Guide to the Sydney Corpus of Television Dialogue (SydTV)

Please cite as:

Bednarek, M. 2018 [last updated February 2021]. *Guide to the Sydney Corpus of Television Dialogue* (*SydTV*). Available at www.syd-tv.com, ACCESS DATE.

1. Preface

The *Sydney Corpus of Television Dialogue* (SydTV) is a new, carefully designed dataset of TV dialogue. SydTV is a small, specialised corpus (~275,000 words), representative of the language variety of contemporary US TV dialogue. TV dialogue is here defined as the dialogue uttered by actors on screen as they are performing characters in fictional TV series, and does not include screen directions, etc. This manual reports on the construction of the corpus.

SydTV could not have been built without School and Faculty funding provided by the University of Sydney. I want to thank the research assistants who helped with building the corpus over several years: Cassandra Liardét (née Fawcett), David Lesslie, Samuel Luke, Ganna Veselovska, and Charlie Revett. For help with making the corpus available to others through the CQPweb user interface, I am very grateful to Chao Sun, Andrew Hardie, and Andressa Rodrigues Gomide. For information on the CQPweb version, see section 5 below.

Information on access to the corpus is provided at the companion website www.syd-tv.com, which also lists the publications that have drawn on the corpus. Frequency and keyness lists are also available for download on this website. Use of SydTV by other researchers is subject to the following conditions:

1. Access is only granted for the purposes of research or scholarship;
2. The corpus cannot be distributed to others;
3. No data from the corpus are permitted to be copied, duplicated or distributed, with the exception of ‘fair use’ in scholarly or educational texts or presentations;
4. Copyright for all material in SydTV remains with the original creators and the material can only be used for criticism, education, and scholarship;
5. The *Sydney Corpus of Television Dialogue* (SydTV) must be acknowledged in any publication(s) and presentation(s) resulting from research on the corpus.
6. Any errors in the corpus should be reported to Monika.BednarekATsydney.edu.au

2. Corpus design

Sinclair recommends that the design of one’s corpus ‘should be documented fully with information about the contents and arguments in justification of the decisions taken’ (Sinclair 2005: 8). In this manual I therefore describe the design of the *Sydney Corpus of Television Dialogue* in some detail.

SydTV contains dialogue from one first-season-episode of 66 different US TV series. The corpus exists in two different versions: the original version (as transcribed) and a partially standardised version, as described in section 4 below. The standardised version (SydTV-Std) is useful for comparing word forms and n-grams across corpora. For example, standardising all instances of *fuckin’* to *fucking* allows the software to treat these as instances of the same word form. However, the original version is useful for analysis of nonstandard language use. In total, the corpus contains about 275,000 words, although its size varies slightly depending on the token definition used (Table 1). A list of all episodes included in SydTV is provided in the appendix. Descriptions of each series can easily be found on the Internet, and plot summaries of most episodes are available in relevant online episode guides.

Table 1 Corpus size of SydTV and SydTV-Std depending on token definitions

Corpus size in number of words		
Token definitions (*Wordsmith 'tokens in text'*)	*SydTV*	*SydTV-Std*
hyphens do not separate words; ' not allowed within word	275,074	276,899
hyphens separate words; ' not allowed within word	276,287	278,112
hyphens do not separate words; ' allowed within word	258,944	260,824
hyphens separate words; ' allowed within word	260,157	262,037

All SydTV files are plain text files (.txt). In relation to tagging, the corpus is predominantly 'raw' text, although speakers were identified as such, either by using angle brackets (i.e. <JACKIE:>) or, in the version made available to other researchers through CQPweb, by using XML-compatible tags (i.e. <u who="JACKIE"> Hey. </u>). The CQPweb version has also been lemmatised and tagged. [**2021 update**: **Note that the token definition and hence resulting number of total words/corpus size differ significantly between CQPweb and Table 1. Raw frequencies should be normalised as per Table 1, unless SydTV is compared with other CQPweb corpora that use the same token definition.]**

The corpus building and hence representativeness of SydTV was constrained by a number of factors: a) legal access to US TV series from the research location (Australia) at the time of corpus construction; b) funding and time needed for transcription and checking of TV dialogue; and c) the time when the corpus was initially designed (e.g. before the 2014 Emmy winners were announced). A full assessment of the corpus is provided in Bednarek (2018).

