

The TV and Movie corpora (released February 2019)



The TV Corpus and the Movie Corpus (part of the BYU suite of corpora: corpus.byu.edu) are the largest available corpora of informal English. The Movie Corpus contains 325 million words in 25,000 movie scripts from 1930-2018, and the TV Corpus contains 200 million words in 75,000 very informal TV shows (e.g. comedies and dramas) from 1950-2018.



By way of comparison, these two corpora are (respectively) 32 times as large and 20 times as large as the 10 million words of data in the “conversation” portion of the British National Corpus (BNC), even including the 2014 BNC update:



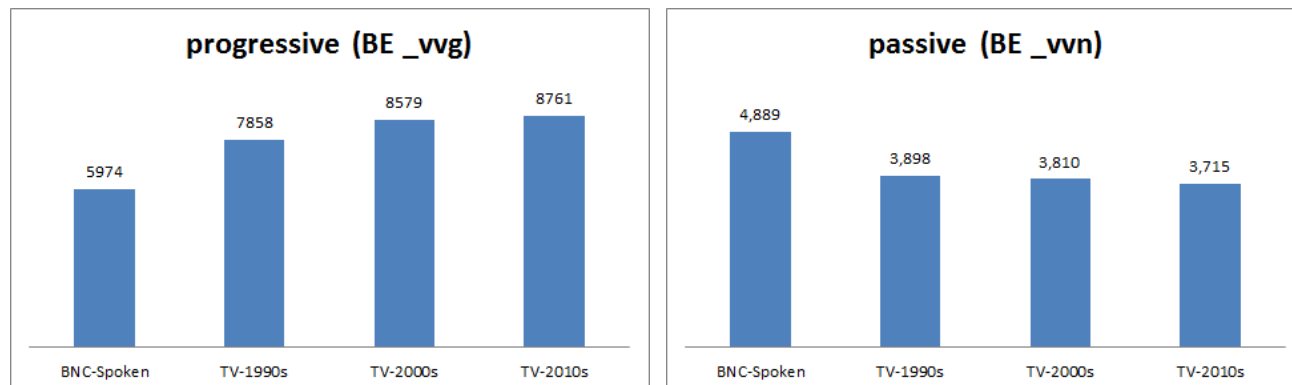
In addition to being very large, the corpora also contain extremely **informal language**. In fact, in many cases it is more informal than the language in actual spoken corpora, like the Spoken portion of the BNC. For example, all of the following phrases are much more common in the TV and Movie corpora than in BNC-Spoken.

. you _vv* me ?	. You heard me? (=subject ellipsis)
, ok okay ?	we're leaving now, OK?
, right ?	you're pretty tired, right?
BE so not ADJ	That is so not possible .
I told you	I told you to get out of here
DO n't get it	I don't get it – how could he even do that?
how can you	How can you even say that?
I totally	I totally get it now!
my God	My God -- she's horrible!
. it 's ADJ .	. It's sad . She's totally forgotten him. (=short phrases)

Even “situational” phrases like the following are much more common in the TV and Movie corpora, showing that these scripts are very oriented to the “here and now”.

hand me * NOUN	Hand me a towel .
. Get out	. Get out before I call the police!
do n't leave	Don't leave! I need you!

Evidence for the highly informal nature of the corpora extends to syntax as well. For example, the progressive (e.g. *I was talking to someone*) occurs more in the Spoken part of the BNC than in any other section (e.g. fiction or newspaper). But the progressive is even more common in the TV and Movie corpora (TV data shown below). Conversely, the passive with *be* (e.g. *the country was colonized in the 18th century*) occurs less in BNC-Spoken than in the other parts of the BNC, and yet it even less common in the TV and Movie corpora (and the scripts in these corpora are becoming even more informal over time).



The TV and Movie corpora can also be used to look at **language change** (TV: 1950-2018; Movies: 1930-2018). In fact, they are the only large corpora that allow us to examine changes in very informal language over the last 70-80 years. (Apologies in advance for the obscenities here; that is also part of the change in the texts over time. Of course in the corpora, none of these entries are censored in any way.)

