.txt	1%
3	spoke out as Mr Hurd called for a wider debate on closer ties with Europe. He			17,075	787	6%	97	9%			0	8%	tabloid.txt	38%
4	would have to stage a full-scale debate on whether to confirm the second			6,566	297	0%	36	5%			0	5%	tabloid.txt	15%
5	Foreign Secretary, called for a wider debate on closer ties with Europe away			24,086	918	9%	237	7%			0	8%	broads.txt	48%
6	ntum to the referendum campaign with a debate in the 92 Group, the largest			23,485	897	6%	223	2%			0	7%	broads.txt	47%
7	Although a debate is taking place within the IRA			42,528	714	1%	478	7%			0	5%	broads.txt	84%
8	of Westminster and for an "intelligent" debate about the single currency. The			24,102	918	8%	237	1%			0	8%	broads.txt	48%
9	complicated the increasingly intense debate about a referendum on the future			23,419	896	8%	223	7%			0	7%	broads.txt	47%
10	that the Labour Party would press for debate on the future of the constitution a			7,964	292	3%	80	1%			0	6%	broads.txt	16%
11	"There is a very serious debate taking place among British			8,042	294	1%	82	5%			0	6%	broads.txt	16%
12	monarchy in the front line of political debate. He said			4,883	186	0%	48	8%			0	0%	broads.txt	10%
13	is no doubt there is a very serious debate taking place among British			5,411	201	6%	58	0%			0	1%	broads.txt	11%
14	Mr Major's hopes of moving the debate away from closer European union			15,056	555	3%	145	0%			0	0%	broads.txt	30%
15	summit in Essen calling for a national debate on the issue of a single currency.			15,671	580	4%	153	2%			0	1%	broads.txt	32%
16	vote at the end of the five-day Budget debate. They warned them			9,744	350	0%	90	5%			0	9%	broads.txt	20%
17	that there should instead be open debate on how Northern Ireland would be			13,468	492	5%	137	3%			0	7%	broads.txt	27%

17 | Set

Why use corpora?

- to find out and model how native speakers speak and write
- to carry out linguistic analyses
- to analyze variation and change in different languages
- to develop authentic language teaching materials
- to get precise data on the target languages for translation purposes
- to look at changes in culture and society
- to use data in natural language processing projects

Why use corpora: YDS example



5. Children who are securely attached to their mother by the time they enter their second year of life are better equipped to ---- new experiences and relationships.

A) depend on

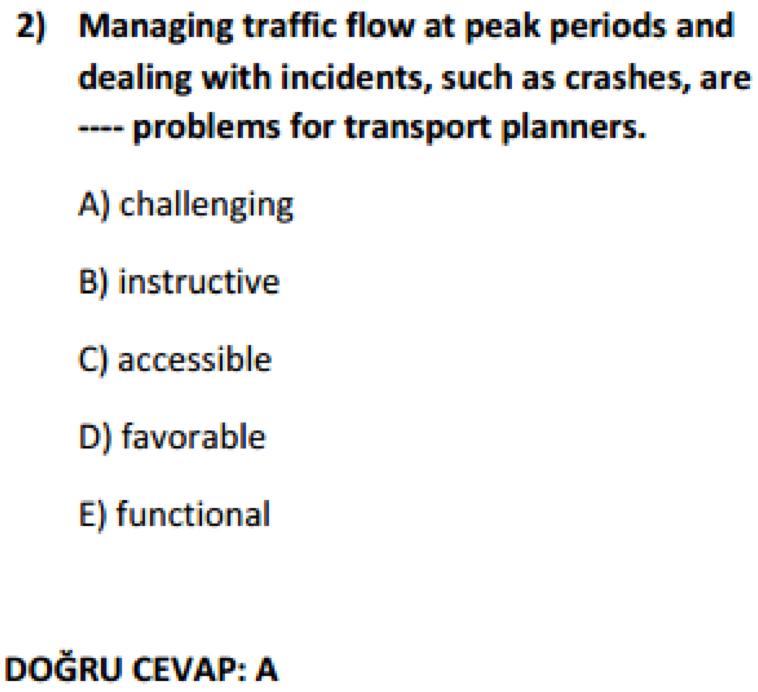
B) turn down

C) refrain from

D) cope with

E) carry out

DOĞRU YANIT: D



2) Managing traffic flow at peak periods and dealing with incidents, such as crashes, are ---- problems for transport planners.

