The TV and Movies corpora

Design, construction, and use

Mark Davies
Brigham Young University

This paper discusses the creation and use of the TV Corpus (subtitles from 75,000 episodes, 325 million words, 6 English-speaking countries, 1950s-2010s) and the Movies Corpus (subtitles from 25,000 movies, 200 million words, 6 English-speaking countries, 1930s–2010s), which are available at English-Corpora.org. The corpora compare well to the BNC-Conversation data in terms of informality, lexis, phraseology, and syntax. But at 525 million words in total size, they are more than 30 times as large as BNC-Conversation (both BNC1994 and BNC2014 combined), which means that they can be used to look at a wide range of linguistic phenomena. The TV and Movies corpora also allow useful comparisons of very informal language across time (containing texts from the 1930s and later for the movies, and from the 1950s onwards for TV shows) and between dialects of English (such as British and American English).

Keywords: TV, movies, diachronic, dialects, speech

1. Introduction

This paper will focus on the design and creation of the TV Corpus and the Movies Corpus (www.english-corpora.org), which are used in some of the other articles in this special issue (Reichelt, Werner). As the sole creator of these two corpora, I can provide some information that might not be available to others. Section 2 of this paper discusses the rationale for these corpora, and Section 3 explains the design and creation of the corpora. Section 4 discusses how the architecture of the corpora allows researchers to focus on specific subsets of the corpora (such as specific movies or TV series) to extract linguistic data particular to those subsets. Section 5 shows how the language of the corpora compares to the language from the spoken portion of other well-known corpora. Section 6 discusses how data from these corpora provides useful information on dialectal

https://doi.org/10.1075/ijcl.00035.dav | Published online: 17 November 2020 | Corrected: 13 January 2021
International Journal of Corpus Linguistics 26:1 (2021), pp. 10–37. issn 1384-6655 | e-issn 1569-9811

variation and historical change in English, as scripted language is the product of a cognitive representation of what people involved in its production see as "natural". Finally, Section 7 offers some general comments about the advantages and shortcomings of the corpora.

2. Rationale for the TV and Movies corpora

Many corpus creators would like to show what is happening in the informal, more "spoken" variety of a language, as opposed to (or at least in addition to) more formal fiction, newspapers, magazines, or academic writing. As corpus creators recognize, however, this is hard to do, since it is very time-consuming and expensive to create a large corpus of the spoken language, because of the effort in recording, transcribing, and then annotating the texts.

As a result, spoken corpora tend to be quite small. For English, for example, the MICASE (Simpson et al., 2002), CALLHOME (Canavan et al., 1997) and CALLFRIEND (Canavan et al., 1996) corpora are all between about one and two million words. This might be adequate for extremely high frequency phenomena (e.g. modals and other auxiliary verbs), but it is far too small to look carefully at medium and lower-frequency words, as well as many syntactic constructions (see Davies, 2015, 2018 for a discussion of corpus size and the range of linguistic phenomena that can be studied with these corpora).

The British National Corpus (2007) is perhaps the only corpus that has a large amount of everyday conversation – about five million words of text from the late 1980s and early 1990s in the BNC1994, as well as 11.5 million more in the 2014 BNC-Spoken update (hereafter BNC2014; see Love et al., 2017).[1] But the BNC is almost a "once-off" type of corpus, since large institutional funding (e.g. generous funding from Oxford University Press) and staffing (a large number of people in the corpus creation team) is not something that is available to most corpus creators. In addition, even though the conversational portion of the BNC corpus is now 16.5 million words (with the 2014 update), that is still more than 30 times smaller than the combined total of the TV and Movies corpora that will be discussed here.

The Corpus of Contemporary American English (COCA) (see Davies, 2008, 2011) is much larger and more recent than these other corpora. COCA contains more than 125 million words of spoken English – four million words each year from 1990 to 2019. These transcripts are for unscripted conversation on TV and

[1] This figure is taken from Love et al. (2017); note that this includes punctuation (Love, 2020), while the figures for the TV and Movies corpora do not include that.

radio programs like *Good Morning American*, the *Today Show*, *All Things Considered*, and *Oprah*. Unfortunately, the conversations often don't deal with "everyday" topics, but rather they often deal with politics, entertainers, the economy, science, business, and other current events.

The problem for corpus creators, then, is that they want to have access to informal language, such as that found in a spoken corpus. But it is almost prohibitively expensive to create a 50 or 100 or 200-million-word corpus of very informal language. There is what is perhaps a fairly easy way to create such corpora, however.

In projects like SUBTLEXus (https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus), rather than using transcriptions of actual recorded speech, data from subtitles of movies and TV are used, on the theory that the dialogue in most TV shows and movies represents the spoken language very well in relation to some lexical and grammatical features (but perhaps not other features like turn-taking patterns or hesitation phenomena). For example, Brysbaert & New (2009); van Heuven et al. (2014), and Brysbaert et al. (2018) all show that the word frequency data from subtitles agrees with native speaker intuitions about their language (as measured by experiments like Lexical Decision Tasks) even better than the data from actual everyday conversation (such as the spoken portion of the BNC). In other words, speakers more readily recognize the words from TV and movies (because they are more commonly used words) than the words from actual spoken corpora. Levshina (2017) and Veirano Pinto (2018) provide similar data and arguments.

Following this line of reasoning, it might make sense to create corpora of subtitles from TV shows and from movies, and we can be quite sure that this data will be a fairly good representation of some aspects of language from actual spoken corpora. (Of course, the language in the two varieties will not be identical, as we will also see throughout this paper; see e.g. Bednarek, 2018; Levshina, 2017 and Forchini, 2012 on differences between TV/movie dialogue and unscripted language.) In addition, an important advantage of these subtitles is that they are readily available. It is quite easy to create 100 million, 200 million, or even 300-million-word corpora of TV shows and movies, which could provide much more data than the spoken portion of the BNC. And of course, this much larger size means that the data can be used to look at a much wider range of features, including medium and low-frequency phenomena in the language. There are other corpora of TV and movie dialogue (see Introduction to this special issue), but they are not as large as the TV and Movies corpora and most often not based on subtitles but rather on transcripts of audio dialogue. Subtitles must be readable by viewers and operate within tight space and time constraints, which may result in reduction of content (e.g. Levshina, 2017; Lugea, 2019). While they are therefore not fully identical to on-screen television/movie dialogue, there are clear similarities between the language of subtitles and transcripts (see Levshina, 2017; for further discussion, see Werner, this issue). The new TV and Movies corpora hence differ from previous corpora both in their size and in their mode.

3. Creating the TV and Movies corpora

So where does one go to get a large amount of subtitles from TV shows and movies? Perhaps the most logical place is the Open Subtitles website (www.opensubtitles.org), which contains subtitles from more than 25,000 movies and more than 75,000 TV episodes. The problem with getting texts from this website, however, is that recently they have incorporated extremely intrusive Javascript code that is designed to prevent users from downloading large amounts of data. For each movie or each TV episode that users attempt to download, the Javascript looks to see whether the mouse has moved to the "download link", which means that it is impossible to use a web browser automator like *Selenium* to download the texts. The only option is to actually click on the links, one by one for each of the 25,000+ movies or 75,000+ TV episodes, and then download the texts via "point/click/save". Even if someone were to do this every 10 seconds for four hours straight in a day (with no breaks), it would take nearly three weeks to download the movies data and nearly two and a half months to download the TV episodes. Obviously, this is not a very inviting proposition.

Luckily, the OPUS Parallel Corpus (opus.nlpl.eu) has already downloaded all of the subtitles data (Lison & Tiedemann, 2016), at least through the end of 2017 (presumably when the Javascript issues were less of a problem for their automated scripts). Best of all, this data is freely available. There is, however, a significant problem in using the data from the OPUS Parallel Corpus: In Open Subtitles, there are at least two sources for the data. First, individual users can submit their version of the subtitles. For example, if someone really likes movie X or TV episode Y, they can watch that movie or episode, transcribe what they hear, and then upload that to Open Subtitles. A second source of data comes from OCR. As Lison & Tiedemann (2016: 926) note, "many subtitles … [were] … automatically extracted via Optical Character Recognition (OCR) from videostreams." What this means is that for a popular TV episode (and even more for a popular movie), there might be several "versions" of the subtitles.

For example, the following are the 20 movies with the most duplicates (as of 2017): *The Lord of the Rings: The Fellowship of the Ring* (137 duplicate texts for the one film), *The Lord of the Rings: The Two Towers* (121 duplicate texts), *The Shawshank Redemption* (88), *The Dark Knight* (87), *Avatar* (87), *Scarface* (81), *Watchmen* (79), *Pulp Fiction* (74), *The Bourne Supremacy* (72), *The Godfather* (71), *Apocalypse Now* (68), *The Last Samurai* (66), *Titanic* (65), *Fight Club* (64), *The Good, the Bad and the Ugly* (64), *House of Flying Daggers* (64), *Pirates of the Caribbean: The Curse of the Black Pearl* (61), *The Girl with the Dragon Tattoo* (61), *The Day the Earth Stood Still* (60), *Braveheart* (58). There are a total of 23,641 movies (out of about 25,000 movies total) that have more than one transcript, and 9,508 movies that have five or more transcripts. Again, this is a serious problem, because we probably wouldn't want all 137 copies of the subtitles for *Lord of the Rings* movies in our corpus.

The following table shows the number of words with and without duplicates for the movies. (While there are duplicates for the TV episodes as well, it is not quite as serious as with the movies.)