	More common 1930-1969 (movies)	More common 1990-2018 (movies)
ADJ	swell, splendid, sore, fond, delighted, dreadful, darn, phony, blasted, satisfactory, snappy, darned, apt, no-good, cockeyed, screwy, disgraceful, crummy, beastly, frightful, double-crossing, phoney, bashful, confounded, shrewd, soapy, daffy	f--king, okay, cool, weird, damn, g--d---, huge, awesome, pregnant, super, sexy, scary, unbelievable, sexual, boring, pathetic, gross, massive, nuclear, creepy, global, creative, magical, intense, ultimate, sh-tty, homeless, random, corporate, pissed
NOUN	darling, fellow, pardon, dough, wagon, headquarters, chap, cigar, railroad, brandy, telegram, corporal, crook, hunch, regiment, squadron, handkerchief, shilling, cinch, butler, skipper, chauffeur, plenty, tailor, sonny, mink, nuisance, mammy, waltz, newspaperman	sh-t, hell, mom, f--k, a-s, b-tch, dude, sex, drug, a--h---, tv, bullsh-t, m-f-r, b-st-rd, girlfriend, relationship, d-ck, computer, video, tape, crap, bro, p-ssy, n-g--, grunt, role, bike, chick, cancer, butt
VERB	shall, suppose, pardon, phone, spoil, frighten, telephone, permit, object, congratulate, oblige, dine, notify, faint, quarrel, acquaint, delight, amuse, intrude, dislike, slug, scam, furnish, sock, darn, consent, tangle, fuss, peddle, double-cross	f--k, suck, screw, p-ss, focus, freak, date, r-pe, pee, film, score, b-tch, sh-t, chill, define, stress, evolve, f-rt, activate, surf, tape, participate, process, monitor, target, manipulate, trigger, puke, initiate, generate

In addition, because the texts in the corpora come from six different English-speaking countries, we can also use the data to **compare these varieties**. For example, the following are words that are more common in American and British TV shows:

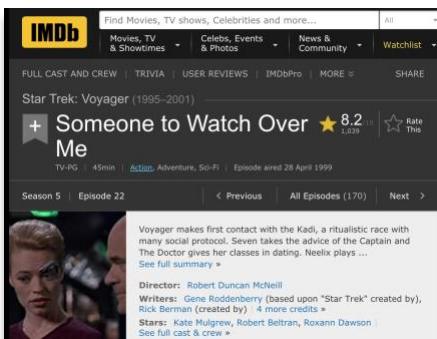
	American	British
ADJ	okay, crazy, damn, awesome, cute, dumb, federal, goddamn, gross, lame, adorable, lousy, crappy, sloppy, phony, downtown, cozy, busted, darn, cranky, high-end, one-time, high-school, canned, cellular, big-time, African-American, goofy, off-limits, old-school, sassy, condescending, puffy, big-a--, sketchy, wordy, charmed, disoriented, kick-a--, bitchy, narcissistic, crummy, self-centered, curt, trashy, whimsical, dorky, scrappy	daft, posh, dodgy, knackered, ruddy, barmy, sodding, poxy, dozy, soppy, mucky, disused, chuffed, tinned, whirly, manky, disorientated, pish, fiddly
NOUN	guy, mom, honey, dude, cop, agent, a--, movie, buddy, apartment, truck, chef, buck, dollar, sweetie, mommy, attorney, mayor, butt, cookie, grandma, a--h---, candy, grade, parking, senator, couch, vacation, closet, homicide, garbage, jerk, baseball, grandpa, elevator, trash, math, thanksgiving, shooter, roommate, bud, assignment, prom, tech, mall, dessert, heck, bout, zombie, soda, motel, halloween, therapist, basketball, counselor, lawsuit, diaper, congressman, chili,	mum, bloke, a-se, quid, rubbish, b-ll-ck, solicitor, railway, vicar, telly, guv, grandad, petrol, ladyship, mammy, shilling, maths, lorry, a---h---, advert, motorway, tosser, tenner, pence, nutter, punter, gearbox, footballer, windscreen, pensioner, barman, pram, tuppence, prat, flatmate, lodger, roundabout, vicarage, workhouse, pillock, sixpence
VERB	guess, figure, kid, damn, date, quit, hire, freak, yell, bust, file, hook, testify, pee, coach, assign, schedule, graduate, violate, practice, dial, jerk, sniffle, participate, brag, party, merge, poop, hustle, reschedule	reckon, fancy, shag, sod, flog, w-nk, queue, burgle, snigger, snog, plod, splutter, clamber

As with all of the other BYU corpora, you can quickly and easily create **“Virtual Corpora”**, which you can then store and search at a later date (and even compare among your different virtual corpora). For example, the Movie corpus allows you to select movies based on year, genre, country, movie rating, IMDB rating, words in the title, the plot, or the script itself, and it creates the corpus in just 1-2 seconds.

