Analele Universității de Vest din Timișoara. Seria Științe Filologice
2024, Nr. 62/1, pp. 43-57
https://doi.org/10.35923/AUTFil.62-1.04

BUILDING *WITS*: CHALLENGES IN CREATING A CORPUS OF WORKPLACE SITCOMS

Karla CSÜRÖS

West University of Timișoara
karla.csuros@e-uvt.ro
https://orcid.org/0000-0002-9556-3724

Building *WITS*: Challenges in creating a corpus of workplace sitcoms

This article introduces the first version of the *Workplace Sitcoms Corpus (WITS)*, consisting of over 2 million words of dialogue lines extracted from eight contemporary American workplace sitcoms: *Ally McBeal* (1997-2002), *Scrubs* (2001-2010), *The Office* (2005-2013), *30 Rock* (2006-2013), *Parks and Recreation* (2009-2015), *Brooklyn Nine-Nine* (2013-2021), *Superstore* (2015-2021) and *Abbott Elementary* (2021-ongoing). The corpus aims to be representative of a specific subgenre of TV dialogue, with a focus on work-related discourses in fictional contexts. Following Bednarek's (2018) definition, television dialogue is understood as all linguistic utterances made by actors performing characters on TV series. The *WITS* corpus is presented in relation to other reference television corpora, such as *SydTV* (Bednarek, 2018) and the *TV Corpus* (Davies 2021). This paper details the corpus design stage, focusing on various selection criteria of titles that are categorized as workplace sitcoms, as well as the data collection and initial transcription stages of building the *WITS* corpus. It also highlights some of the key challenges in gathering a large amount of television dialogue data: difficulty in finding sources and means of assuring their accuracy, adjustable data cleaning techniques, proper annotation methods and others. Notably, in line with Quaglio's (2008) findings, I argue that fan transcriptions are the most reliable source of data when it comes to television dialogue, with some noteworthy caveats that will impact my further research.

KEYWORDS:

American television; comedy series; corpus design; telecinematic discourse; workplace sitcom.

Introduction

Television sitcoms, particularly those set in workplace environments, can provide a rich source of data for linguistic and discourse analysis due to their reliance on dialogue to convey humor, character development, and narrative progression. In this context, this paper introduces the *Workplace Sitcoms Corpus (WITS)*, a newly developed corpus comprising over 2 million words of dialogue from eight prominent American workplace sitcoms. This corpus is designed to be representative of a specific subgenre of television dialogue, with a focus on discourses related to work and professional environments in fictional settings. By building this corpus, I aim to provide a resource for examining the distinctive language features and conversational structures that characterize workplace sitcoms, contributing to broader linguistic studies of television dialogue. This particular paper outlines the methodological framework behind the creation of the corpus, including the selection criteria for the shows, data collection processes, and initial transcription stages. Furthermore, I discuss several key challenges encountered during the construction of the *WITS* corpus, including difficulties in sourcing reliable transcripts, implementing accurate data-cleaning techniques, and developing appropriate annotation strategies. These considerations are essential for ensuring the accuracy and reliability of the *WITS* corpus as a resource for linguistic analysis.

1. Corpus-based incursions into television dialogue

Television dialogue, as defined by Bednarek (2018, 6–9), refers to all speech uttered on-screen by characters or narrators, as part of serialized scripted fictional narratives. This includes voice-overs, monologues, dialogues, and interactions between multiple speakers. When analyzing television dialogue from a linguistic point of view, any screen directions or extra-linguistic elements (e.g., location, time, special effects, character movements or actions, etc.) are generally discarded. The main interest that linguists may have in television dialogue, as Bednarek points out, stems from its design principles; i.e., television dialogue is "designed to be spoken as if not written" (2018, p. 19), meaning that it is scripted in a way that reflects spoken language.

Even so, there are undoubtedly some features of television dialogue that distinguish it from unscripted spoken language. Richardson (2010, p. 6) states that television dialogue is designed to be "easily listened to"; similarly, Bednarek (2018, p. 19–20) describes it as "intelligible, accessible, and comprehensible" to audience members. This entails a certain fluency and "unnatural coherence and focus" (Toolan 2011, 181) of television dialogue that is generally not found in unscripted language, meaning that there are less speech errors, overlaps, interruptions or hesitations (Bednarek, 2018, p. 20).

Many other differences between television dialogue and natural language, however, can be attributed to the design principles of available spoken corpora, as per Bednarek (2018, p. 21). To expand on this, unscripted data is not collected from intimate settings in which interlocutors may become emotional (Dose, 2013). Speakers may also choose not to use as many taboo words when they are being recorded as in their daily lives (McEnery *et al.*, 2000). Finally, recordings of unscripted conversations generally do not include the initial and final greetings between speakers (Quaglio, 2009).

Despite this, television dialogue is not wholly inauthentic to natural language. One function of television dialogue, as discussed by Bednarek (2012b, p. 43) and Quaglio (2009, p. 120), is to reflect natural language, particularly everyday conversations, in a realistic manner. Moreover, television dialogue is bound by the same linguistic resources and principles as natural language, which implies the existence of a certain similarity between the two (Richardson, 2010, p. 106). There are several studies that have successfully compared television dialogue to unscripted language through multiple approaches (Al-Surmi, 2012; Bednarek, 2010, 2011, 2012a; Dose, 2013; Gregori-Signes, 2017; Heyd, 2010; Quaglio, 2008, 2009; Sardinha & Pinto, 2017). In large, television dialogue is believed to exemplify "core linguistic features that typify [natural] conversation" (Quaglio, 2009, p. 68), while also containing more emotional, informal and customary language (e.g., greetings, pleasantries). Sitcoms have also been found to be the television genre most similar to natural conversations (Al-Surmi, 2012; Heyd, 2010; Quaglio, 2009). However, none of the above-mentioned studies are replicable, in the sense that their television corpora are not widely available.

1.1. Corpora of television dialogue

Despite the existence of multiple studies (discussed above) that focus on analyzing the linguistic dimension of one or several TV series, there are very few existing TV corpora that are accessible for educators. The largest corpus, in both size and scope, is the aptly named *TV Corpus*, containing over 325 million words extracted from the subtitles of 75 thousand episodes that aired over the past 75 years. Although the *TV Corpus* was built from the best-rated subtitles readily available online (Davies, 2021, p. 13–14), its substantial size means that the accuracy of its data cannot be fully guaranteed. Furthermore, the *TV Corpus* does not contain any information about the speakers, only the dialogue lines themselves. The *TV Corpus* is available for consultation and paid offline download on the website english-corpora.org.