Table 1. Duplicates in OPUS Parallel Corpus and Open Subtitles

	Size with duplicates (words)	Without duplicates
1930s	12,003,555	4,574,125
1940s	20,508,362	6,767,339
1950s	28,110,259	8,985,292
1960s	36,784,117	11,903,773
1970s	43,250,227	13,462,814
1980s	58,142,264	14,768,207
1990s	111,292,642	23,471,814
2000s	275,024,411	58,760,647
2010s	182,935,165	45,216,076
Total	**768,051,004**	**187,910,087**

OPUS (the possible source for our subtitles) has all of the duplicate versions from Open Subtitles, with seemingly no way to distinguish among them, or even any way to know that they come from the same TV episode or movie. The filenames in OPUS are simply the Open Subtitles numbers (e.g. 3792253 or 4007229 or 9722836), and all of these filenames would refer to the same TV episode or movie. We wouldn't want 10 or 20 copies of the same movie in our corpus, and so there needs to be some way to eliminate this redundancy. Luckily, there is a solution.

In the metadata for each subtitles page at Open Subtitles, there is a link to the IMDb (Internet Movie Database; www.imdb.com), which contains extensive metadata on more than 100,000 movies and TV episodes – title, year, actors, directors, plot, user ratings, and so on. Because there is only one IMDb entry for each movie or TV episode, we can use the IMDb information at Open Subtitles to find all of the duplicate subtitles that refer to a given TV episode or movie. The downside is that this requires downloading each of the duplicate files (more than 600,000 of them) and searching for the IMDb code. And that is precisely what the intrusive Javascript at the Open Subtitles site prevents us from doing.

There is a solution for this as well, however. The Open Movie Database (www.omdbapi.com) allows us to run automated queries against a huge database that contains detailed information on TV episodes or movies, either by IMDb number or by Open Subtitles number. Using automated queries, a user can run more than 200,000 queries in just two or three hours. Crucially, the information from the Open Movie Database contains both the IMDb number and the Open Subtitles number. If we scrape that information and put it into a relational database, we can then easily identify all of the duplicate versions of a movie or TV episode.

In addition to identifying duplicates, we can even find the "best" of the many duplicate entries. In Open Subtitles, each of the subtitles are "ranked" by other users, according to the perceived accuracy of the subtitles. And those "user rankings" are also available in the Open Movie Database. It is simply a matter of using a GROUP BY statement in the database and then selecting MAX (userRanking) to find which is the best subtitles file for a given movie or TV episode, and then that would be the one that we use in our final corpus. But crucially, in order to wade through the duplicate entries and select the "best" subtitles in the OPUS corpus and the Open Subtitles files – and then compare these to the Internet Movie Database – we probably need to use relational databases or something with equivalent functionality.

In our case all of the corpora from English-Corpora.org (formerly the "BYU Corpora") are built on top of relational databases, and so in just one or two seconds we can sort through information on hundreds of thousands of subtitles files to find the "most accurate" subtitles – one per movie or TV episode. In addition, we also have all of the metadata from IMDb, which we can use to limit our searches to particular sections of the corpus or compare between sections of the corpus (see Section 4 below).

To actually create the corpus, I simply took the "best" file for all of the TV episodes and movies included in OPUS, cleaned it by removing headers and footers in the text, and then tagged the files for part of speech (using the *CLAWS 7* tagger; Rayson & Garside, 1998). I then input the files (with one word + PoS tag per line) into the relational database architecture that I have used for all of the corpora from English-Corpora.org. So, for a 325-million-word corpus (as with the TV Corpus), there would be a database with 325 million rows of data. This is then linked to a number of other tables and databases, including lexicons, frequency by section, and a [sources] table with metadata (from IMDb) for each of the TV episodes or movies (see Davies, 2018 for a description of the corpus architecture).

At this point, perhaps it would be useful to provide a handful of short extracts from the corpora, to show what the actual subtitles data looks like; see Examples (1a) to (1c).

(1) a. All right, that makes more sense. You should have said that at the beginning When you said, "I read a book about anthropology." I don't really know why you're **screaming** at me right now. – I'm not scream – I'm not screaming. That's Meredith's cake. It's her birthday. I don't care. I have an appetite for life! Mmm. Mmm! Oh, god. That's lemon. Good for you, man. Good for you. (TV: The Office: US, 2010)

b. (SCREAMING) Shawn! Cory, what are you doing? Shoving everyone down the elevator shaft. Guess who's next? (SCREAMING) Rachel! Rachel…; (**SCREAMING**) Angela, come on. Everybody's doing it. Doing what? This. (SCREAMING) Hi, Cory. Lauren? What are you doing here? I'm over you. You shouldn't be here. I'm not Lauren. Then who are you? I'm everything you're giving up. (TV: Boy Meets World: US, 1999)

c. (Tracersignal) What? Dad, it's here. (Growling) (Gunfire) (Grunting) (Yelling) No! No! (Gunjams) Oh, my God! (Growls) (**Screaming**) No! No! God help me! (Gunfire) No! (Growling) Oh! Dad! (Screams) (Yells) (Growls) Dad! Nicole…; – Dad! – Nicole. Kill – Kill – Dad? You can still – What? I love you, pumpkin. No. I'm sorry. (Movies: Shaktopus: US, 2010)

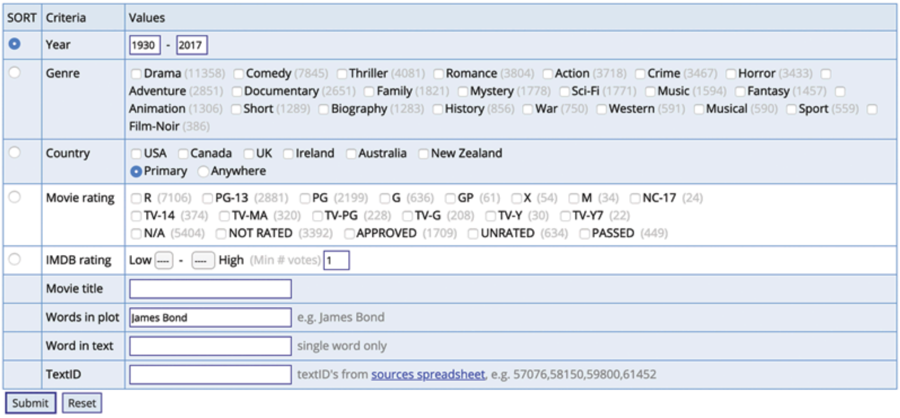
All three of these extracts were taken from the subtitles, in contexts near the word *screaming*. In many cases, as in (1a), the word is simply part of the spoken dialogue, as would be any other word. In other cases, it represents the tone or style of speech, as in (1b) and (1c). In some cases, as in (1c), there are almost as many cases of these elements as actual speech, but passages like this are quite rare. Importantly, nearly all of these "non-speech" tokens are surrounded by parentheses in the displayed text, and they can be eliminated by including the "NOT" operator plus parenthesis in the search, e.g. "-(screaming -)".

4. Using metadata to create "Virtual Corpora"

As discussed in the previous section, one of the advantages of using the Internet Movie Database is that it allows us to remove duplicates from the Open Subtitles data in the OPUS Parallel Corpus – so that instead of having 137 copies of transcripts for *The Lord of the Rings: The Fellowship of the Ring*, for example, we only have one. But there is another important advantage of including the IMDb data in the architecture for the TV and Movies corpora, and that relates to the creation of "Virtual Corpora".

All of the corpora from English-Corpora.org allow users to quickly and easily create "Virtual Corpora", which they can then store and search at a later date (and even compare among their different virtual corpora). For example, in the Wikipedia corpus (www.english-corpora.org/wiki), users can create a "biology" or "investments" corpus, and in the (currently) nine-billion-word NOW corpus (www.english-corpora.org/now) they could, for example, create a corpus of articles from *The Guardian* (UK) from 1 Nov 2019 to 31 Dec 2019 that have *refugees* in the article title or in the text of the article itself.

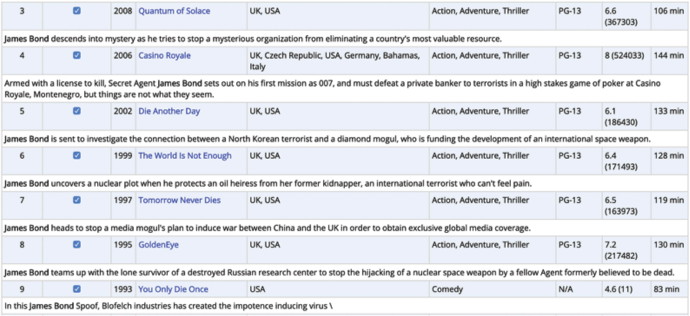
In the TV and Movies corpora, researchers can use the rich metadata from IMDb for each of the 25,000+ movies and 75,000+ TV episodes. For example, as shown in Figure 1, the Movies Corpus allows users to select movies based on year, genre, country, movie rating, IMDb rating, words in the title, the plot, or the text itself, and it takes only 1–2 seconds to find the matching movies in the corpus.