A) challenging

B) instructive

C) accessible

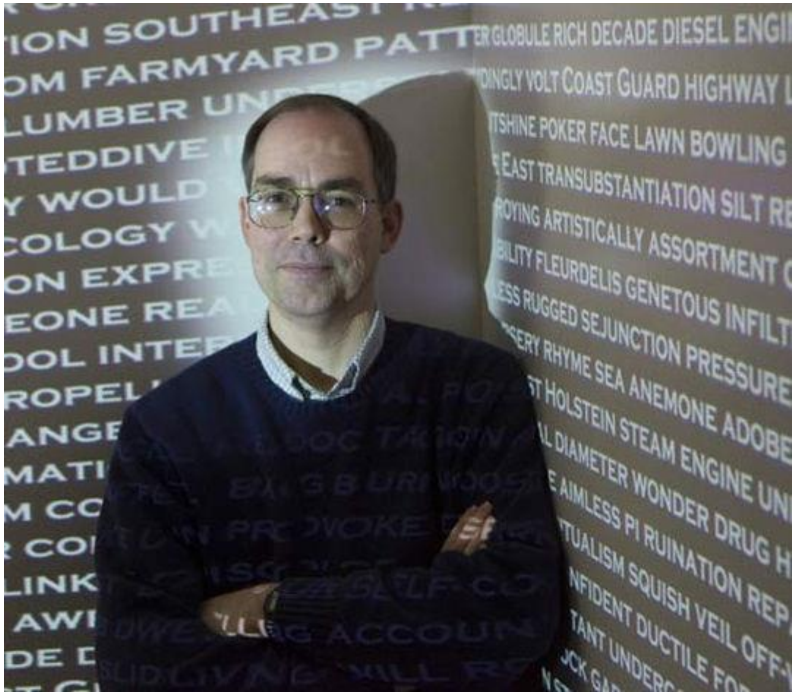
D) favorable

E) functional

DOĞRU CEVAP: A

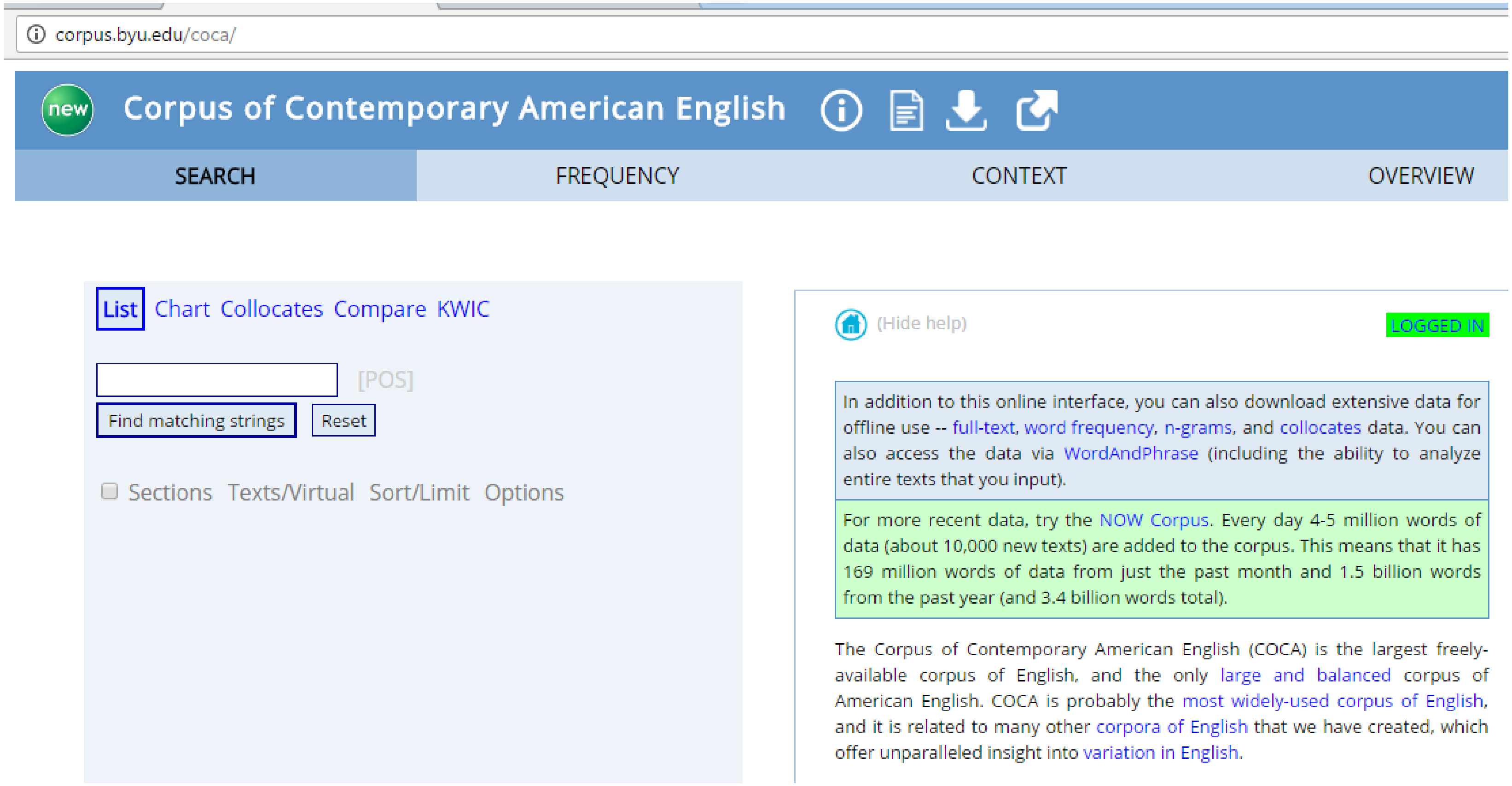
COCA: The Corpus of Contemporary American English