SydTV ("The Sydney Corpus of Television Dialogue") was created to ensure high accuracy of data and a great level of representativeness for TV dialogue as a variety of English. *SydTV* comprises around 275 thousand words from sixty-six shows, with one episode per show being selected. Bednarek (2018, p. 81–85) aimed to strike a balance between drama and comedy series, as well as between "mainstream" and "quality" shows. She also took into consideration the textual time, i.e., she processed a mix of pilots, finales, and episodes throughout the shows' first seasons. Frequency and keyness lists, as well as examples of linguistic analyses done with *SydTV* are available on the website syd-tv.com.

Other examples of TV corpora, this time representative of specific genres, are *SOAP* and *CATS*. *SOAP*, also known as the "Corpus of American Soap Operas", consists of 100 million words extracted from the scripts of ten soap operas (Davies, 2011). Like the *TV Corpus*, *SOAP* is accessible on english-corpora.org. On the other hand, *CATS* ("Corpus of American Television Series") contains approximately 160 thousand words from twenty-eight episodes taken from four drama series. Dose (2012, p. 106–109) used fan transcripts to build *CATS*, annotating the data with extra information related to character actions and scene settings. The *CATS* corpus, as opposed to other TV corpora, is not yet accessible online.

To sum up, there are TV corpora meant to be representative of TV dialogue as a whole (*TV Corpus*, *SydTV*) or of a particular genre (*SOAP* for soap operas, *CATS* for drama series). The manner of data collection for such corpora also varies greatly: some use subtitles, while others use (fan) transcripts. To the best of my knowledge, there have not been any corpora built specifically to examine the linguistic characteristics of workplace sitcoms as a television (sub)genre, despite its consistent worldwide success (Schneider, 2023). To fill this gap, my project's aim is to design and collect the *WITS* ("**w**orkplace s**it**com**s**") corpus, meant to be a representative sample of English language dialogue spoken in workplace sitcoms, i.e., comedies that primarily take place in the workplace, as defined by Charney (2005).

2. Corpus design

The following section details the preliminary corpus design stage of *WITS*, focusing on various selection criteria of titles that are included in the *WITS* corpus, as well as the data collection and initial transcription stages of building the *WITS* corpus. It also highlights some of the key challenges in gathering a large amount of television dialogue data: difficulty in finding sources and means of assuring their accuracy, adjusting data cleaning techniques, proper annotation methods, and others.

2.1. Data selection

When designing the *WITS* corpus, the first question was related to what titles should be included. Given the large number of comedy series that have aired over the decades and to ensure the high cultural relevancy of the corpus, I used a variety of criteria to select the shows for *WITS*. First, I created a list of all TV series that were categorized as belonging to the "comedy" genre and that included the keywords "workplace" or "work" on the Internet Movie Database (henceforth IMDb, www.imdb.com).

Since I am primarily concerned with American English, based on the worldwide proliferation of American pop culture in the past decades, I chose only US productions with primarily American writers and casts, excluding any British workplace comedies, such as *The IT Crowd* (2006-2013). Furthermore, in order to take as much of an inclusive approach as possible to contemporary language, I decided to focus on series that have initially aired over the past 25 years, as this timeframe spans from the global rise and new "golden age" of American television in the 2000s (Bednarek, 2018, p. 82) up to the present.

Moreover, there were other discourse-related criteria that impacted my corpus design. I restricted my selection to series that were first broadcasted on non-premium television, omitting the likes of HBO's *Veep* (2012-2019) and *Silicon Valley* (2014-2019). This restriction acknowledges Bednarek's findings that there are usually major differences in language use between network and subscription-based outlets, particularly when it comes to the censorship of taboo words (2018, p. 7). As one focus of my study is work-related discourse in workplace sitcoms, I decided to exclude any comedies that are set in a workplace, but in which "working" does not play an important role, as is the case in *It's Always Sunny in Philadelphia* (2005-ongoing).

Even so, the list of workplace sitcoms included in *WITS* is not exhaustive and may be further expanded in the future.

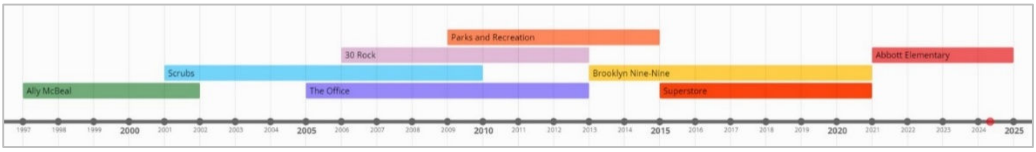


Figure 1. Timeline of WITS series

Taking all aspects into consideration, at the end of the corpus design stage, I had eight valid workplace comedy series that could be included in the *WITS* corpus (see Figure): *Ally McBeal* (1997-2002), *Scrubs* (2001-2010), *The Office* (2005-2013), *30 Rock* (2006-2013), *Parks and Recreation* (2009-2015), *Brooklyn Nine-Nine* (2013-2021), *Superstore* (2015-2021) and *Abbott Elementary* (2021-ongoing). Over the course of their initial broadcasting, six of the series were exclusively broadcast to one channel, with *Scrubs* and *Brooklyn Nine-Nine* as the only two to appear on multiple channels throughout their run. Six of the eight aired on NBC, two on ABC, and two on Fox. Based on Bednarek's classification of "quality television", i.e., shows that have been nominated and/or have won an Emmy or a Golden Globe for "best/outstanding series" (2018, 83), seven of the *WITS* titles can be considered "quality" series. These series also enjoy high audience ratings, with an average score of 7.9 out of 10 per episode (see Figure 2).