SORT	Criteria	Values
●	Year	1930 - 2017
○	Genre	□ Drama (11358) □ Comedy (7845) □ Thriller (4081) □ Romance (3804) □ Action (3718) □ Crime (3467) □ Horror (3433) □ Adventure (2851) □ Documentary (2651) □ Family (1821) □ Mystery (1778) □ Sci-Fi (1771) □ Music (1594) □ Fantasy (1457) □ Animation (1306) □ Short (1289) □ Biography (1283) □ History (856) □ War (750) □ Western (591) □ Musical (590) □ Sport (559) □ Film-Noir (386)
○	Country	□ USA □ Canada □ UK □ Ireland □ Australia □ New Zealand ● Primary ○ Anywhere
○	Movie rating	□ R (7106) □ PG-13 (2881) □ PG (2199) □ G (636) □ GP (61) □ X (54) □ M (34) □ NC-17 (24) □ TV-14 (374) □ TV-MA (320) □ TV-PG (228) □ TV-G (208) □ TV-Y (30) □ TV-Y7 (22) □ N/A (5404) □ NOT RATED (3392) □ APPROVED (1709) □ UNRATED (634) □ PASSED (449)
○	IMDB rating	Low ---- - ---- High (Min # votes) 1
	Movie title	
	Words in plot	James Bond e.g. James Bond
	Word in text	single word only
	TextID	textID's from sources spreadsheet, e.g. 57076,58150,59800,61452

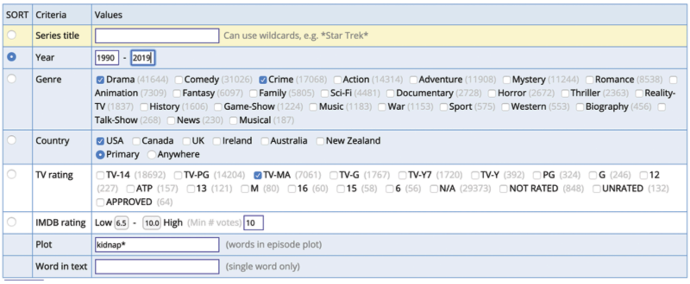
Figure 1. Creating "Virtual Corpus" in the Movies Corpus

Using this metadata, users could for example limit their search to the genre of [comedies] from the US in the 1970s–1990s that are rated R (US MPAA rating, "Under 17 requires accompanying parent or adult guardian; contains some adult material") and which have very poor user ratings in the IMDb – to look at the language of really bad comedies during this period. Or they could quickly and easily create a "Virtual Corpus" of all James Bond movies, resulting in a Virtual Corpus like that shown in Figure 2. Likewise, in the TV Corpus, users could search for crime/drama shows from the 1990s to the present, from the US, which are apparently quite violent (being rated MA-14), and whose plot description mentions "kidnapping", as demonstrated in Figure 3.



3	☑	2008	Quantum of Solace	UK, USA	Action, Adventure, Thriller	PG-13	6.6 (367303)	106 min
James Bond descends into mystery as he tries to stop a mysterious organization from eliminating a country's most valuable resource.								
4	☑	2006	Casino Royale	UK, Czech Republic, USA, Germany, Bahamas, Italy	Action, Adventure, Thriller	PG-13	8 (524033)	144 min
Armed with a license to kill, Secret Agent James Bond sets out on his first mission as 007, and must defeat a private banker to terrorists in a high stakes game of poker at Casino Royale, Montenegro, but things are not what they seem.								
5	☑	2002	Die Another Day	UK, USA	Action, Adventure, Thriller	PG-13	6.1 (186430)	133 min
James Bond is sent to investigate the connection between a North Korean terrorist and a diamond mogul, who is funding the development of an international space weapon.								
6	☑	1999	The World Is Not Enough	UK, USA	Action, Adventure, Thriller	PG-13	6.4 (171493)	128 min
James Bond uncovers a nuclear plot when he protects an oil heiress from her former kidnapper, an international terrorist who can't feel pain.								
7	☑	1997	Tomorrow Never Dies	UK, USA	Action, Adventure, Thriller	PG-13	6.5 (163973)	119 min
James Bond heads to stop a media mogul's plan to induce war between China and the UK in order to obtain exclusive global media coverage.								
8	☑	1995	GoldenEye	UK, USA	Action, Adventure, Thriller	PG-13	7.2 (217482)	130 min
James Bond teams up with the lone survivor of a destroyed Russian research center to stop the hijacking of a nuclear space weapon by a fellow Agent formerly believed to be dead.								
9	☑	1993	You Only Die Once	USA	Comedy	N/A	4.6 (11)	83 min
In this James Bond Spoof, Blofelch industries has created the impotence inducing virus \								

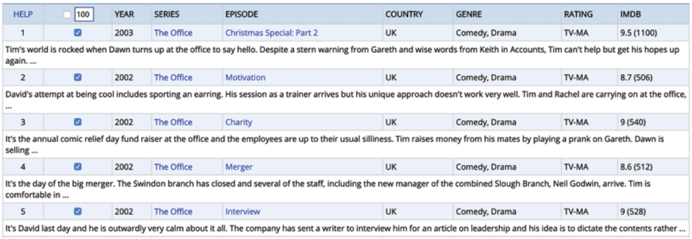
Figure 2. Partial list of texts of a Virtual Corpus in the Movies Corpus



SORT	Criteria	Values
○	Series title	Can use wildcards, e.g. *Star Trek*
●	Year	1990 - 2019
○	Genre	☑ Drama (41644) ☐ Comedy (31026) ☑ Crime (17068) ☐ Action (14314) ☐ Adventure (11908) ☐ Mystery (11244) ☐ Romance (8538) ☐ Animation (7309) ☐ Fantasy (6097) ☐ Family (5805) ☐ Sci-Fi (4481) ☐ Documentary (2728) ☐ Horror (2672) ☐ Thriller (2363) ☐ Reality-TV (1837) ☐ History (1606) ☐ Game-Show (1224) ☐ Music (1183) ☐ War (1153) ☐ Sport (575) ☐ Western (553) ☐ Biography (456) ☐ Talk-Show (268) ☐ News (230) ☐ Musical (187)
○	Country	☑ USA ☐ Canada ☐ UK ☐ Ireland ☐ Australia ☐ New Zealand ● Primary ○ Anywhere
○	TV rating	☐ TV-14 (18692) ☐ TV-PG (14204) ☑ TV-MA (7061) ☐ TV-G (1767) ☐ TV-Y7 (1720) ☐ TV-Y (392) ☐ PG (324) ☐ G (246) ☐ 12 (227) ☐ ATP (157) ☐ 13 (121) ☐ M (80) ☐ 16 (60) ☐ 15 (58) ☐ 6 (56) ☐ N/A (29373) ☐ NOT RATED (848) ☐ UNRATED (132) ☐ APPROVED (64)
○	IMDB rating	Low 6.5 - 10.0 High (Min # votes) 10
	Plot	kidnap* (words in episode plot)
	Word in text	(single word only)

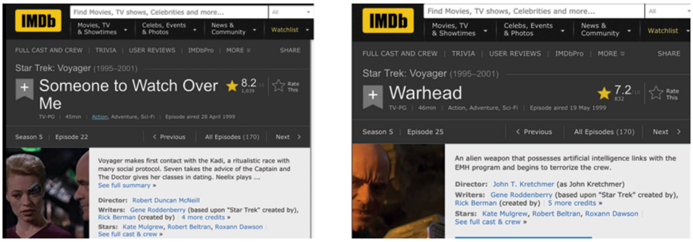
Figure 3. Creating a Virtual Corpus in the TV Corpus

Perhaps the most intuitive use of the metadata is to create a Virtual Corpus of a given TV show, such as *Star Trek: Next Generation*, *Doctor Who*, *Friends*, or *The Office* (UK). Figure 4 shows a partial listing of some episodes in a Virtual Corpus from *The Office* (UK, 2001–2003). Users can also click on any episode in the list to see the IMDb entry for that show, as in the two episodes of *Star Trek: Next Generation* shown in Figure 5.



HELP	☐ 100	YEAR	SERIES	EPISODE	COUNTRY	GENRE	RATING	IMDB
1	☑	2003	The Office	Christmas Special: Part 2	UK	Comedy, Drama	TV-MA	9.5 (1100)
Tim's world is rocked when Dawn turns up at the office to say hello. Despite a stern warning from Gareth and wise words from Keith in Accounts, Tim can't help but get his hopes up again. ...								
2	☑	2002	The Office	Motivation	UK	Comedy, Drama	TV-MA	8.7 (506)
David's attempt at being cool includes sporting an earring. His session as a trainer arrives but his unique approach doesn't work very well. Tim and Rachel are carrying on at the office, ...								
3	☑	2002	The Office	Charity	UK	Comedy, Drama	TV-MA	9 (540)
It's the annual comic relief day fund raiser at the office and the employees are up to their usual silliness. Tim raises money from his mates by playing a prank on Gareth. Dawn is selling ...								
4	☑	2002	The Office	Merger	UK	Comedy, Drama	TV-MA	8.6 (512)
It's the day of the big merger. The Swindon branch has closed and several of the staff, including the new manager of the combined Slough Branch, Neil Godwin, arrive. Tim is comfortable in ...								
5	☑	2002	The Office	Interview	UK	Comedy, Drama	TV-MA	9 (528)
It's David last day and he is outwardly very calm about it all. The company has sent a writer to interview him for an article on leadership and his idea is to dictate the contents rather ...								

Figure 4. Partial list of texts of a Virtual Corpus in the TV Corpus



Star Trek: Voyager (1995–2001)

Someone to Watch Over Me ★ 8.2/10 1,039

TV-PG | 45min | Action, Adventure, Sci-Fi | Episode aired 28 April 1999

Season 5 | Episode 22

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 ...
See full summary »

Director: Robert Duncan McNeill
Writers: Gene Roddenberry (based upon "Star Trek" created by), Rick Berman (created by) | 4 more credits »
Stars: Kate Mulgrew, Robert Beltran, Roxann Dawson | See full cast & crew »

Star Trek: Voyager (1995–2001)

Warhead ★ 7.2/10 832

TV-PG | 46min | Action, Adventure, Sci-Fi | Episode aired 19 May 1999

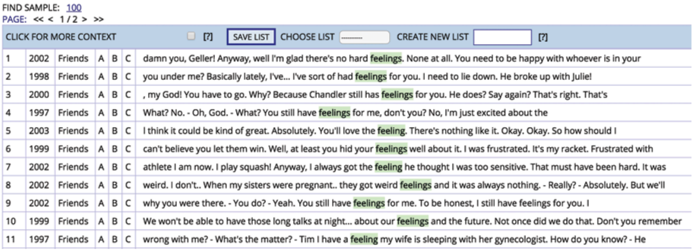
Season 5 | Episode 25

An alien weapon that possesses artificial intelligence links with the EMH program and begins to terrorize the crew.