SORT	Criteria	Values
<input checked="" type="radio"/>	Year	1930 - 2018
<input type="radio"/>	Genre	<input type="checkbox"/> Drama (11358) <input type="checkbox"/> Comedy (7845) <input type="checkbox"/> Thriller (4081) <input type="checkbox"/> Romance (3804) <input type="checkbox"/> Action (3718) <input type="checkbox"/> Crime (3467) <input type="checkbox"/> Horror (3433) <input type="checkbox"/> Adventure (2851) <input type="checkbox"/> Documentary (2651) <input type="checkbox"/> Family (1821) <input type="checkbox"/> Mystery (1778) <input type="checkbox"/> Sci-Fi (1771) <input type="checkbox"/> Music (1594) <input type="checkbox"/> Fantasy (1457) <input type="checkbox"/> Animation (1306) <input type="checkbox"/> Short (1289) <input type="checkbox"/> Biography (1283) <input type="checkbox"/> History (856) <input type="checkbox"/> War (750) <input type="checkbox"/> Western (591) <input type="checkbox"/> Musical (590) <input type="checkbox"/> Sport (559) <input type="checkbox"/> Film-Noir (386)
<input type="radio"/>	Country	<input type="checkbox"/> USA <input type="checkbox"/> Canada <input type="checkbox"/> UK <input type="checkbox"/> Ireland <input type="checkbox"/> Australia <input type="checkbox"/> New Zealand <input checked="" type="radio"/> Primary <input type="radio"/> Anywhere
<input type="radio"/>	Movie rating	<input type="checkbox"/> R (7106) <input type="checkbox"/> PG-13 (2881) <input type="checkbox"/> PG (2199) <input type="checkbox"/> G (636) <input type="checkbox"/> GP (61) <input type="checkbox"/> X (54) <input type="checkbox"/> M (34) <input type="checkbox"/> NC-17 (24) <input type="checkbox"/> TV-14 (374) <input type="checkbox"/> TV-MA (320) <input type="checkbox"/> TV-PG (228) <input type="checkbox"/> TV-G (208) <input type="checkbox"/> TV-Y (30) <input type="checkbox"/> TV-Y7 (22) <input type="checkbox"/> N/A (5404) <input type="checkbox"/> NOT RATED (3392) <input type="checkbox"/> APPROVED (1709) <input type="checkbox"/> UNRATED (634) <input type="checkbox"/> PASSED (449)
<input type="radio"/>	IMDB rating	Low <input type="checkbox"/> - <input type="checkbox"/> High (Min # votes) 1
	Movie title	<input type="text"/>
	Words in plot	<input type="text"/>
	Word in text	<input type="text"/>
		<input type="button" value="Submit"/> <input type="button" value="Reset"/>

The TV corpus allows you to create similar Virtual Corpora from the 75,000 TV episodes in the corpus. For example, you could quickly create virtual corpora from TV series like Star Trek Next Generation, Dr Who, Friends, or The Office, and then easily compare between these corpora. You can even click on any episode to see the IMDB entry for that show.

STARTREK_VOYAGER (RENAME) DELETE ADD TO MOVE TO --SELECT-- (SEE ALL VIRTUAL CORPORA) SHARE LIST ?									
HELP	<input type="checkbox"/> 100	YEAR	SERIES	EPISODE	COUNTRY	GENRE	RATING	IMDB	
1	<input type="checkbox"/>	1999	Star Trek: Voyager	Someone to Watch Over Me	USA	Action, Adventure, Sci-Fi	TV-PG	8.1 (704)	Voyager makes first contact with the Kadi, a ritualistic race with many social protocol. Seven takes the advice of the Captain and The Doctor gives her classes in dating. Neelix plays ...
2	<input type="checkbox"/>	1999	Star Trek: Voyager	Warhead	USA	Action, Adventure, Sci-Fi	TV-PG	7.2 (677)	An alien warhead or missile that possesses Artificial Intelligence links with the EMH program and begins to terrorize the crew.
3	<input type="checkbox"/>	1999	Star Trek: Voyager	Bride of Chaotical	USA	Action, Adventure, Sci-Fi	N/A	7.4 (456)	While Tom Paris & Harry Kim are running an episode of their \
4	<input type="checkbox"/>	1999	Star Trek: Voyager	Tinker Tenor Doctor Spy	USA	Action, Adventure, Sci-Fi	TV-PG	8.5 (710)	The Doctor's experiment with daydreaming gets out of control when his program is compromised by an alien race. The aliens bully passing ships for supplies, but, before doing so, they first ...
5	<input type="checkbox"/>	1999	Star Trek: Voyager	The Disease	USA	Action, Adventure, Sci-Fi	N/A	6.6 (365)	Voyager encounters a group of xenophobic nomads, at space for 400 years, with serious ship wide malfunctions. This offer to help leads to serious consequences. Strait laced Ensign Kim gets entangled in a forbidden romance.
6	<input type="checkbox"/>	1999	Star Trek: Voyager	Juggernaut	USA	Action, Adventure, Sci-Fi	TV-PG	6.8 (572)	Voyager responds to a distress call of a heavily-damaged Malon freighter. Torres, Neelix, Chakotay, and the only 2 surviving Malon have 6 hours to stop a theta-radiation fallout which will ...