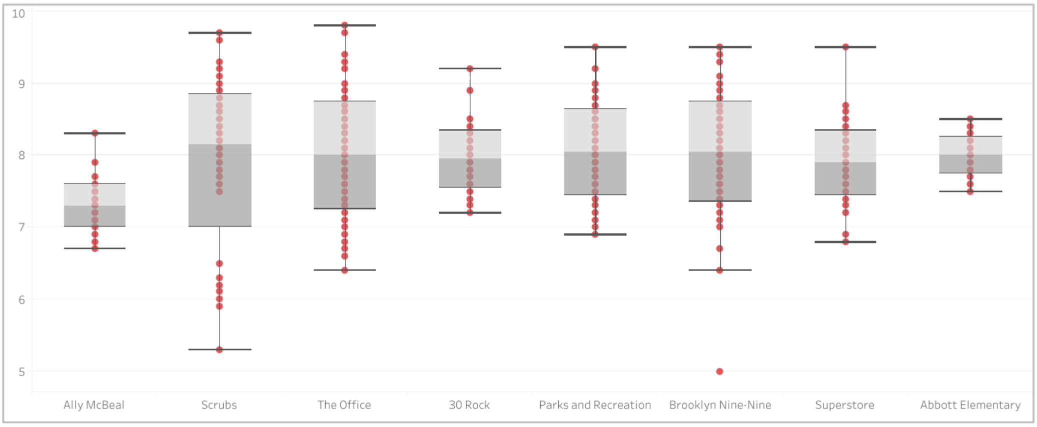
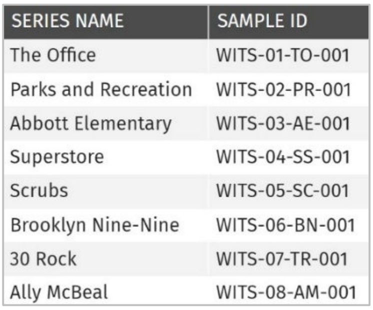


Figure 2. IMDb ratings per TV series

2.2. Included metadata

Before the data collection stage, I extracted metadata for the eight workplace sitcoms in *WITS* from their IMDb pages. Every episode that officially aired on television at least once was included in the metadata table, excluding unaired pilots and web-exclusive specials. Generally, I maintained the official production numbering of the episodes as per IMDb, except for two-part episodes that originally aired on the same date, in order to retain their narrative integrity.

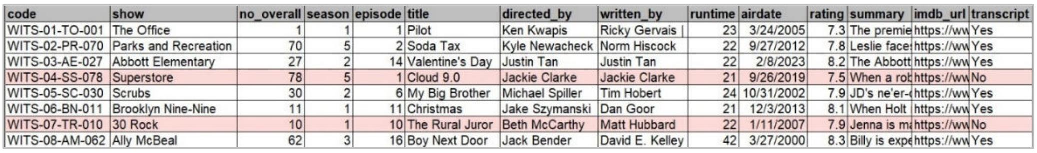
Each episode was attributed a unique identification number, i.e. ID code (see Figure 3). Each code follows the same format: corpus name (*WITS*, short for ***w****orkplace* ***sit****com****s***), number in the corpus (from 01 to 08), subcorpus (two representative letters), and episode number (starting from 001). The corpus number is not chronological, i.e., based on the series initial airdate; rather, it represents the arbitrary order in which I have decided to process the corpus data. The name of the subcorpora were generally chosen to represent two letters from the series' names, ideally the capitalized letters. Attributing IDs to all episodes ensures that the corpus is expandable in the future, as more episodes are released (in the case of *Abbott Elementary*) or more transcriptions are published online.



SERIES NAME	SAMPLE ID
The Office	WITS-01-TO-001
Parks and Recreation	WITS-02-PR-001
Abbott Elementary	WITS-03-AE-001
Superstore	WITS-04-SS-001
Scrubs	WITS-05-SC-001
Brooklyn Nine-Nine	WITS-06-BN-001
30 Rock	WITS-07-TR-001
Ally McBeal	WITS-08-AM-001

Figure 3. Sample ID codes for WITS series

Afterwards, I created a list of all IMDb pages related to the episodes in my corpus. IMDb contains information about the episodes and provides relevant insight into various aspects related to the production and reception of the series. These characteristics were collected by web scraping each respective IMDb link. Web scraping is an automatic extraction method in which website data is accessed, mined, and exported in a spreadsheet format (Barton, 2024). Subsequently, the metadata spreadsheet of the *WITS* corpus includes the above-mentioned details, such as unique ID code, series title, episode number overall (same as episode number used in the ID code), season and episode number in season, episode title, director(s), writer(s), runtime, original airdate, IMDb rating, summary, IMDb URL, as well as a column that automatically displays whether an episode transcript is added to the *WITS* corpus database. As Figure 4 shows, the spreadsheet is also conditionally formatted to display rows that do not have any transcript data attributed. The storing of the metadata in spreadsheet format entails the ability to sort and filter the given information based on any number of criteria, which will prove useful in my data analysis, as well as offer the possible addition of other metadata categories in the future.



code	show	no_overall	season	episode	title	directed_by	written_by	runtime	airdate	rating	summary	imdb_url	transcript
WITS-01-TO-001	The Office	1	1	1	Pilot	Ken Kwapis	Ricky Gervais |	23	3/24/2005	7.3	The premie	https://ww	Yes
WITS-02-PR-070	Parks and Recreation	70	5	2	Soda Tax	Kyle Newacheck	Norm Hiscock	22	9/27/2012	7.8	Leslie face	https://ww	Yes
WITS-03-AE-027	Abbott Elementary	27	2	14	Valentine's Day	Justin Tan	Justin Tan	22	2/8/2023	8.2	The Abbott	https://ww	Yes
WITS-04-SS-078	Superstore	78	5	1	Cloud 9.0	Jackie Clarke	Jackie Clarke	21	9/26/2019	7.5	When a rob	https://ww	No
WITS-05-SC-030	Scrubs	30	2	6	My Big Brother	Michael Spiller	Tim Hobert	24	10/31/2002	7.9	JD's ne'er-c	https://ww	Yes
WITS-06-BN-011	Brooklyn Nine-Nine	11	1	11	Christmas	Jake Szymanski	Dan Goor	21	12/3/2013	8.1	When Holt	https://ww	Yes
WITS-07-TR-010	30 Rock	10	1	10	The Rural Juror	Beth McCarthy	Matt Hubbard	22	1/11/2007	7.9	Jenna is ma	https://ww	No
WITS-08-AM-062	Ally McBeal	62	3	16	Boy Next Door	Jack Bender	David E. Kelley	42	3/27/2000	8.3	Billy is expe	https://ww	Yes

Figure 4. Metadata categories for WITS

2.3. Data collection

The next step in building the *WITS* corpus was gathering dialogue lines from at least ten episodes of each of the eight abovementioned workplace sitcoms. It must be noted there is no agreed-upon "best" source of television dialogue data, since there are multiple aspects to be considered, and researchers may value each one differently. For instance, if the desired result is collecting a *large* dataset of "raw" unannotated lines, then automatically extracting subtitles is the easiest method (Davies, 2021, p. 15). However, as Bednarek (2015, p. 69) and Bywood *et al.* (2013, p. 598) point out, there are several disadvantages to using subtitles: they rarely include the names of the speakers, they may reflect standardized versions of the English language (e.g., grammar, spelling, etc.), and they tend to "correct" characters' oral mistakes. Thus, subtitles may not be entirely accurate to dialogue spoken on-screen.