Director: John T. Kretchmer (as John Kretchmer)
Writers: Gene Roddenberry (based upon "Star Trek" created by), Rick Berman (created by) | 5 more credits »
Stars: Kate Mulgrew, Robert Beltran, Roxann Dawson | See full cast & crew »

Figure 5. IMDb entry for texts in a Virtual Corpus

After creating a Virtual Corpus, users can delete entries from their Virtual Corpus, assign entries to user-defined categories (such as genre, time period, or country), or move or copy entries (texts) from one Virtual Corpus to another. The real value of the Virtual Corpora is that they allow users to limit their search to a particular set of movies or TV series or episodes. For example, they could search for the word *feeling*(*s*) in the TV series *Friends* (Figure 6). They could generate KWIC lines for the phrase *why don't* in the James Bond movies (Figure 7). Or they could search for collocates of *memory* in any *Star Trek* episode (Figure 8).

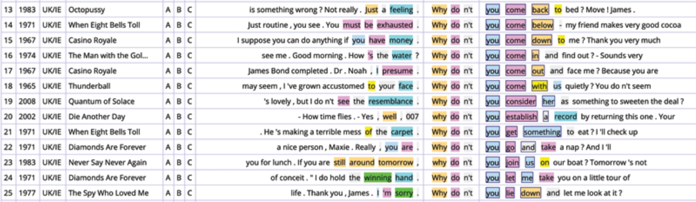


FIND SAMPLE: 100
PAGE: << < 1 / 2 > >>

CLICK FOR MORE CONTEXT □ [?] SAVE LIST CHOOSE LIST ---------- CREATE NEW LIST [?]

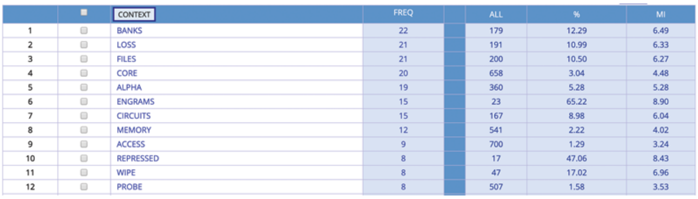
1	2002	Friends	A	B	C	damn you, Geller! Anyway, well I'm glad there's no hard feelings. None at all. You need to be happy with whoever is in your
2	1998	Friends	A	B	C	you under me? Basically lately, I've... I've sort of had feelings for you. I need to lie down. He broke up with Julie!
3	2000	Friends	A	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

Figure 6. KWIC entries from a Virtual Corpus: *feelings* in *Friends*



13	1983	UK/IE	Octopussy	A	B	C	is something wrong ? Not really . Just a feeling .	Why do n't	you come back to bed ? Move ! James .
14	1971	UK/IE	When Eight Bells Toll	A	B	C	Just routine , you see . You must be exhausted .	Why do n't	you come below - my friend makes very good cocoa
15	1967	UK/IE	Casino Royale	A	B	C	I suppose you can do anything if you have money .	Why do n't	you come down to me ? Thank you very much
16	1974	UK/IE	The Man with the Gol...	A	B	C	see me . Good morning . How 's the water ?	Why do n't	you come in and find out ? - Sounds very
17	1967	UK/IE	Casino Royale	A	B	C	James Bond completed . Dr . Noah , I presume .	Why do n't	you come out and face me ? Because you are
18	1965	UK/IE	Thunderball	A	B	C	may seem , I 've grown accustomed to your face .	Why do n't	you come with us quietly ? You do n't seem
19	2008	UK/IE	Quantum of Solace	A	B	C	's lovely , but I do n't see the resemblance .	Why do n't	you consider her as something to sweeten the deal ?
20	2002	UK/IE	Die Another Day	A	B	C	- How time flies . - Yes , well , 007	why do n't	you establish a record by returning this one . Your
21	1971	UK/IE	When Eight Bells Toll	A	B	C	. He 's making a terrible mess of the carpet .	Why do n't	you get something to eat ? I 'll check up
22	1971	UK/IE	Diamonds Are Forever	A	B	C	a nice person , Maxie . Really , you are .	Why do n't	you go and take a nap ? And I 'll
23	1983	UK/IE	Never Say Never Again	A	B	C	you for lunch . If you are still around tomorrow ,	why do n't	you join us on our boat ? Tomorrow 's not
24	1971	UK/IE	Diamonds Are Forever	A	B	C	of conceit . " I do hold the winning hand .	Why do n't	you let me take you on a little tour of
25	1977	UK/IE	The Spy Who Loved Me	A	B	C	life . Thank you , James . I 'm sorry .	Why do n't	you lie down and let me look at it ?

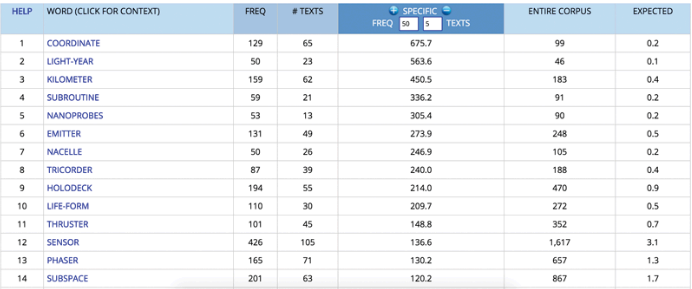
Figure 7. Re-sortable KWIC entries from a Virtual Corpus: *why don't* in James Bond movies



	□	CONTEXT	FREQ		ALL	%	MI
1	□	BANKS	22		179	12.29	6.49
2	□	LOSS	21		191	10.99	6.33
3	□	FILES	21		200	10.50	6.27
4	□	CORE	20		658	3.04	4.48
5	□	ALPHA	19		360	5.28	5.28
6	□	ENGRAMS	15		23	65.22	8.90
7	□	CIRCUITS	15		167	8.98	6.04
8	□	MEMORY	12		541	2.22	4.02
9	□	ACCESS	9		700	1.29	3.24
10	□	REPRESSED	8		17	47.06	8.43
11	□	WIPE	8		47	17.02	6.96
12	□	PROBE	8		507	1.58	3.53

Figure 8. Collocates from Virtual Corpus: *memory* in *Star Trek*

In just one to two seconds, users can also generate "keywords" from a Virtual Corpus, as with the noun keywords from *Star Trek: Next Generation* shown in Figure 9. (The keywords are generated by comparing the words in the Virtual Corpus to the rest of the TV or Movies corpus; similar to log likelihood comparisons.) Users can then click on any of these keywords to see the KWIC lines for that word in just the Star Trek Virtual Corpus (or of course any Virtual Corpus that they have created).



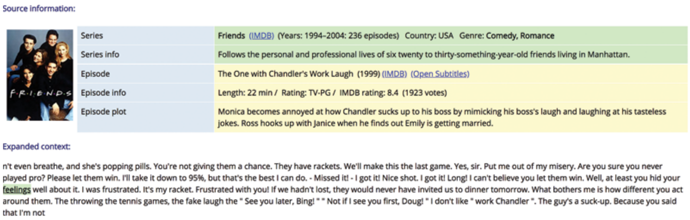
HELP	WORD (CLICK FOR CONTEXT)	FREQ	# TEXTS	SPECIFIC FREQ 50 5 TEXTS	ENTIRE CORPUS	EXPECTED
1	COORDINATE	129	65	675.7	99	0.2
2	LIGHT-YEAR	50	23	563.6	46	0.1
3	KILOMETER	159	62	450.5	183	0.4
4	SUBROUTINE	59	21	336.2	91	0.2
5	NANOPROBES	53	13	305.4	90	0.2
6	EMITTER	131	49	273.9	248	0.5
7	NACELLE	50	26	246.9	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

Figure 9. Keyword list from a Virtual Corpus: *Star Trek Voyager*

In addition to limiting searches to particular groups of movies, TV series, or TV episodes, it is also possible to compare across one's own Virtual Corpora. For example, one could compare the frequency of the word *love* in *Friends* or *The Office* or *Seinfeld*, or the frequency of a form of *kill* in movies from the 1930s or 1950s, or US Westerns from the 1950s–1960s, or R-rated crime movies from the 1990s, or all of the James Bond movies.

All of the preceding examples show how the IMDb metadata can be used to create Virtual Corpora, which is essentially a "corpus within a corpus". Previously, researchers needed to somehow find, download and clean all of the episodes of a given TV show (or set of movies) by themselves, and then begin the entire process again if they wanted to compare that to another set of data. With the TV and Movies Corpora, they can create these corpora in several seconds. This feature should be of interest to corpus linguistic research, which has often analyzed a particular series or franchise, such as *Friends* (Quaglio, 2009) or *Star Trek* (Csomay & Young, this issue).

Another use of the IMDb metadata is to simply see more information about a certain movie, series, or episode, from within the KWIC view. Users can click on any entry to see an "Expanded KWIC display" for the word or phrase, as in Figure 10. But in addition, they can see what the episode or movie was about, which might provide useful information on why a particular word or phrase or construction was used.



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

Figure 10. Metadata in expanded KWIC display

The use of metadata to create Virtual Corpora for particular TV series and movies showcases another potential use of the TV and Movies corpora: to study 'telecinematic discourse' (Piazza et al., 2011) in its own right (see Introduction to this issue). This allows us to study language use in specific series, movies, or genres, to analyze variation over time (see Werner, this issue), or to use the corpora as baseline against which other television series or movies can be compared (see Reichelt, this issue).

5. Informal nature of the language in the TV and Movies corpora

As was discussed in Section 2, one purpose of the TV and Movies corpora is to provide data on very informal language – hopefully similar to the type of data that is available from sources like the BNC-Spoken. As this section will show, in many cases the TV and Movies data is in fact quite comparable to BNC-Spoken, in terms of its informality. This would seem to support the findings of the psycholinguistic experiments that were discussed in Section 2, which show that people recognize the language of subtitles as being more "everyday" and "familiar" than the data from actual spoken corpora like that in the BNC.