The small size of SydTV means that the corpus design is crucial. My primary aim was to design a specialised corpus that is maximally representative of the language variety of US TV dialogue. Representativeness refers to the 'extent to which a sample [the corpus] includes the full range of variability in a population' (Biber 1993: 243). The relevant population, or language variety, is contemporary US TV dialogue, that is dialogue from TV series whose country of origin is the United States, and which were first broadcast between 2000 and 2012. This specific time frame was adopted because the first decade of the 21st century was characterised by the global rise of American TV series (*Scripted Series Report* 2010/2011), and has been labelled the new 'golden age of television'.[1] New seasons of some of the TV series included in SydTV are still being produced at the time of writing (e.g. *Veep*, *The Big Bang Theory*, *NCIS*) while most others are being shown as repeats or are available via services such as Netflix, Amazon, or Itunes. All TV series included in the corpus are clearly identified as 'American' and/or have United States as country of origin in their Wikipedia entry. Animated series, soap operas and series targeted exclusively at children or teenagers are not included in SydTV.

To design a representative corpus, I used a mix of production and reception criteria: As far as **production** is concerned, the corpus needs to mimic the variability of contemporary US TV series. That is, the aim was to include as many different TV series as possible and to include a balance of comedy and drama genres, because this is one of the major distinctions made in the TV industry (as evident from Emmy and Golden Globe award categories). The corpus thus contains dialogue from 66 TV series, with about half classified as comedy genres and the other half as drama genres (using the genre labels provided by the Internet Movie Database/IMDb). It must be noted that this two-fold distinction simplifies matters somewhat, since many contemporary TV series are a mix of comedy and drama,[2] or otherwise 'hybrids' (Dunn 2005: 138). The drama category therefore includes genre combinations or hybrid genres such as crime/drama, drama/fantasy or action/drama. Similarly, the comedy category includes TV series that are only classified as 'comedy' by the IMDb (often sitcoms) and 'comedy hybrids' that are labeled genre-wise *first* as comedy by IMDb, with other genre labels also present (e.g. comedy/drama, comedy/crime or comedy/romance).

Another important production variable takes into account the serial nature of TV narratives. TV series are conceptualised and produced as seasons with a particular number of episodes. Even series that typically resolve storylines within an episode often have ongoing stories across episodes. It was hence considered important to include pilot episodes, final episodes and other episodes occurring towards the beginning, middle, and end of the respective season – that is, representing different moments of textual time within the season. I used percentages as a rough-and-ready shortcut for calculating textual time given that the number of episodes per season varies (traditional network TV series have around 22-26 episodes per season, but cable series may have small seasons with only 7-13 episodes). For example, the third episode of a total of 24 episodes represents 12.5% of the season, while the fourth episode of a total of 13 episodes represents 31% of the season; both would be considered as episodes occurring toward the beginning of the season.[3] This consideration of textual time aims to ensure representation across the season to avoid a potential influence of particular kinds of episodes, especially pilot and final episodes which are atypical and have very specific functions (Thompson 2003: 62, Douglas 2011: 53, Mittell 2015: 55-85). Pilots are also written before the writing team is employed. While the corpus thus contains a mix of episodes from different points in the season, only dialogue from the *first* season of a TV series is included, to avoid introducing another variable that might impact on findings.

Reception-based criteria were also taken into account in designing SydTV, namely critical acclaim and popularity. The notion of critical acclaim links to that of quality television. While many acknowledge a rise in quality television programs and would agree on examples such as *Breaking Bad*, *The Wire* or *Mad Men*, quality television can in fact be defined in many different ways (see Thompson 2003, McCabe & Akass 2007, Richardson 2010, Hassler-Forest 2014). Mittell (2015: 211) suggests that 'there is rarely any analytic clarity as to what precisely counts as quality television.' In designing SydTV, quality was therefore solely defined on the basis of Emmy or Golden Globe award nominations or wins for 'best/outstanding' TV series or 'outstanding writing.' The only criterion on the basis of which an episode is labeled as 'quality' is whether the TV series has been nominated or won one or more of the awards listed below:

- Golden Globe nominees or winners (2000-2014) in the categories: *Best TV Series, Drama* or *Best TV Series, Musical or Comedy*; available at www.goldenglobes.com
- Emmy nominees or winners (2000-2013) in the categories: *Outstanding Writing for a Drama Series*; *Outstanding Writing for a Comedy Series*; *Outstanding Comedy Series* or *Outstanding Drama series*; available at www.emmys.com

These awards recognise either writing specifically or the TV series overall, rather than non-dialogue related aspects such as performance. It is entirely possible, for instance, for a TV series to feature superb performances or costume design but mediocre writing. A TV series that was nominated for or won an award for performance, make-up, directing, etc, but not for writing or overall series is thus not labelled 'quality'. The aim was for half the corpus to include dialogue from 'quality' series, with the other half coming from other series (called 'mainstream' from now on, for lack of a better label).

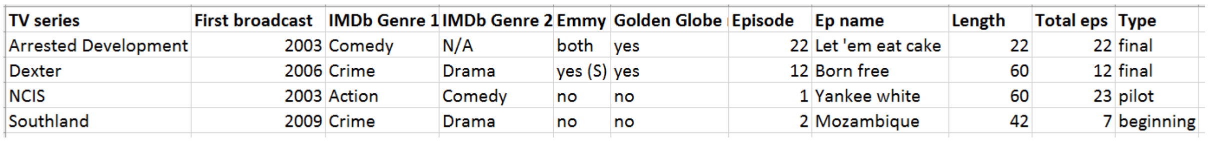
Reception was further taken into account by selecting TV series from lists of bestselling or otherwise popular programmes, in particular:

1. Amazon's 'Bestsellers in TV shows' (www.amazon.com, updated hourly, accessed 4:34 pm Australian Eastern time, 8 November 2010), and Amazon's 'Bestsellers in Movies & TV' (includes DVD, Blu-Ray and Amazon Instant Video, www.amazon.com, updated hourly, accessed 4.52 pm Australian Eastern time, 20 June 2014);
2. Popular series in the Internet Movie Database (IMDb; http://www.imdb.com/), for example, 'Best Action TV Series With At Least 1,000 Votes' (accessed 8 Nov 2010); top 100 'Most popular by genre: Television and Mini-Series' (genre chosen = comedy, accessed 20 June 2014)

The 2010 access date for two of these lists derives from the fact that they were used in the creation of the television corpus (Bednarek 2014), on which SydTV builds. To find additional examples of comedy series (in order to achieve balance) I also consulted a list of US comedies provided by Wikipedia (http://en.wikipedia.org/wiki/List_of_comedies#United_States, accessed 23 June 2014). The lists used as sampling frames (i.e. lists of *potential* series to select for the corpus) can include children's series, reality TV series, comedy sketch-shows, animated series, mini-series, or soap operas, but such entries were disregarded.

The different lists (award nominations/wins; bestsellers; popular series; Wikipedia list of US comedies) were taken as starting point for selecting texts for inclusion in SydTV. The selection was in turn determined by the general goal to include a large number of different TV series and to achieve balance in terms of comedy vs. drama, textual time, and ‘quality’ vs. ‘mainstream’, as outlined above. [4] This includes intersections of variables – for example, care was taken to ensure that not all pilot episodes derive from ‘quality’ series or all comedy series are ‘mainstream’. While the ultimate aim was to achieve a rough balance in the number of words, length in minutes was used during the design stage in the attempt to achieve this aim, since the number of words was then unknown (as it ultimately depends on episode length, amount of dialogue, and speed of delivery).

The IMDb was used systematically to ascertain the two main genre labels for each TV series (as explained above), the number of episodes in the season, the typical length of episodes, and the year of first broadcast. The IMDb is not free from problems, for example with respect to its genre labels (inaccuracies and changes over time are possible), but is used here following Richardson (2010), who also draws on this resource for background information about TV series.Wikipedia was consulted for further contextualisation. An excel file was created to document all relevant information for each episode/TV series that could *potentially* be included in the corpus, information which then informed the *actual* selection of episodes. Figure 1 shows an example of some of the information documented for each potential corpus file: the year of first broadcast, the first and second IMDb genre, award nominations/wins, the episode to be included, the typical length of episodes in minutes, the number of total episodes in the season and the position of the episode in the context of the whole season (textual time).