Another potential source of dialogue data, official scripts, has also been found to be inadequate. Bednarek notes that scripts are oftentimes available solely in PDF format, which means that their conversion is problematic even with the best software available. She also explains that online scripts can be earlier working drafts, thus featuring significant changes from the on-air version of the shows, and that a "definitive" script usually does not exist (2015, p. 70–71). For instance, since some sitcoms are known to allow some actors to improvise certain lines on set (Greene 2020, 169), it is impossible for scripts to be fully representative of on-air dialogue. These observations are in line with the findings of Taylor (2004) and Forchini (2012), which underline the changes that occur between written scripts and the final on-screen products.

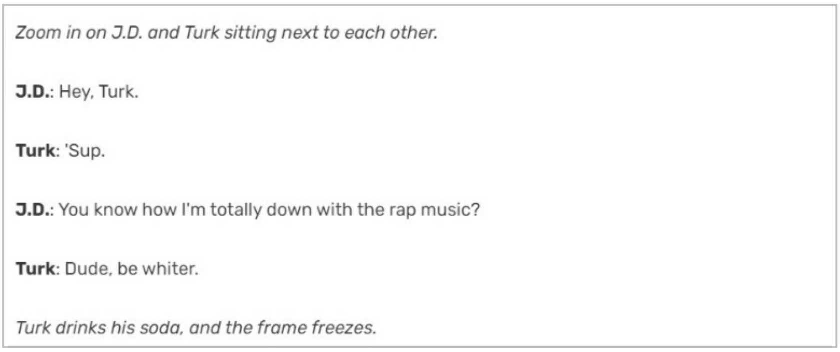
The final option for collecting television dialogue is transcribing the data. Transcripts, as opposed to scripts, are written post-factum, i.e., after the airing of an episode, and consist of the lines of dialogue heard in the episode (Quaglio, 2009, p. 30). Transcripts can be created by researchers themselves, but this endeavor is unpractical, time-consuming, and costly for small research teams (Bednarek, 2015, p. 72).

Instead, a potential alternative is offered by fan transcripts, i.e., transcriptions done by fan clubs and made available online. Quaglio found that fan transcripts are more reliable than official scripts or subtitles, claiming that they "were not only fairly accurate, but also extremely detailed, including several features that scripts are not likely to present: hesitators …, pauses …, repeats …, contractions …, and even descriptions of the scenes and actors' performances" (2009, p. 30).

Taking all concerns into consideration, the *WITS* corpus was built on fan transcripts. The main sources for fan transcripts were *Fandom* (www.fandom.com), the largest wiki for the entertainment sector, and *ForeverDreaming* (www.transcripts.foreverdreaming.com), a forum featuring fan transcripts for a large number of telecinematic products. These two were used alternatively for seven of the series. Another source, that was solely used for *Ally McBeal* transcripts, was a Japanese website (www.upp.so-net.ne.jp) from the early 2000s that offered English and Japanese transcripts of the show (dubbed *Ally, my love*) for English-learning purposes. This website is no longer active, but it has been archived on the *Wayback Machine* (web.archive.org), making it still accessible today.

2.4. Data processing and storage

Even though the eight selected shows are fully transcribed on the websites mentioned above, not all transcriptions follow the same format. A small number of episodes are extremely detailed, with narrative sequences describing character actions, as in Figure 5. The majority have only allocated speaker names to each dialogue line, occasionally describing character actions between square brackets (see Figure 6). Many others, however, did not have any additional information other than the spoken dialogue lines (see Figure 7). Because of the lack of standardization in the transcripts and in order to be able to add a large number of episodes to the *WITS* corpus, I decided to process my data using only speaker names and their respective spoken lines, as in Figure 6 below.



Zoom in on J.D. and Turk sitting next to each other.

J.D.: Hey, Turk.

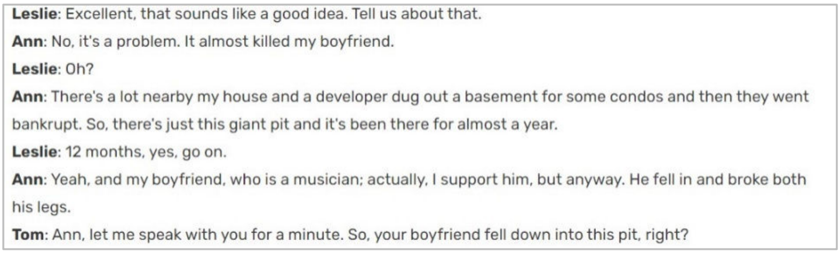
Turk: 'Sup.

J.D.: You know how I'm totally down with the rap music?

Turk: Dude, be whiter.

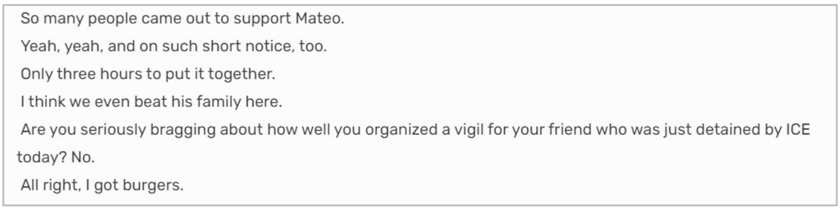
Turk drinks his soda, and the frame freezes.

Figure 5. Example of transcript with speaker names and narrative details



Leslie: Excellent, that sounds like a good idea. Tell us about that.
Ann: No, it's a problem. It almost killed my boyfriend.
Leslie: Oh?
Ann: There's a lot nearby my house and a developer dug out a basement for some condos and then they went bankrupt. So, there's just this giant pit and it's been there for almost a year.
Leslie: 12 months, yes, go on.
Ann: Yeah, and my boyfriend, who is a musician; actually, I support him, but anyway. He fell in and broke both his legs.
Tom: Ann, let me speak with you for a minute. So, your boyfriend fell down into this pit, right?

Figure 6. Example of transcript with speaker names



So many people came out to support Mateo.
Yeah, yeah, and on such short notice, too.
Only three hours to put it together.
I think we even beat his family here.
Are you seriously bragging about how well you organized a vigil for your friend who was just detained by ICE today? No.
All right, I got burgers.