In terms of lexical data, Table 2 shows examples of phrases that are more common in the TV and Movies corpora than in BNC-Spoken. In each case, the table shows the search string (for the version of the BNC1994 at English-Corpora .org), a sample sentence, the raw frequency and normalized frequency (per million words; pmw) in both the BNC-Spoken and the TV Corpus, as well as a number (the rightmost column) showing how much more frequent the word is in the TV Corpus than in BNC-Spoken. (Similar data was found in the Movies Corpus, but for reasons of space, only the TV Corpus data is shown here.) Crucially, the TV data is just for the 7.3 million words of data from the UK in the 1980s and 1990s in the 325-million-word TV Corpus, which permits a good comparison to the BNC1994. For example, the normalized frequency of [, OK/okay?] is about nine times as frequent in the TV corpus than it is in BNC-Spoken.

Table 2. Frequency of informal phrases in BNC-Spoken and TV Corpus

Search string	Example	BNC	BNC pmw	TV	TV pmw	TV/BNC
my God	My God – she's horrible!	572	57.4	991	135.8	2.4
, ok|okay?	we're leaving now, OK?	344	34.5	439	60.1	1.7
I told you	I told you to leave	1,252	12.52	687	94.1	7.5
, right?	You're pretty tired, right?	274	27.5	602	82.5	3.0
. it 's ADJ.	. It's sad. She's gone now	126	12.7	561	76.8	6.1
do n't leave	Don't leave! I need you	39	3.9	76	10.4	2.7
. Get out	. Get out right now!	23	2.3	155	21.2	9.2
hand me * NOUN	Hand me a towel.	2	0.2	155	2.1	10.3

The last three rows are particularly interesting. Each of these are very much oriented towards the "here and now" (aligning with findings on 'discourse immediacy' and 'interaction in the here-and-now' reported in Quaglio, 2009; Bednarek, 2018, respectively, for US television series). The fact that they are more common in the TV Corpus than the BNC-Spoken shows that the TV Corpus is highly situational – rather than more abstract and theoretical discussion of politics or other current events, such as what one might find in COCA-Spoken.

Evidence for the highly informal nature of the corpora extends to syntax as well. For example, Figure 11 shows the normalized frequency (per million words) of the progressive (BE _v?g; e.g. *I was talking to someone*) in the 1980s–1990s UK portion of the TV and Movies corpora (these sections were selected so that they would be comparable to the BNC both for country and time period). It also shows the normalized frequency in the five million words of BNC-Conversation ("BNC SPOK +C" in the chart; what www.natcorp.ox.ac.uk/corpus/creating.xml calls "Spoken: Demographic") and the five million words of BNC: Context-Governed ("BNC SPOK –C" e.g. courtroom, classroom, or sermons; see www.natcorp.ox.ac.uk/corpus/creating.xml). Finally, it shows the frequency of the progressive in the three other "macro-genres" of the BNC (fiction, newspapers, and academic), as well as the 125 million words COCA-Spoken (which is taken from unscripted conversations on national TV and radio programs).

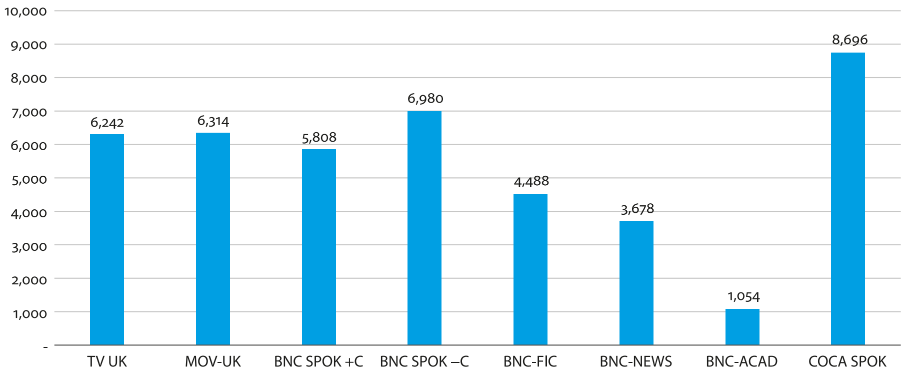


Figure 11. Frequency of progressive constructions

As the data in Figure 11 indicates, the progressive is a feature of more informal language. In the BNC, it occurs the most in spoken, and then fiction, newspaper, and (least of all) in academic (this compares well with the data in Biber et al., 1999: 461–463). The most important data from this figure is that the frequency of the progressive in TV and Movies (again, limited just to UK 1980s–1990s) places it between BNC: Conversation and BNC: Context-Governed.

Conversely, the passive with *be* (BE _v?n; e.g. *the country* *was colonized* *in the 18th century*) occurs the least in spoken, and then fiction, news, and (most frequently) in academic (see Figure 12). This again agrees with the data in Biber et al. (1999: 475–481). And again, the TV and Movies data (UK, 1980s–1990s) patterns fairly well with BNC-Spoken; its frequency places it between BNC: Conversation and BNC: Context-Governed (and certainly closer to BNC: Conversation in the case of the Movies corpus).

Finally, consider the frequency of NOUN + NOUN (e.g. *county council*, *car park*, *back door*, *washing machine*, *living room*, *dinner time*) in the various sections of the BNC and in the TV and Movies corpora shown in Figure 13. As Biber et al. (1999: 589–594) note, this is more common in newspaper texts (due to space constraints) and academic texts than in fiction and spoken, and the data from the BNC agrees with this quite well. Most importantly for our purposes here, we see that the frequency of NOUN + NOUN in the TV and Movies corpora patterns more with BNC: Conversation than with BNC: Context-Governed, and certainly more than with COCA-Spoken or the other genres of the BNC.

As the data in Table 2 indicates, the TV and Movies corpora are very informal in terms of phraseology, and Figures 11–13 show that the data from the TV and Movies corpora patterns well with BNC-Spoken in terms of syntax. Obviously, the

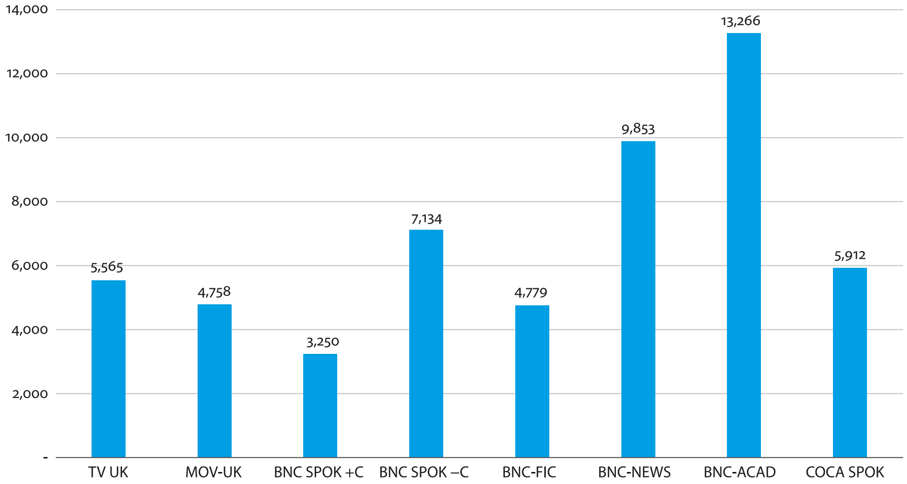


Figure 12. Frequency of passive constructions

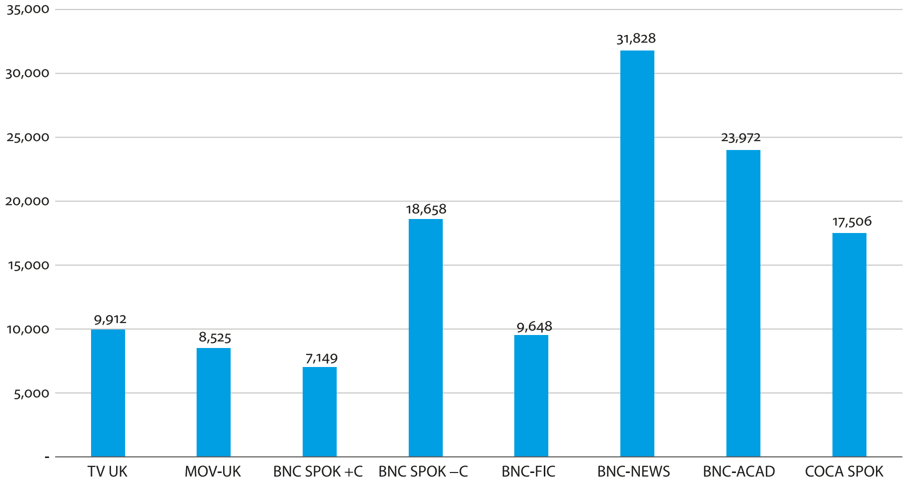


Figure 13. Frequency of NOUN + NOUN constructions

TV and Movies language is scripted, rather than being naturally occurring conversation. And yet it is quite striking how close the scriptwriters were to actual conversation (at least as measured by BNC: Conversation).

