



[introduction](#) [format/samples](#) [corpora](#) [related sites](#) [get data](#)

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.

NOW corpus: the samples below are just for 2010-2016, but the full-text data continues to grow by 130-150 million words each month. The last update is for **May 2024**. ([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 (1.5 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		<table><tr><th>textID</th><th>ID</th><th>wordID</th></tr><tr><td>2002364</td><td>153180333</td><td>69</td></tr><tr><td>2002364</td><td>153180334</td><td>3</td></tr><tr><td>2002364</td><td>153180335</td><td>978</td></tr><tr><td>2002364</td><td>153180336</td><td>8880</td></tr><tr><td>2002364</td><td>153180337</td><td>8047</td></tr><tr><td>2002364</td><td>153180338</td><td>12</td></tr><tr><td>2002364</td><td>153180339</td><td>3</td></tr><tr><td>2002364</td><td>153180340</td><td>351</td></tr><tr><td>2002364</td><td>153180341</td><td>19630</td></tr><tr><td>2002364</td><td>153180342</td><td>134</td></tr><tr><td>2002364</td><td>153180343</td><td>6720</td></tr><tr><td>2002364</td><td>153180344</td><td>38</td></tr><tr><td>2002364</td><td>153180345</td><td>42</td></tr></table>	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	2002364	153180344	38	2002364	153180345	42	<table><tr><th>word</th><th>lemma</th><th>PoS</th></tr><tr><td>But</td><td>but</td><td>ccb</td></tr><tr><td>the</td><td>the</td><td>at</td></tr><tr><td>huge</td><td>huge</td><td>jj</td></tr><tr><td>bonus</td><td>bonus</td><td>nn1</td></tr><tr><td>prize</td><td>prize</td><td>nn1</td></tr><tr><td>is</td><td>be</td><td>vbz</td></tr><tr><td>the</td><td>the</td><td>at</td></tr><tr><td>real</td><td>real</td><td>jj</td></tr><tr><td>draw</td><td>draw</td><td>nn1@</td></tr><tr><td>--</td><td>--</td><td>x</td></tr><tr><td>announced</td><td>announce</td><td>vvn</td></tr><tr><td>by</td><td>by</td><td>ii</td></tr><tr><td>an</td><td>a</td><td>at1</td></tr></table>	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	by	by	ii	an	a	at1	##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 serious slot player
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																																																																																						
2002364	153180344	38																																																																																						
2002364	153180345	42																																																																																						
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																																																																																						
by	by	ii																																																																																						
an	a	at1																																																																																						

		<table><tr><td>2002364</td><td>153180346</td><td>3355</td></tr><tr><td>2002364</td><td>153180347</td><td>3923</td></tr><tr><td>2002364</td><td>153180348</td><td>52</td></tr><tr><td>2002364</td><td>153180349</td><td>10985</td></tr><tr><td>2002364</td><td>153180350</td><td>3</td></tr><tr><td>2002364</td><td>153180351</td><td>44306</td></tr><tr><td>2002364</td><td>153180352</td><td>3792</td></tr><tr><td>2002364</td><td>153180353</td><td>22</td></tr><tr><td>2002364</td><td>153180354</td><td>3</td></tr><tr><td>2002364</td><td>153180355</td><td>809</td></tr><tr><td>2002364</td><td>153180356</td><td>449</td></tr><tr><td>2002364</td><td>153180357</td><td>3531</td></tr></table>	2002364	153180346	3355	2002364	153180347	3923	2002364	153180348	52	2002364	153180349	10985	2002364	153180350	3	2002364	153180351	44306	2002364	153180352	3792	2002364	153180353	22	2002364	153180354	3	2002364	153180355	809	2002364	153180356	449	2002364	153180357	3531	<table><tr><td>electronic</td><td>electronic</td><td>jj</td></tr><tr><td>display</td><td>display</td><td>nn1</td></tr><tr><td>that</td><td>that</td><td>cst_dd1</td></tr><tr><td>resembles</td><td>resemble</td><td>wvz</td></tr><tr><td>the</td><td>the</td><td>at</td></tr><tr><td>ticking</td><td>ticking</td><td>jj</td></tr><tr><td>wheel</td><td>wheel</td><td>nn1</td></tr><tr><td>on</td><td>on</td><td>ii</td></tr><tr><td>the</td><td>the</td><td>at</td></tr><tr><td>TV</td><td>tv</td><td>nn1</td></tr><tr><td>game</td><td>game</td><td>nn1</td></tr><tr><td>show</td><td>show</td><td>nn1_vv0</td></tr></table>	electronic	electronic	jj	display	display	nn1	that	that	cst_dd1	resembles	resemble	wvz	the	the	at	ticking	ticking	jj	wheel	wheel	nn1	on	on	ii	the	the	at	TV	tv	nn1	game	game	nn1	show	show	nn1_vv0	touches the pull handle on a one-armed bandit .
2002364	153180346	3355																																																																										
2002364	153180347	3923																																																																										
2002364	153180348	52																																																																										
2002364	153180349	10985																																																																										
2002364	153180350	3																																																																										
2002364	153180351	44306																																																																										
2002364	153180352	3792																																																																										
2002364	153180353	22																																																																										
2002364	153180354	3																																																																										
2002364	153180355	809																																																																										
2002364	153180356	449																																																																										
2002364	153180357	3531																																																																										
electronic	electronic	jj																																																																										
display	display	nn1																																																																										
that	that	cst_dd1																																																																										
resembles	resemble	wvz																																																																										
the	the	at																																																																										
ticking	ticking	jj																																																																										
wheel	wheel	nn1																																																																										
on	on	ii																																																																										
the	the	at																																																																										
TV	tv	nn1																																																																										
game	game	nn1																																																																										
show	show	nn1_vv0																																																																										