TV series	First broadcast	IMDb Genre 1	IMDb Genre 2	Emmy	Golden Globe	Episode	Ep name	Length	Total eps	Type
Arrested Development	2003	Comedy	N/A	both	yes	22	Let 'em eat cake	22	22	final
Dexter	2006	Crime	Drama	yes (S)	yes	12	Born free	60	12	final
NCIS	2003	Action	Comedy	no	no	1	Yankee white	60	23	pilot
Southland	2009	Crime	Drama	no	no	2	Mozambique	42	7	beginning

Figure 1 Information about television episodes (Excel 2016)

TV series and episodes were then non-randomly chosen for inclusion in the corpus, taking into account the overall aim for a balanced design. One full episode per TV series was selected, with the aim of building a ‘full-text’ rather than ‘sample’ corpus (where the sampling unit might be a 2000-word sample per episode). This means that the integrity of the text is respected (c.f. Sinclair 2005) and is important for episode-based discourse and stylistic analysis. For each chosen episode, dialogue was transcribed from scratch (48 episodes) or on the basis of existing transcripts or scripts (18 episodes) as explained in section 3. Table 2 shows the composition of the final corpus in number of episodes and words, alongside the variables of textual time, ‘quality’ vs. ‘mainstream’ and drama vs. comedy. The table indicates that the corpus is fairly balanced, since it contains 116,295 words from drama genres and 158,779 words from comedy genres as well as 135,887 words from ‘quality’ and 139,187 words from ‘mainstream’ TV series, in addition to a healthy mix of different types of episodes in terms of textual time.

Table 2 Composition of SydTV in number of episodes and words according to Wordsmith (‘tokens in text’), showing the variables of textual time, ‘quality’ vs. ‘mainstream’ and drama vs. comedy (token definition: hyphens do not separate words; ‘ not allowed within word)

SydTV: Number of episodes and words				
	‘Quality’		*‘Mainstream’*	
Textual time	*Drama*	*Comedy*	*Drama*	*Comedy*
pilot episodes	0/0	7/26,671	2/10,053	5/16,779
final episodes	2/10,334	3/10,539	1/3,664	4/14,019
episodes at the beginning	2/8,675	3/9,812	¼,958	3/12,540
episodes in the middle	5/20,314	4/15,272	5/24,065	3/13,361
episodes at the end	3/13,900	5/20,370	4/20,332	4/19,416
Total	12/53,223	22/82,664	13/63,072	19/76,115
	135,887		139,187	

Finally, the corpus contains a mix of broadcast/network television (42 episodes), and cable television (basic cable: 11 episodes, premium cable: 13 episodes). This was not used as a systematic variable during the design stage, as it would have introduced too much complexity. It is nevertheless important that a corpus contain series from such diverse distributors, since differences between these may impact on language use (especially on the use of swear/taboo words because relevant regulations only apply to network television). Further, a corpus that only includes commercial network television series will not be representative of contemporary TV, since shows by HBO (premium cable) or AMC (basic cable) have become important cultural products, including series such as *The Wire* and *Breaking Bad* (both included in SydTV). SydTV contains no Netflix or Amazon originals, since these were not as widespread during corpus design as they are now, and Netflix was not available in Australia until 2015. However, many of the programmes included in SydTV are now distributed, if not created, via such platforms.

3. Information on transcription and transcription conventions

The following sections provide information on the transcription process and conventions.

3.1 Dialogue that was *not* transcribed

Dialogue in recaps ('previously on') and previews/teasers was not transcribed, with the exception of *Arrested Development*, which includes a fake 'teaser' at the end of the episode (*On the next season of Arrested Development…*). These fake scenes continue the episode's story, do not actually occur in future episodes (Mittell 2015: 21) and are hence considered as part of the relevant episode's dialogue. Further, where recaps or flashbacks are part of the narrative they *are* included in the transcription (e.g. in *24* or at the beginning of the *Royal Pains* episode).