In a sense, this is probably not overly surprising. As Levshina (2017) has shown, subtitles contain many features of involved informal communication and are "remarkably close to real informal language" (Levshina, 2017: 336). Imagine if a contemporary TV script had a character saying *with <u>whom</u> did you go out last night*, or *we <u>must</u> go now*, or *Who is it? It's <u>I</u>*. It's hard to imagine even getting the actor to repeat these lines, without it sounding extremely formal and awkward. Scriptwriters are fairly sophisticated, and they will write scripts that model actual conversation quite well, and that is reflected in the TV and Movies subtitles data (for insights into scriptwriters' language awareness, see the interviews with Hollywood TV writers in Bednarek, 2019). The results also partially align with previous work on US television dialogue that analyzes transcripts rather than subtitles and is based on much smaller datasets (e.g. Bednarek, 2018) and/or individual series (e.g. Quaglio, 2009). For instance, some of the informal phrases listed in Table 2 (*my god*, *it's okay*, *told you*) were identified as "key" in US television dialogue compared to unscripted American English in Bednarek (2018), while Quaglio (2009) has suggested that the dialogue of the sitcom *Friends* is more informal than unscripted American conversation. These overlaps confirm Levshina's (2017: 330) assumption that there are similarities between subtitles and transcripts. However, a full comparison of informality in subtitles compared to transcripts or of informality in different series or types of TV narratives and movies is beyond the scope of this article.

6. Dialectal and historical variation in English

One issue with many spoken corpora is that they are often limited in terms of time and space. An advantage of the TV and Movies corpora is that they contain data from several different dialects and time periods (decades), extending back to the 1950s (TV) and the 1930s (Movies). Tables 3 and 4 summarize the amount of data for the different countries and decades. (Note that Misc. includes co-productions from other countries.)

Table 3. Size of Movies Corpus by country and decade

Movies	US/CA	UK/IE	AU/NZ	Misc.	Total
1950s	2,012,631	20,740	–	–	2,033,371
1960s	6,728,110	2,168,841	–	5,727	8,902,678
1970s	5,717,836	3,063,468	–	–	8,781,304
1980s	11,905,793	3,054,673	49,263	1,814	15,011,543
1990s	26,825,820	4,373,746	78,769	228,645	31,506,980
2000s	71,570,270	14,511,570	997,291	464,778	87,543,909
2010s	141,039,715	25,959,596	4,015,203	1,406,977	172,421,491
Total	**265,800,175**	**53,152,634**	**5,140,526**	**2,107,941**	**326,201,276**

Table 4. Size of TV Corpus by country and decade

TV	US/CA	UK/IE	AU/NZ	Misc.	Total
1930s	6,013,722	445,980	2,245	104,255	6,566,202
1940s	8,679,722	1,077,429	–	51,151	9,808,302
1950s	8,570,819	1,826,174	21,777	197,173	10,615,943
1960s	5,851,067	2,687,175	6,594	557,976	9,102,812
1970s	6,972,688	2,060,309	112,715	958,968	10,104,680
1980s	10,739,129	2,153,349	308,640	917,461	14,118,579
1990s	19,259,078	2,983,322	384,607	1,986,577	24,613,584
2000s	38,572,824	6,970,252	793,610	4,893,749	51,230,435
2010s	48,649,187	8,705,479	1,337,876	4,626,223	63,318,765
Total	**153,308,236**	**28,909,469**	**2,968,064**	**14,293,533**	**199,479,302**

6.1 Dialectal differences

The 525 million words of data (from TV and Movies combined) is more than 100 times as much data as the spoken corpora (for multiple countries) in other corpora, such as in the International Corpus of English (ICE; Greenbaum, 1996). Of course, the data in ICE is from actual spoken English. Because the corpus has been very carefully designed and constructed, it offers some advantages over the TV and Movies subtitles. On the other hand, the much larger TV and Movies corpora allow a wide range of searches – especially lexically oriented searches – where a small two to three-million-word corpus (e.g. the combined spoken sections from the UK, Ireland, Australia, and New Zealand in ICE) would be quite inadequate.

As Baker (2009, 2011) notes, there is often not enough data in a small two to four-million-word corpus to look at lexical phenomena, such as what words are more common in one country than another. But with the TV and Movies subtitles corpora, this is quite easy to do. For example, the 266 million words of data from the US and the 53 million words of data from the UK in the TV corpus allows us to find those words that are at least 10 times as frequent in one dialect than in the other (Table 5). (Table 5 also shows that there are spelling differences between the different countries' sections of the corpus – e.g. in the NOUN row: *mom* vs *mum* – something users should keep in mind when searching the whole corpus for particular words.)

Table 5. Informal words in American and British sections of the TV Corpus

	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-ass, sketchy, wordy, charmed, disoriented, kick-ass, 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, ass, movie, buddy, apartment, truck, chef, buck, dollar, sweetie, mommy, attorney, mayor, butt, cookie, grandma, asshole, 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, bollock, solicitor, railway, vicar, telly, guv, grandad, petrol, ladyship, mammy, shilling, maths, lorry, arsehole, 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, wank, queue, burgle, snigger, snog, plod, splutter, clamber

6.2 Change over time

The TV and Movies corpora can also be used to look at language change (TV: 1950–present; Movies: 1930–present). Other corpora such as the Corpus of Historical American English (COHA; Davies, 2012) allow us to look at hundreds of millions of words of data from the past 200 years. (COHA has 400 million words from 1810–2009 and more than 200 million words from just the 1930s to the 2000s.) But COHA doesn't really have any "spoken" texts. The TV and Movies corpora, however, provide us with more than 525 million words of highly informal language from the 1930s-2010s. As the data in Table 6 indicates, this allows us to find words that are at least 10 times as frequent in texts from the 1930s–1960s (left) and the 1990s–2000s (right).

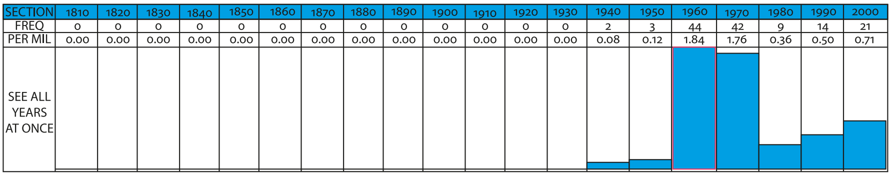
Table 6. Informal words in 1930s–1960s and 1990s–2010s sections of Movies Corpus

	More common 1930s–1960s	More common 1990s–2010s
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	fucking, okay, cool, weird, damn, goddamn, huge, awesome, pregnant, super, sexy, scary, unbelievable, sexual, boring, pathetic, gross, massive, nuclear, creepy, global, creative, magical, intense, ultimate, shitty, 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	shit, hell, mom, fuck, ass, bitch, dude, sex, drug, asshole, TV, bullshit, motherfucker, bastard, girlfriend, relationship, dick, computer, video, tape, crap, bro, pussy, nigger, 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, scram, furnish, sock, darn, consent, tangle, fuss, peddle, double-cross	fuck, suck, screw, piss, focus, freak, date, rape, pee, film, score, bitch, shit, chill, define, stress, evolve, fart, activate, surf, tape, participate, process, monitor, target, manipulate, trigger, puke, initiate, generate

Note that many of these words from the 1990s–2010s may have been more frequent in earlier decades in actual speech, but censorship on movies and TV shows in earlier periods means that they simply don't appear in the corpora. For additional insights into this matter, Werner (this issue) investigates changes in the frequency of swear words in the TV and Movie corpora over time.

Another advantage of very large, informal corpora in terms of looking at lexical change relates to granularity. As is discussed in Davies (2018), lexical change can occur quite fast, and to catch relevant developments it is often not sufficient to sample the language only every 25–30 years, such as in 1931, 1961, and 1991 (as with the Brown family of corpora) or in the late 1980s and then again in 2014 (as with the BNC1994 and BNC2014). Any changes that take place in between these years are essentially "invisible", and in terms of lexical change, this is often too long of a gap.

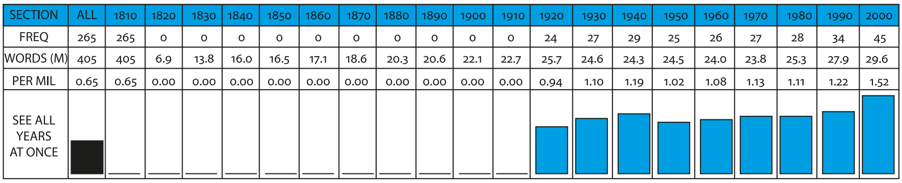
Let us briefly consider two examples related to granularity, which are representative of any number of words over time. First, let us consider the frequency for *groovy* in COHA, as shown in Figure 14.



SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	44	42	9	14	21
PER MIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.12	1.84	1.76	0.36	0.50	0.71

Figure 14. Frequency of *groovy* by decades in COHA

Imagine that our two corpora contained texts 30 years apart – from 1955 and 1985. In this case, it would appear (based on the COHA data from the 1950s and the 1980s) that *groovy* is on the increase. While it has increased slightly in these 30 years, we would miss entirely the steep decrease from the 1960s/1970s to the 1980s. Second, consider the case of *normalcy*, shown in Figure 15.



