Full-text corpus data

The full-text corpus data is available in three different formats. When you purchase the data, you purchase the rights to all three formats, and you can download whichever ones you want.

Samples: The sample data that is linked to below is taken completely at random from each of the corpora (usually about 1/100th the total number of texts). No attempt has been made to "clean up" this sample data in any way. If you're happy with the sample data that you download, you should be equally as happy with the complete set of data.

Note that the size shown in the first column is the total amount of words that you can download, after purchasing the data. The size in the other three columns is for the samples. Note also that the shared files below (sources and lexicon) are just for the sample texts.

NOW corpus: the samples below are just for 2010-2016, but the full-text data continues to grow by 200-220 million words each month. The last update is for . (More info)
Coronavirus corpus: the samples below are just for Jan-May 2020, but the full-text data continues to grow by about 100 million words each month. The last update is for . (More info)

Corpus (size of complete full-text data)		Database (more)	Word/lemma/PoS	Linear text
iWeb (14 billion) COCA (950 million) COHA (385 million) GloWbE (1.8 billion) NOW (billion) Coronavirus (billion) Wikipedia (1.8 billion) TV (310 million) Movies (190 million) SOAP (95 million)	Shared files (See) Lexicon Sources Lexicon Sources Lexicon Sources Lexicon Sources Lexicon Sources Lexicon Sources Lexicon Sources Lexicon Sources Lexicon Sources	Samples (See): 14 mw COCA: 8.9 mw COHA: 3.6 mw GloWbE: 2.1 mw NOW: 1.7 mw Corona: 3.2 mw Wiki: 1.8 mw TV: 2.1 mw Movies: 1.6 mw SOAP: 2.1 mw	Samples (See): 14 mw COCA: 8.9 mw COHA: 3.6 mw GloWbE: 2.1 mw NOW: 1.7 mw Corona: 3.2 mw Wiki: 1.8 mw TV: 2.1 mw Movies: 1.6 mw SOAP: 2.1 mw	Samples: (See): 14 mw COCA: 8.9 mw COHA: 3.6 mw GloWbE: 2.1 mw NOW: 1.7 mw Corona: 3.2 mw Wiki: 1.8 mw TV: 2.1 mw Movies: 1.6 mw SOAP: 2.1 mw
Spanish (1.8 billion) Portuguese (1 billion)	Lexicon Sources Lexicon Sources	Spanish: 2.0 mw Portuguese: 9.5 mw	Spanish: 2.0 mw Portuguese: 9.5 mw	Spanish: 2.0 mw Portuguese: 9.5 mw
Explanation and notes		Most robust format, but requires knowledge of SQL. Allows for powerful JOINs across corpus, lexicon, and sources tables.	Word, lemma, and part of speech in vertical format; can be imported into a database. In most of the corpora, texts are separated by a line with ## and the textID. In COHA, each text is its own file).	This format provides a textID for each text, and then the entire text on the same line. In this format, words are not annotated for part of speech or lemma. In addition, contracted words like <can't> are separated into two parts (ca n't) and punctuation is separated from words (eye level As her).
Short sample		textID / ID / wordID 2002364 / 153180333 / 69 2002364 / 153180334 / 3 2002364 / 153180335 / 978 2002364 / 153180336 / 8880 2002364 / 153180337 / 8047 2002364 / 153180338 / 12 2002364 / 153180339 / 3 2002364 / 153180340 / 351 2002364 / 153180341 / 19630 2002364 / 153180342 / 134 2002364 / 153180343 / 6720	word / lemma / PoS But / but / ccb the / the / at huge / huge / jj bonus / bonus / nn1 prize / prize / nn1 is / be / vbz the / the / at real / real / jj draw / draw / nn1@ -- / -- / x announced / announce / vvn	##2002364 But the huge bonus prize is the real draw -- announced by an electronic display that resembles the ticking wheel on the TV game show , placed just above eye level . As her losses mounted to more than $200 , Budz fed the machine $5 tokens , pressing the Spin button almost rhythmically -- no

			2002364	153180344	38	by	by	ii
			2002364	153180345	42	an	a	at1
			2002364	153180346	3355	electronic	electronic	jj
			2002364	153180347	3923	display	display	nn1
			2002364	153180348	52	that	that	cst_dd1
			2002364	153180349	10985	resembles	resemble	vvz
			2002364	153180350	3	the	the	at
			2002364	153180351	44306	ticking	ticking	jj
			2002364	153180352	3792	wheel	wheel	nn1
			2002364	153180353	22	on	on	ii
			2002364	153180354	3	the	the	at
			2002364	153180355	809	TV	tv	nn1
			2002364	153180356	449	game	game	nn1
			2002364	153180357	3531	show	show	nn1_vv0

serious slot player touches the pull handle on a one-armed bandit .