Figure 7. Example of raw transcript

Figure 8 charts the number of transcribed episodes that have been processed in the *WITS* corpus as of late 2024. Overall, 628 out of 1048 total episodes (roughly 60%) in the eight series were processed so far. *The Office* and *Parks and Recreation* are the only two series fully transcribed in *WITS*. *Scrubs*, *Superstore,* and *Abbott Elementary* have more than 68% of their total episodes transcribed. *Ally McBeal* sits at 45%, and *Brooklyn Nine-Nine* at 17%. *30 Rock* is the series with the lowest percentage of episodes transcribed (7%), as it is the only TV series that did not have any transcripts with allocated speakers whatsoever. Thus, in order to process it, 10 episodes were randomly selected and manually assigned speaker names through parallel viewing. Although this was a time-consuming effort, I aim to add at least ten more episodes of *30 Rock* to the *WITS* corpus in order to reach at least 15% of the total episodes transcribed for each series.

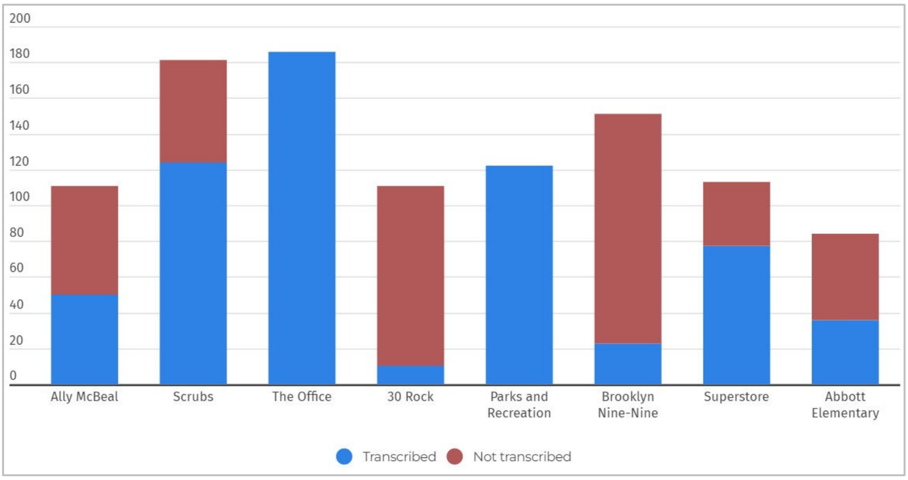


Figure 8. Number of episodes in WITS

During the processing stage, the *WITS* data was stored and managed exclusively in spreadsheets. At first, every episode was allotted its own sheet in which the transcriptions were manually copied from the source websites, each row of a sheet representing one line of spoken dialogue. Using the data separator tools, I divided the names of the speakers and their lines into two different columns based on the use of colons. While this was largely a success, some lines had to be manually divided due to the erroneous use of semi-colons as a separator. Via a series of filters and formulas, I then cleaned and organized the data following a set of principles, which are described below.

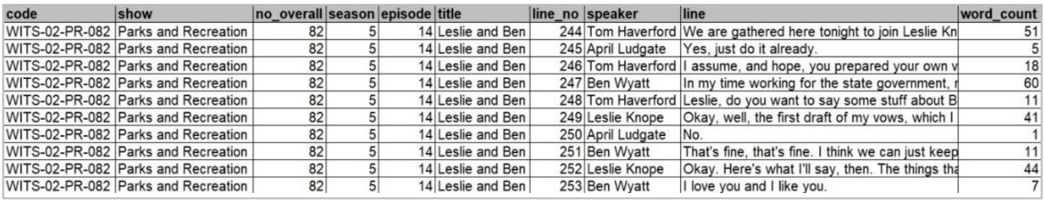
First, the "speaker" column was normalized. Multiple names referring to the same character were standardized to make filtering easier. As a rule, the speakers are all formatted as first name (or nickname) and last name, with the exception of unnamed extras or characters with no full name given. All honorifics, titles, and positions (e.g., "Dr.", "Mr.", "Captain") were removed. For example, in the *Brooklyn Nine-Nine* subcorpus, the speaker names "Holt", "Ray Holt", "Raymond Holt", "Captain Holt" appear in the transcripts, but in *WITS*, they were all changed to "Raymond Holt".

Then, all non-dialogue parts from "speaker" and "line" columns were removed. Details related to character actions, narrative sequences, and locations (see Figure 5) were removed because they were enclosed in brackets (round, square, and curly) or written in all caps. Afterward, data was also manually checked to ensure no accidental omissions of non-dialogue information.

Any missing data was manually added. Transcribers used three question marks ("???") or an equal sign ("=") to indicate places in which their transcription is incomplete. They also color-coded instances where they were unsure of their transcriptions in red or yellow. This generally occurred when characters used rare or specialized vocabulary, or when there were overlaps in the conversations. In both such instances, I viewed the corresponding episodes and manually filled in the missing information. Lines spoken in foreign languages (e.g., Spanish, Tagalog) were only partially transcribed, based on consulting multiple sources.

Finally, all bleeped instances were coded. As they were initially broadcast in the US, the series follow strict regulations when it comes to the use of swear words, with the most profane of them being bleeped. Transcribers marked bleeps using two or more asterisks (e.g., "Holy ****" or "Holy s**t"). In the *WITS* corpus, all bleeped words are marked using the code [BLEEP].

Lastly, the data was combined in a single spreadsheet. Basic metadata categories (episode code, series name, season and number, title) were added in order to allow easy filtering. Each line of dialogue was attributed a line number to ensure that the chronological order of each episode is still intact, even when the database is filtered. Finally, using a pre-existing formula, the word count of each line appears automatically in the final column of the dataset. Figure displays an example of processed dialogue lines from *Parks and Recreation*.



code	show	no_overall	season	episode	title	line_no	speaker	line	word_count
WITS-02-PR-082	Parks and Recreation	82	5	14	Leslie and Ben	244	Tom Haverford	We are gathered here tonight to join Leslie Kn	51
WITS-02-PR-082	Parks and Recreation	82	5	14	Leslie and Ben	245	April Ludgate	Yes, just do it already.	5
WITS-02-PR-082	Parks and Recreation	82	5	14	Leslie and Ben	246	Tom Haverford	I assume, and hope, you prepared your own v	18
WITS-02-PR-082	Parks and Recreation	82	5	14	Leslie and Ben	247	Ben Wyatt	In my time working for the state government, r	60
WITS-02-PR-082	Parks and Recreation	82	5	14	Leslie and Ben	248	Tom Haverford	Leslie, do you want to say some stuff about B	11
WITS-02-PR-082	Parks and Recreation	82	5	14	Leslie and Ben	249	Leslie Knope	Okay, well, the first draft of my vows, which I	41
WITS-02-PR-082	Parks and Recreation	82	5	14	Leslie and Ben	250	April Ludgate	No.	1
WITS-02-PR-082	Parks and Recreation	82	5	14	Leslie and Ben	251	Ben Wyatt	That's fine, that's fine. I think we can just keep	11
WITS-02-PR-082	Parks and Recreation	82	5	14	Leslie and Ben	252	Leslie Knope	Okay. Here's what I'll say, then. The things tha	44
WITS-02-PR-082	Parks and Recreation	82	5	14	Leslie and Ben	253	Ben Wyatt	I love you and I like you.	7