SECTION	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	265	265	0	0	0	0	0	0	0	0	0	0	24	27	29	25	26	27	28	34	45
WORDS (M)	405	405	6.9	13.8	16.0	16.5	17.1	18.6	20.3	20.6	22.1	22.7	25.7	24.6	24.3	24.5	24.0	23.8	25.3	27.9	29.6
PER MIL	0.65	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	1.10	1.19	1.02	1.08	1.13	1.11	1.22	1.52

Figure 15. Frequency of *normalcy* by decades in COHA

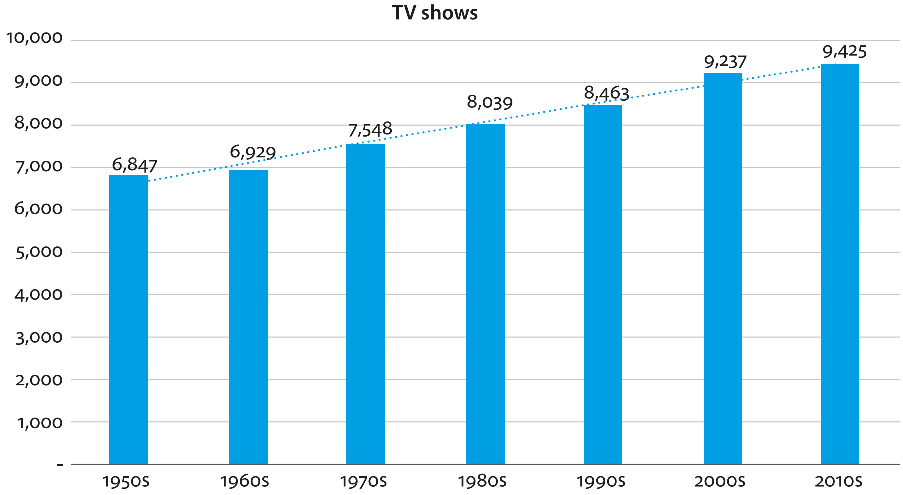
This word was famously "rescued" from obscurity by President Warren G. Harding in 1920, who (according to purists) mistakenly used it instead of the more "correct" *normality*. The word caught on with a public tired of World War I and other foreign involvements, and Harding went on to win the election. But imagine that we only had two corpora from 1915 and 1935 (roughly the same amount of time as with the BNC1994 and the BNC2014). There would obviously be a large increase in frequency between 1915 and 1935, but there would be no way to know if that predated Harding, whether his campaign caused the increase in usage, or whether it was after his time. In summary, corpora that have texts that are spaced decades apart may be adequate for looking at much more gradual grammatical change, but they are much more problematic in looking at lexical change, which can occur quite suddenly.

There is no such problem with the TV or Movies data. As the data in Table 7 shows, there are no "gaps" in the data from year to year. This table shows the number of words in the TV Corpus for each year from 1987 (roughly when the BNC1994 began to be created) through the next 30 years – a total of 283 million words of data for these 30 years. And this is just for the TV corpus; there are an additional 140 million words of data from the Movies corpus for this same 30-year period.

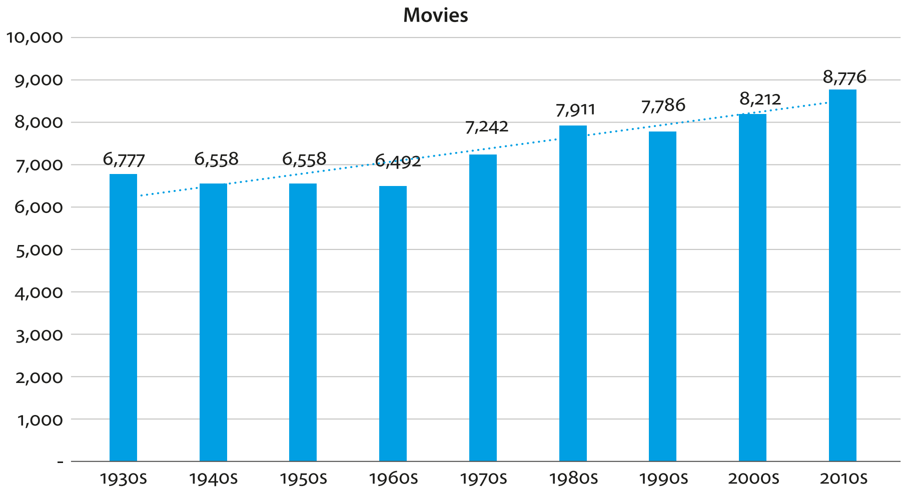
Table 7. Number of words in TV Corpus by year, 1987–2016

1987	2,080,511	1997	3,821,834	2007	11,642,166
1988	1,715,698	1998	4,242,221	2008	11,137,597
1989	2,554,744	1999	4,505,438	2009	15,367,913
1990	1,968,905	2000	4,590,593	2010	19,205,273
1991	2,135,182	2001	5,506,332	2011	21,167,200
1992	2,181,034	2002	6,131,648	2012	21,854,565
1993	2,466,673	2003	6,672,996	2013	22,377,615
1994	3,055,304	2004	7,468,196	2014	23,022,413
1995	3,474,276	2005	9,094,251	2015	24,793,373
1996	3,656,113	2006	9,932,217	2016	25,077,851

In addition to lexical change, the corpora can also be used to look at many other types of linguistic change, such as syntactic change. For example, Figure 16 shows the frequency of the progressive over time (the data labels indicate the normalized frequency per million words in each decade, and this is based on 2,963,000 tokens in the TV corpus and 1,590,000 tokens in the Movies corpus). As was discussed previously (see Figure 11), the progressive occurs more in informal genres. Data from 3,241,000 tokens in COHA (Figure 17) also shows that the progressive is increasing overall, at least in American English.



a.



b.

Figure 16. Frequency of the progressive construction by decade in TV and Movies corpora

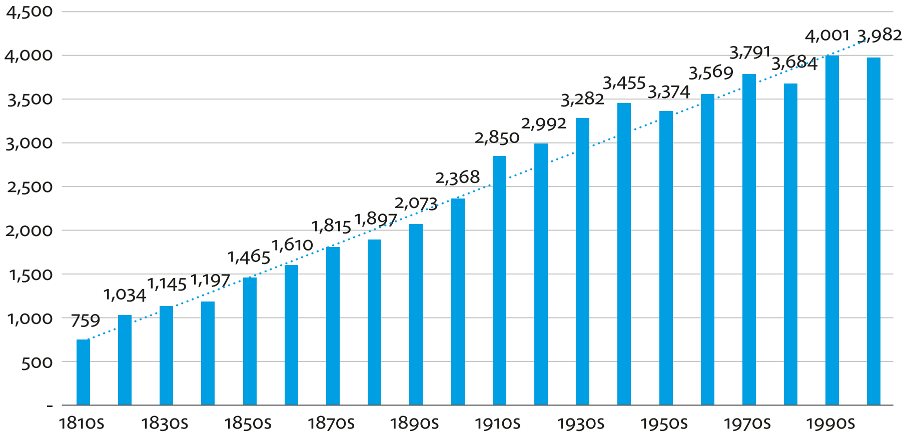


Figure 17. Frequency of progressive construction by decade in COHA

Both the TV and Movies data, as well as the COHA data, show that the progressive is becoming more frequent over time. (Note also that for every decade, the frequency is much higher in the TV and Movies corpora than in COHA, which is to be expected, since these corpora are more informal than COHA overall. In addition, the progressive is much more prominent in speech, which is not centrally represented in COHA.) It appears that the TV and Movies corpora probably reflect quite well the changes that were actually occurring in the language during this time. Additional evidence for increasingly informal language comes from the passive construction. As was discussed previously (see Figure 12), the passive occurs less in informal genres. Data from 3,241,000 tokens in COHA (Figure 18) also shows that the BE passive is decreasing overall, at least in American English. Figure 19 from the TV and Movies corpora is based on 1,415,000 tokens of the passive in the TV corpus and 786,000 tokens in the Movies corpus.

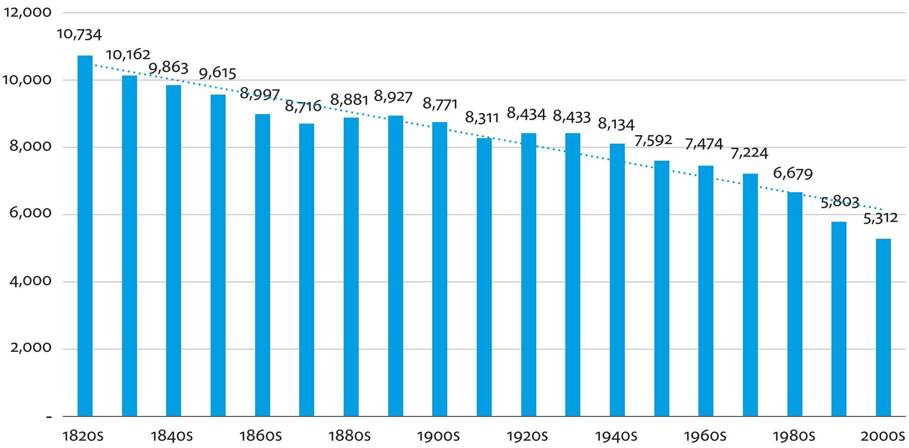
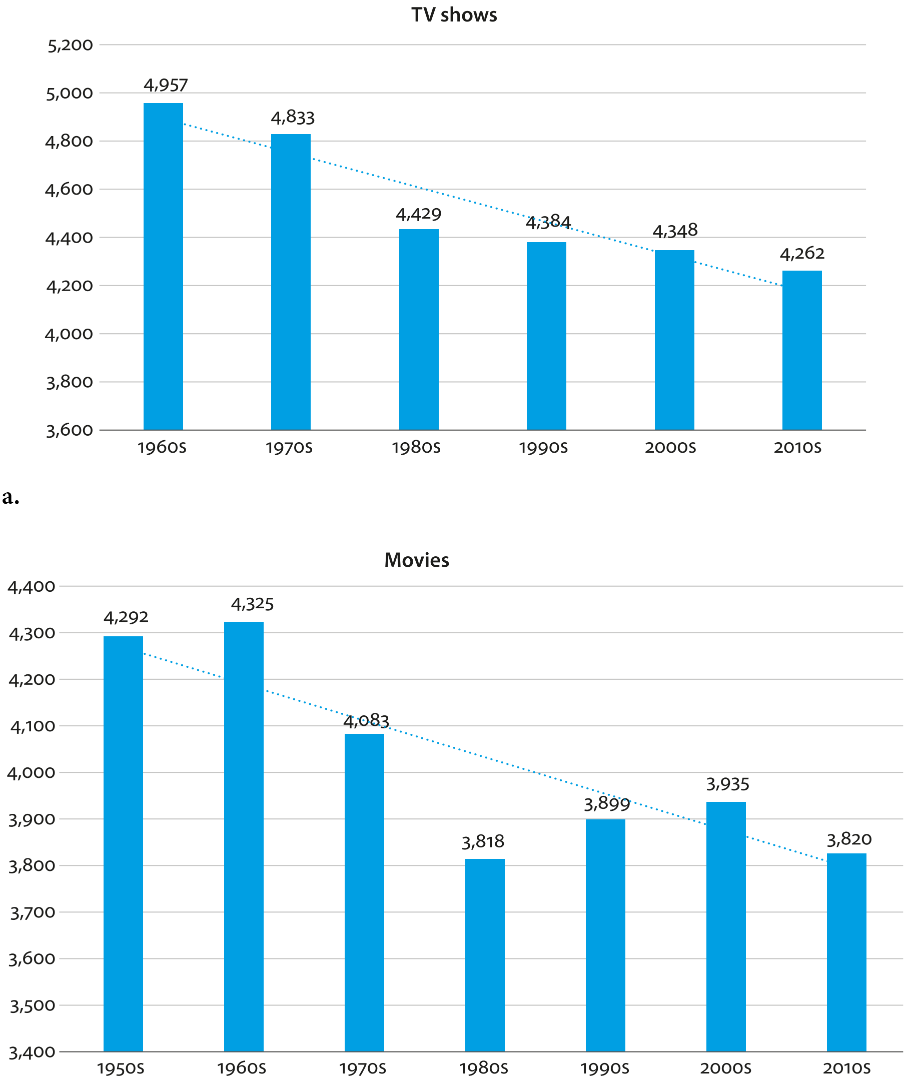