In 1-2 seconds more, you can generate “keywords” from your Virtual Corpus, as with the following noun keywords from Star Trek Next Generation:

HELP	WORD (CLICK FOR CONTEXT)	FREQ	# TEXTS	SPECIFIC FREQ 50 5 TEXTS	ENTIRE CORPUS	EXPECTED
1	COORDINATE	129	65	675.7	99	0.2
2	LIGHT-YEAR	50	23	563.6	46	0.1
3	KILOMETER	159	62	450.5	183	0.4
4	SUBROUTINE	59	21	336.2	91	0.2
5	NANOPROBES	53	13	305.4	90	0.2
6	EMITTER	131	49	273.9	248	0.5
7	NACELLE	50	26	246.9	105	0.2
8	TRICORDER	87	39	240.0	188	0.4
9	HOLODECK	194	55	214.0	470	0.9
10	LIFE-FORM	110	30	209.7	272	0.5
11	THRUSTER	101	45	148.8	352	0.7
12	SENSOR	426	105	136.6	1,617	3.1
13	PHASER	165	71	130.2	657	1.3
14	SUBSPACE	201	63	120.2	867	1.7

For any search, you can of course see excerpts from the TV or movie scripts (as in these entries for the word *feelings* in the TV series *Friends*):

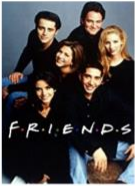
FIND SAMPLE: [100](#)
 PAGE: << < 1/2 > >>

CLICK FOR MORE CONTEXT [?] CHOOSE LIST CREATE NEW LIST [?]

1	2002	Friends	A	B	C	damn you, Geller! Anyway, well I'm glad there's no hard feelings . None at all. You need to be happy with whoever is in your
2	1998	Friends	A	B	C	you under me? Basically lately, I've... I've sort of had feelings for you. I need to lie down. He broke up with Julie!
3	2000	Friends	A	B	C	, my God! You have to go. Why? Because Chandler still has feelings for you. He does? Say again? That's right. That's
4	1997	Friends	A	B	C	What? No. - Oh, God. - What? You still have feelings for me, don't you? No, I'm just excited about the
5	2003	Friends	A	B	C	I think it could be kind of great. Absolutely. You'll love the feeling . There's nothing like it. Okay. Okay. So how should I
6	1999	Friends	A	B	C	can't believe you let them win. Well, at least you hid your feelings well about it. I was frustrated. It's my racket. Frustrated with
7	2002	Friends	A	B	C	athlete I am now. I play squash! Anyway, I always got the feeling he thought I was too sensitive. That must have been hard. It was
8	2002	Friends	A	B	C	weird. I don't.. When my sisters were pregnant.. they got weird feelings and it was always nothing. - Really? - Absolutely. But we'll
9	2002	Friends	A	B	C	why you were there. - You do? - Yeah. You still have feelings for me. To be honest, I still have feelings for you. I
10	1999	Friends	A	B	C	We won't be able to have those long talks at night... about our feelings and the future. Not once did we do that. Don't you remember
11	1997	Friends	A	B	C	wrong with me? - What's the matter? - Tim I have a feeling my wife is sleeping with her gynecologist. How do you know? - He

Finally, there is a wealth of information from IMDB for each of the 25,000 movie scripts and 75,000 TV episodes, and you can always click in the display from the BYU corpora to see more information from IMDB itself.

Source information:

	Series	Friends (IMDB) (Years: 1994–2004: 236 episodes) Country: USA Genre: Comedy, Romance
	Series info	Follows the personal and professional lives of six twenty to thirty-something-year-old friends living in Manhattan.
	Episode	The One with Chandler's Work Laugh (1999) (IMDB) (Open Subtitles)
	Episode info	Length: 22 min / Rating: TV-PG / IMDB rating: 8.4 (1923 votes)
	Episode plot	Monica becomes annoyed at how Chandler sucks up to his boss by mimicking his boss's laugh and laughing at his tasteless jokes. Ross hooks up with Janice when he finds out Emily is getting married.

Expanded context:

n't even breathe, and she's popping pills. You're not giving them a chance. They have rackets. We'll make this the last game. Yes, sir. Put me out of my misery. Are you sure you never played pro? Please let them win. I'll take it down to 95%, but that's the best I can do. - Missed it! - I got it! Nice shot. I got it! Long! I can't believe you let them win. Well, at least you hid your **feelings** well about it. I was frustrated. It's my racket. Frustrated with you! If we hadn't lost, they would never have invited us to dinner tomorrow. What bothers me is how different you act around them. The throwing the tennis games, the fake laugh the " See you later, Bing! " " Not if I see you first, Doug! " I don't like " work Chandler ". The guy's a suck-up. Because you said that I'm not

In summary, the TV Corpus (325 million words) and the Movie Corpus (200 million words) allow you to quickly and easily search through extremely large amounts of data, to gain unparalleled insight into informal, colloquial English.