Figure 9. Example from WITS dataset

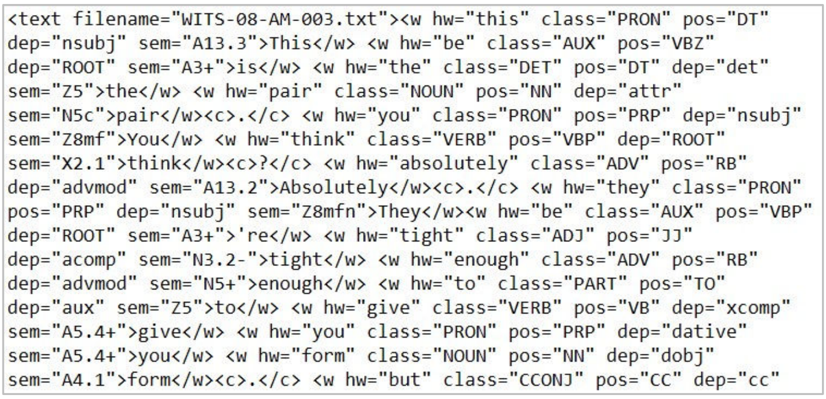
Since spreadsheets cannot generally be used as data sources for corpus tools, the *WITS* corpus has also been exported into UTF-8 encoded **.txt* files. This was done in the episodic format, meaning that each **.txt* file is named after one episode code and contains all dialogue lines that are attributed to that specific episode, thus allowing a chronological examination of linguistic data, with the possibility of looking into each series individually, similar to Davies' approach to virtual corpora in *The TV corpus* (2021, p. 16–21)An additional benefit of storing the data as such is the ease with which various subcorpora can be created. For instance, files can be exported to contain all dialogue lines attributed to one "speaker" to delve into corpus-driven examinations of character identity (cf. Bednarek, 2023).

2.5. Data accuracy and linguistic annotation

To test and ensure the accuracy of the fan transcripts found on these websites, I selected five random episodes of each series, totaling 40 episodes, and compared them manually with their respective episodes. In each case, the transcripts were found to be satisfactory, accurately reflecting dialogue lines, including the aspects that Quaglio mentioned related to hesitation, repetitions of words or sounds, and contracted verb forms (2009, p. 30). However, they were not entirely accurate, which prompted me to create two separate categories of the *WITS* corpus: the large *WITS* corpus, containing all available transcripts, which will be used to compare the language of workplace sitcoms to a larger reference corpus of natural language, and a smaller *WITS* corpus, more balanced in terms of episodes per series (10-20 for each TV show) and manually checked for complete accuracy, which will be used for pedagogical activities.

On a linguistic level, annotation refers to the act of applying supplementary information to corpus data related to its analysis of various levels: grammatical, semantical, phonetic, etc. (Baker *et al.*, 2006, p. 66–67). Some types of annotations, such as part-of-speech (POS) tagging or semantic tagging, can be done automatically through the use of taggers. For POS tagging, the most well-known example would be CLAWS ("Constituent-Likelihood Automatic Word-Tagging System"), which assigns an indicator of grammatical class to each word based on its context. The CLAWS system has been perfected over time, achieving around 96-7% accuracy (Garside, 1990, p. 31). Another example of a frequently used automatic tagger is USAS ("UCREL Semantic Analysis System"), which groups words based on their meanings being more or less related to the same conceptual framework with a 96-8% accuracy (Rayson *et al.*, 2004, p. 7–8).

In its main format, the *WITS* corpus data is not annotated for its linguistic features, i.e., it does not feature any linguistic analysis encoded in the data itself. However, by using LancsBox X (Brezina & Platt, 2024), the *WITS* corpus has been automatically annotated through POS and semantic tagging. LancsBox X allows files to be exported after being processed through the application, meaning that the annotated version of the *WITS* data is also readily available under the format shown in Figure 10. Other linguistic annotations may be possible in the future.



```
<text filename="WITS-08-AM-003.txt"><w hw="this" class="PRON" pos="DT"
dep="nsubj" sem="A13.3">This</w> <w hw="be" class="AUX" pos="VBZ"
dep="ROOT" sem="A3+">is</w> <w hw="the" class="DET" pos="DT" dep="det"
sem="Z5">the</w> <w hw="pair" class="NOUN" pos="NN" dep="attr"
sem="N5c">pair</w><c>.</c> <w hw="you" class="PRON" pos="PRP" dep="nsubj"
sem="Z8mf">You</w> <w hw="think" class="VERB" pos="VBP" dep="ROOT"
sem="X2.1">think</w><c>?</c> <w hw="absolutely" class="ADV" pos="RB"
dep="advmod" sem="A13.2">Absolutely</w><c>.</c> <w hw="they" class="PRON"
pos="PRP" dep="nsubj" sem="Z8mfn">They</w><w hw="be" class="AUX" pos="VBP"
dep="ROOT" sem="A3+">'re</w> <w hw="tight" class="ADJ" pos="JJ"
dep="acomp" sem="N3.2-">tight</w> <w hw="enough" class="ADV" pos="RB"
dep="advmod" sem="N5+">enough</w> <w hw="to" class="PART" pos="TO"
dep="aux" sem="Z5">to</w> <w hw="give" class="VERB" pos="VB" dep="xcomp"
sem="A5.4+">give</w> <w hw="you" class="PRON" pos="PRP" dep="dative"
sem="A5.4+">you</w> <w hw="form" class="NOUN" pos="NN" dep="dobj"
sem="A4.1">form</w><c>.</c> <w hw="but" class="CCONJ" pos="CC" dep="cc"
```

Figure 10. POS and semantic annotation of WITS file

3. Meeting the *WITS* corpus

As of late 2024, the large *WITS* corpus contains over 2 million words of dialogue lines from eight contemporary American workplace sitcom series that are considered to be representative of the genre. The data was extracted from 628 episodes across the eight series. The total runtime of the episodes included in *WITS* is 262 hours and 37 minutes, with an average runtime per episode of 25 minutes. When it comes to its sample size, if counting a word as being separated by two white spaces, *WITS* measures 2.086 million words, averaging 3,322 words per episode. Figure 11 illustrates the overall word count distribution of the series in *WITS*.