Figure 18. Frequency of passive constructions by decade in COHA

The data shows that the passive is becoming less common over time, which closely agrees with the data from COHA. (Note also that for every decade, the frequency is much lower in the TV and Movies corpora than in COHA, which is to be expected, since this is more informal language than COHA overall.) Again, the TV and Movies corpora probably reflect quite well the changes that were actually occurring in the language during this time. These corpora can thus also be used to confirm and further probe results from sociolinguistic studies that investigate linguistic innovation and change (based on limited data), which have proposed that television dialogue reflects and sometimes enhances ongoing language change (see overview in Bednarek, 2018: 28–31). On the other hand, the TV and Movies corpora can also be a basis for analyzing whether telecinematic discourse itself is a dynamic or stable variety (see Veirano Pinto, 2014; Werner, this issue).

7. Conclusion

Subtitles data from movies and TV shows provide us with the ability to obtain large amounts of informal data at a very low cost. It can be quite expensive to



a.

b.

Figure 19. Frequency of passive constructions by decade in TV and Movies corpora

create a good spoken corpus with everyday conversation, which is evidenced by the fact that most spoken corpora are quite small (one to two million words, as with the Switchboard (Godfrey & Holliman, 1993) or CALLHOME corpora from the Linguistic Data Consortium). And such corpora are quite limited in terms of the phenomena that they can consider (see Davies, 2015). Larger spoken corpora like that of the British National Corpus or the International Corpus of English can be extremely expensive to collect, clean, and transcribe. But even here, the corpora are rather small – five million words of informal conversation in the BNC1994, and less than 2.5 million words of speech in the ICE corpora from the US, Canada, UK, Ireland, Australia, and New Zealand combined.

TV and Movies subtitles corpora are essentially the best of all worlds. As Section 5 indicates and other research confirms, they model conversation very well. But they are extremely inexpensive to create – basically just the time involved in downloading and categorizing the data, as was discussed in Section 3. And they offer a huge advantage over actual conversation, in that the subtitles data can be (and in fact is) much larger than in actual conversation. For example, the TV and Movies corpora are (respectively) about 20 times and 12 times as large as BNC-Conversation (the combined total from both the BNC1994 and BNC2014), and the disparity is even greater for ICE.

Obviously, the subtitles data are not a perfect substitute for the actual spoken language in these other corpora. For example, it is possible that there are some features of actual speech, such as dysfluencies, hesitations, errors, repairs, syntactic blends, prefaces, and tags (see Biber et al., 1999: 1037–1126) that may not appear as much in the subtitles data as in actual speech, or which have a different distribution. Subtitles are limited by spatial constraints and condense or cut portions of dialogue, which can affect various interpersonal and stylistic features such as discourse markers, formulaic politeness expressions, hesitations, false starts, phatics, or sentential tags (Lugea, 2019). Levshina (2017) suggests that the language of subtitles is less vague, narrative and spontaneous, but more dynamic and emotional than unscripted language. We will leave it to future researchers to investigate this in more detail.

On the other hand, the immense size of the subtitles data means that we can look at a much wider range of linguistic phenomena with this data, as well as having huge amounts of informal data to look at language change and dialectal variation. In summary, both the actual spoken data and the subtitles data can be valuable tools to allow us to look at variation in very informal English. In addition, the TV and Movies corpora allow us to analyze telecinematic discourse (in the form of subtitles) in its own right, across countries and over time.

References

Baker, P. (2009). The BE06 corpus of British English and recent language change. *International Journal of Corpus Linguistics*, *14*(3), 312–337. https://doi.org/10.1075/ijcl.14.3.02bak

Baker, P. (2011). Times may change but we'll always have money: A corpus driven examination of vocabulary change in four diachronic corpora. *Journal of English Linguistics*, *39*(1), 65–88. https://doi.org/10.1177/0075424210368368

Bednarek, M. (2018). *Language and Television Series: A Linguistic Approach to TV Dialogue*. Cambridge University Press. https://doi.org/10.1017/9781108559553

Bednarek, M. (2019). *Creating Dialogue for TV: Screenwriters Talk Television*. Routledge. https://doi.org/10.4324/9780429029394

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.

BNC Consortium. (2007). British National Corpus (version 3, BNC XML ed.). http://www.natcorp.ox.ac.uk

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Canavan, A., & Zipperlen, G. (1996). *CALLFRIEND American English-Non-Southern Dialect (LDC96S46)*. Linguistic Data Consortium https://catalog.ldc.upenn.edu/LDC96S46.

Canavan, A., Graff, D., & Zipperlen, G. (1997). *CALLHOME American English Speech (LDC97S42)*. Linguistic Data Consortium https://catalog.ldc.upenn.edu/LDC97S42.

Davies, M. (2009). the 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190. https://doi.org/10.1075/ijcl.14.2.02dav

Davies, M. (2011). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–465. https://doi.org/10.1093/llc/fqq018

Davies, M. (2012). Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English. *Corpora*, *7*(2), 121–157. https://doi.org/10.3366/cor.2012.0024

Davies, M. (2015). Corpora: An introduction. In D. Biber & R. Reppen (Eds.), *Cambridge Handbook of English Corpus Linguistics* (pp. 11–31). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.002

Davies, M. (2017). Using large online corpora to examine lexical, semantic, and cultural variation in different dialects and time periods. In E. Friginal (Ed.), *Studies in Corpus-Based Sociolinguistics* (pp. 19–82). Routledge. https://doi.org/10.4324/9781315527819-2

Davies, M. (2018). Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In C. Suhr, T. Nevalainen & I. Taavitsainen (Eds.), *From Data to Evidence in English Language Research* (pp. 34–55). Brill. https://doi.org/10.1163/9789004390652_004

Forchini, P. (2012). *Movie Language Revisited: Evidence from Multi-Dimensional Analysis and Corpora*. Peter Lang. https://doi.org/10.3726/978-3-0351-0325-0

Greenbaum, S. (1996). *Comparing English Worldwide: The International Corpus of English*. Clarendon Press.

Godfrey, J. J., & Holliman, E. (1993). *Switchboard-1 Release 2 (LDC97S62)*. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC97S62

Van Heuven, W., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Levshina, N. (2017). Online film subtitles as a corpus: An *n*-gram approach. *Corpora*, *12*(3), 311–338. https://doi.org/10.3366/cor.2017.0123

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16).* European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L16-1147/

Love, R. (2020). *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. Routledge.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, *22*(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Lugea, J. (2019). The intralingual subtitling of *The Wire*: Changes of style and substance. *Journal of Applied Linguistics and Professional Practice*, *12*(1), 23–49. https://doi.org/10.1558/jalpp.24620

Piazza, R., Bednarek, M., & Rossi, F. (Eds.) (2011). *Telecinematic Discourse: Approaches to the Language of Films and Television Series*. John Benjamins. https://doi.org/10.1075/pbns.211

Quaglio, P. (2009). *Television Dialogue: The Sitcom Friends vs. Natural Conversation*. John Benjamins. https://doi.org/10.1075/scl.36

Rayson, P., & Garside, R. (1998). The CLAWS web tagger. *ICAME Journal*, *22*(4), 121–123.

Simpson, R., Briggs, L., Ovens, J., & Swales, J. (2002). *The Michigan Corpus of Academic Spoken English*. The Regents of the University of Michigan.

Tiedemann, J. (2016). OPUS – parallel corpora for everyone. *Baltic Journal of Modern Computing*, *4*(2), 384.

Veirano Pinto, M. (2014). Dimensions of variation in North American movies. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional Analysis, 25 Years on: A Tribute to Douglas Biber* (pp. 109–146). John Benjamins. https://doi.org/10.1075/scl.60.04vei

Veirano Pinto, M. (2018). Variation in movies and television programs: The impact of corpus sampling. In V. Werner (Ed.), *The Language of Pop Culture* (pp. 139–161). Routledge. https://doi.org/10.4324/9781315168210-7

Address for correspondence

Mark Davies
Department of Linguistics
Brigham Young University
Provo UT 84602
USA

mark_davies@byu.edu

Publication history

Published online: 17 November 2020
Corrected: 13 January 2021

In the original Online-First version of this article published on 17 November 2020, statements about the size of the Spoken BNC2014 were incorrect. These have been updated in the current version of the article.