- created by Lingustics professor Mark Davies, BYU
- freely-available
- more than 520 million words of texts between 1990-2015
- equally divided among spoken, fiction, popular magazines, newspapers, and academic texts



Let's try it out!

http://corpus.byu.edu/coca/



In addition to this online interface, you can also download extensive data for offline use -- full-text, word frequency, n-grams, and collocates data. You can also access the data via WordAndPhrase (including the ability to analyze entire texts that you input).

For more recent data, try the NOW Corpus. Every day 4-5 million words of data (about 10,000 new texts) are added to the corpus. This means that it has 169 million words of data from just the past month and 1.5 billion words from the past year (and 3.4 billion words total).

The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English, and the only large and balanced corpus of American English. COCA is probably the most widely-used corpus of English, and it is related to many other corpora of English that we have created, which offer unparalleled insight into variation in English.

List function

Matching forms + frequencies

- Search for **how individual words are used**: innovative
- Search for **how phrases are used**: freak out, FREAK OUT
- Search for **all forms of a word**: EAT; UNDERSTAND
- Search with **parts of speech** : well-established NOUN (or well-established [nn*]); speak ADV; speak [rp*]; poached NOUN
- Search for **patterns matching or including a word (wildcards)**: *new*
- Search for **any word in phrases**: as * to; as * by
- Search by **synonyms**: =fundamentally; in =particular NOUN
- Search with **alternants**: ADJ perspectives|outcomes; essential|necessary NOUN
- Search by **excluding**: pretty –NOUN
- Search by **combining**: =significant findings; we =aim to; these findings =show

Chart function

Section-/Genre- & time-specific matching forms + frequencies

- Search for **lexical variation between genres:** You have to know; sort of; type of; the cream of the crop; the best ; busted; unattractive ; lots of; [get] [vvn*]; look up; research

- Search for **lexical change over time:** BE likely a|the; Recep Tayyip Erdogan; Ataturk; Ankara; Middle East Technical University

Collocates function

words that occur with other words; nearby words

- Search by **distance and grammatical category**: expand (word) NOUN (collocates) | mistakenly (word) VERB (collocates) | opportunity (word) ADJ (collocates)

- Search **by distance only for any word**: controversial (word) * (collocates)

- Search **by distance and for synonyms of a word**: opportunity (word) =great (collocates)

Compare function

compare the collocates of two words

- Search for **differences in usage and meaning**: big/further /less (word1) tall/farther /fewer (word2) NOUN (collocates) |man (word1) woman (word2) ADJ (collocates)

KWIC function

keyword in context

- Search by **sorting the words to the left or right**: LOOK into (123R) | money (321) | sick of (123R)

TED ÜNİVERSİTESİ

Thank you!

Questions/Comments?

Aysel Saricaoglu
aysel.saricaoglu@tedu.edu.tr