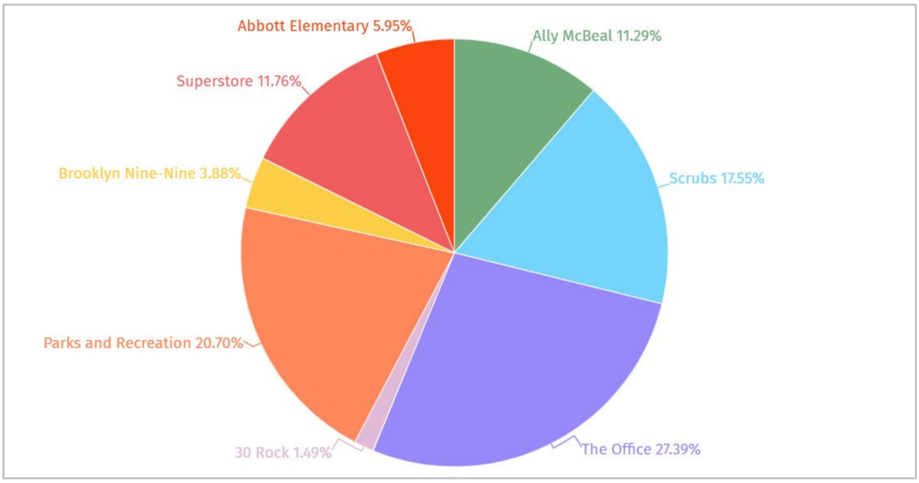


Figure 11. Word count distribution of WITS

3.1. Limitations of the WITS corpus

The size and scope of the *WITS* corpus were both impacted by time and resource constraints. While *WITS* aims to be representative of the TV subgenre of contemporary American workplace sitcoms, it is not comprehensive in the list of series included. *WITS* is also not yet entirely balanced when it comes to the length distribution of the data (McEnery & Hardie, 2012, p. 9). The amount of data collected so far is different for each series due to the low number of existing fan transcripts with allocated speaker names for some shows, such as *Brooklyn Nine-Nine* and *30 Rock*. Additionally, since the series themselves vary in number of seasons, episodes per season, and runtime length, some series may be overrepresented in the raw frequency lists of the *WITS* corpus. That is why it is important to consider comparing the subcorpora among themselves, as well as look at the relative frequency of certain linguistic items across the entire corpus (McEnery & Hardie, 2012, p. 247).

However, a benefit of the current design of *WITS* is that, by watching the shows and designing the corpus to include extra information such as episode summaries, my level of familiarity with all the series can be considered high. Thus, I ensure that this study is not merely focused on counting "decontextualized data" and that I do not start "from the position of tabula rasa" (Baker, 2006, p. 25). Instead, I will be able to provide meaningful insight into the relationships between linguistic data, character identities, and storyline developments. Furthermore, because of the current size and scope of the corpus, I am able to easily access the source videos of the series, be it on subscription-based platforms (e.g., *Netflix*, *Max*) or DVDs. McEnery & Hardie note that the level of ease of access to corpus data is an important aspect that corpus builders need to take into consideration (2012, p. 66). Moreover, the availability of the source videos means that I can undertake multimodal analyses of key scenes based on the linguistic data in the *WITS* corpus.

Another impact that time constraints had on the *WITS* corpus relates to the level of detail and type of linguistic information included. In spite of it containing spoken data, the corpus does not contain typical mark-ups of spoken corpora, such as details about extra-linguistic information (gestures, actions), prosodic elements (rhythm, stress, intonation), or phonetic transcriptions (Knight & Adolphs, 2022, p. 28–29). Therefore, on its own, *WITS* can only be used to identify lexical and grammatical aspects of characters' speech. It is also notable that while *WITS* primarily features American English, there are certain characters who also speak in other languages, such as Spanish (e.g., Carla Espinosa in *Scrubs*), Tagalog (e.g., Mateo Liwanag in *Superstore*), or German (e.g., Dwight Schrute in *The Office*). As explained in Section 2.4, not all such dialogue was transcribed, which means that *WITS* is not a suitable source to analyze multilingualism in TV series.

The corpus also does not contain other extra-linguistic information other than speaker names and corresponding lines. In the future, I plan to add more detailed annotation when it comes to the locations of the scenes, in order to be able to filter between conversations that take place in the workplace and outside. In addition, I plan to add certain annotations for speakers, i.e., extra information about the socio-demographic profile of some of the key characters in the series, including their gender, sexuality, ethnicity, nationality, profession and employment status. This will allow me to link the results of my quantitative language analysis to "individual and social character traits as well as character relationships and social stereotypes and norms" (Bednarek 2023, 29).

4. Conclusion. Next steps

Overall, this paper has introduced the first version of the *WITS* corpus, detailing the design principles, selection criteria, and main steps in the process of building a corpus of dialogue lines extracted from fan transcriptions of eight contemporary American workplace sitcoms. Moving forward, I plan to expand the corpus to its full potential, e.g., include enough data from all sources and add more relevant linguistic and extralinguistic annotation. Thus, I will be able to perform high-quality corpus-based analyses (e.g., frequency-based, keyness analyses of word and n-gram units) and develop pedagogical materials for learners of English as a foreign language.

References:

Al-Surmi, M. (2012). Authenticity and TV Shows: A Multidimensional Analysis Perspective. *TESOL Quarterly*, *46*(2), 671–694. https://doi.org/10.1002/tesq.33.

Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh University Press.

Barton, D. (2024, April 19). What is web scraping? *Apify Blog*. https://blog.apify.com/what-is-web-scraping/

Bednarek, M. (2010). *The language of fictional television: Drama and identity*. Continuum International Pub. Group.

Bednarek, M. (2011). The language of fictional television: A case study of the 'dramedy' *Gilmore Girls*. *English Text Construction*, *4*(1), 54–84. https://doi.org/10.1075/etc.4.1.04bed.

Bednarek, M. (2012a). Constructing 'nerdiness': Characterisation in The Big Bang Theory. *Multilingua*, *31*(2–3), 199–229. https://doi.org/10.1515/multi-2012-0010.

Bednarek, M. (2012b). “Get us the hell out of here”: Key words and trigrams in fictional television series. *International Journal of Corpus Linguistics*, *17*(1), 35–63. https://doi.org/10.1075/ijcl.17.1.02bed.

Bednarek, M. (2015). Corpus-Assisted Multimodal Discourse Analysis of Television and Film Narratives. In P. Baker & T. McEnery (Eds.), *Corpora and discourse studies: Integrating discourse and corpora* (pp. 63–87). Palgrave Macmillan.

Bednarek, M. (2018). *Language and television series: A linguistic approach to TV dialogue*. Cambridge University Press.

Bednarek, M. (2023). *Language and characterisation in television series: A corpus-informed approach to the construction of social identity in the media*. John Benjamins Publishing Company.

Beier, S., Berlow, S., Boucaud, E., Bylinskii, Z., Cai, T., Cohn, J., Crowley, K., Day, S. L., Dingler, T., Dobres, J., Healey, J., Jain, R., Jordan, M., Kerr, B., Li, Q., Miller, D. B., Nobles, S., Papoutsaki, A., Qian, J., … Wolfe, B. (2022). Readability Research: An Interdisciplinary Approach. *Foundations and Trends® in Human–Computer Interaction*, *16*(4), 214–324. https://doi.org/10.1561/1100000089.

Brezina, V., & Platt, W. (2024). *#LancsBox X* (4.0.0) [Computer software]. lancsbox.lancaster.ac.uk

Bywood, L., Volk, M., Fishel, M., & Georgakopoulou, P. (2013). Parallel subtitle corpora and their applications in machine translation and translatology. *Perspectives*, *21*(4), 595–610. https://doi.org/10.1080/0907676X.2013.831920.

Charney, L. (2005). Television Sitcoms. In M. Charney (Ed.), *Comedy: A geographic and historical guide* (Vol. 2, pp. 586–600). Praeger.

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)* [Computer software]. https://www.english-corpora.org/coca/.

Davies, M. (2011). *Corpus of American Soap Operas* [dataset]. https://www.english-corpora.org/soap/

Davies, M. (2021). The TV and Movies corpora: Design, construction, and use. *International Journal of Corpus Linguistics*, *26*(1), 10–37. https://doi.org/10.1075/ijcl.00035.dav.

Dose, S. (2012). Scripted speech in the EFL classroom: The Corpus of American Television Series for teaching spoken English. In J. Thomas & A. Boulton (Eds.), *Input, Process and Product: Developments in Teaching and Language Corpora* (pp. 103–121). Masarykova univerzita; CEEOL. https://www.ceeol.com/search/chapter-detail?id=837377.

Dose, S. (2013). Flipping the script: A Corpus of American Television Series (CATS) for corpus-based language learning and teaching. *Varieng: Studies in Variation, Contacts and Change in English*. https://varieng.helsinki.fi/series/volumes/13/dose/.

Forchini, P. (2012). *Movie language revisited: Evidence from multi-dimensional analysis and corpora.* Peter Lang.

Garside, R. (1990). The CLAWS Word-tagging System. In R. Garside & G. Sampson (Eds.), *The computational analysis of English: A corpus-based approach* (1 publ., 2. impr, pp. 30–41). Longman.

Greene, A. (2020). *The office: The untold story of the greatest sitcom of the 2000s*. Dutton.

Gregori-Signes, C. (2017). “Apparently, women don’t know how to operate doors": A corpus-based analysis of women stereotypes in the TV series *3rd Rock from the Sun*. *International Journal of English Studies*, *17*(2), 21. https://doi.org/10.6018/ijes/2017/2/257311.

Heyd, T. (2010). HOW YOU GUYS Doin’? STAGED ORALITY AND EMERGING PLURAL ADDRESS IN THE TELEVISION SERIES *FRIENDS*. *American Speech*, *85*(1), 33–66. https://doi.org/10.1215/00031283-2010-002.

Knight, D., & Adolphs, S. (2022). Building a spoken corpus: What are the basics? In A. O’Keeffe & M. J. McCarthy, *The Routledge Handbook of Corpus Linguistics* (2nd ed., pp. 21–34). Routledge. https://doi.org/10.4324/9780367076399-3.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2022). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 319–344. https://doi.org/10.1075/ijcl.22.3.02lov.

McEnery, T., Baker, P., & Hardie, A. (2000). Ssessing claims about language use with corpus data – swearing and abuse. In J. M. Kirk (Ed.), *Corpora Galore: [Analyses and techniques in describing English]; papers from the nineteenth International Conference on English Language Research on Computerised Corpora (ICAME 1998)* (pp. 45–55). Rodopi.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Nørgaard, N., Montoro, Rocío., & Busse, B. (2010). *Key terms in stylistics*. Continuum International Pub. Group.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Quaglio, P. (2008). Television dialogue and natural conversation: Linguistic similarities and functional differences. In A. Ädel & R. Reppen (Eds.), *Corpora and discourse: The challenges of different settings* (pp. 189–210). John Benjamins Pub. Co.

Quaglio, P. (2009). *Television dialogue: The sitcom Friends vs. natural conversation*. John Benjamins Pub. Co.

Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). The UCREL semantic analysis system. *Proceedings of the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP Tasks in Association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 7–12.

Richardson, K. (2010). *Television Dramatic Dialogue*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195374056.001.0001.

Sardinha, T. B., & Pinto, M. V. (2017). American television and off-screen registers: A corpus-based comparison. *Corpora*, *12*(1), 85–114. https://doi.org/10.3366/cor.2017.0110.

Schneider, J. (2023, August 28). How the workplace became the star of TV. *BBC*. https://bbc.com/worklife/article/20230824-how-the-workplace-became-the-star-of-tv.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. John Benjamins Pub.

Taylor, C. (2004). The Language of Film: Corpora and Statistics in the Search for Authenticity. Notting Hill (1998) – A Case Study. *Miscelánea*, *30*, 71–86. Dialnet.

Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora*, *8*(1), 81–113. https://doi.org/10.3366/cor.2013.0035.

Toolan, M. (2011). Chapter 9. "I don't know what they're saying half the time, but I'm hooked on the series": Incomprehensible dialogue and integrated multimodal characterisation in The Wire. In R. Piazza, M. Bednarek, & F. Rossi (Eds.), *Pragmatics & Beyond New Series* (Vol. 211, pp. 161–183). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.211.12too.