Also not transcribed were dialogue fragments that interrupt or are part of the title song (e.g. *Help me* in *True Calling*; *hah*; *oh baby* in *According to Jim*), cases where the character sings part of the title song (e.g. *New Girl*), or statements such as *Hot in Cleveland is recorded in front of a live studio audience*. In contrast, the following dialogue *was* transcribed (where (V) indicates a character speaking in voice-over, and <VOICE:> indicates a voice-over narrator):

- <JANE (V):> See that aspiring model there? That was me, Deb, until the day I died. I thought I'd go straight to Heaven, but there was a bit of a mix-up and I woke up in someone else's body. So now, I'm Jane, a super busy lawyer with my very own assistant. I got a new life, a new wardrobe, and the only people who really know what's going on with me are my girlfriend Stacy and my guardian angel friend. I used to think everything happens for a reason... […] Well, I sure hope I was right. (*Drop Dead Diva*; repeated at the beginning of each episode)
- <EARL (V):> You know the kind of guy who does nothin' but bad things and then wonders why his life sucks? Well, that was me. Every time somethin' good happened to me somethin' bad was waiting around the corner. Karma. That's when I realized I had to change, so I made a list of everything bad I've ever done and one by one I'm gonna make up for all my mistakes. I'm just tryin' to be a better person. My name is Earl. (*My Name is Earl*; repeated at the beginning of each episode)
- <GOSSIP GIRL (V):> Gossip Girl here. Your one and only source into the scandalous lives of Manhattan's elite. [recapped scenes; not transcribed] And who am I? That's a secret I'll never tell. You know you love me. XOXO, Gossip Girl. (*Gossip Girl*; the last four sentences are repeated at the beginning of each episode)
- <VOICE:> Now the story of a wealthy family who lost everything and the one son who had no choice but to keep them all together. It's Arrested Development. (*Arrested Development*)
- <JACK (V):> Right now, terrorists are plotting to assassinate a presidential candidate. My wife and daughter have been targeted, and people that I work with may be involved in both. I'm federal agent Jack Bauer and today is the longest day of my life.
  <VOICE:> The following takes place between seven PM and eight PM on the day of the California presidential primary. (*24*)

The following sections describe the version of SydTV as it has been transcribed originally, while the process of standardisation that subsequently produced SydTV-Std is described in section 4.

3.2 General information and caveats

The transcription conventions used in the creation of the *Sydney Corpus of Television Dialogue* (SydTV) were inspired by and adapted from: ADVICe (itself based on MICASE);[5] CHILDES;[6] and SWITCHBOARD.[7]

Because of the expensive (time-consuming) nature of detailed transcription, a mainly orthographic transcript was produced, with some marked linguistic variants transcribed, as explained below. Transcribers (research assistants) checked every transcript against its video at least once and then after some time double-checked each transcript at least once, too. They were asked to be consistent and to watch for common spelling confusions such as *its* and *it's, they'r*e and *there* and *their*, *by* and *bye*. Much care was taken to ensure the accuracy of the transcription.[8] Nevertheless, some minor inconsistencies remain and human error is still a possibility. Known inconsistencies in the transcripts, which were not corrected, include punctuation: While the transcripts include punctuation symbols (e.g. question marks, exclamation marks), punctuation has not been consistently used to identify aspects such as intonation or speed of delivery. This was initially a desired goal and was used in some of the transcripts, but ultimately had to be abandoned because of funding constraints, as it was too time-consuming. Some partial words (in interruptions) may have been transcribed (e.g. *Lorel*) and occasionally words may start with a capital letter when they should not or are not capitalised when they should be.

Moreover, transcribers might disagree on particular dialogue lines, for example because of speaker overlap, mumbling, speedy delivery, etc. As Adams points out 'each hearing is authentic' (Adams 2013: 232), and in some instances it can indeed be difficult to determine which variant is used (Adams 2013: 233). The following two dialogue lines are a case in point:

<BUBBLES:> I ain't, I didn't asked for nothin'. (*The Wire*)
<LETTIE:> I ain't eat anything all day, like you said. (*True Blood*)

When I played these two examples as audio files to an audience of American speakers, there was no clear consensus about what exactly is being said by the two characters here, especially in relation to *ain't* compared to a (reduced) *didn't*. The details of what viewers heard are discussed in Bednarek (2018). When I played the same audio files to an audience in Australia (with various L1 varieties), results were even more mixed.[9] This is obviously not the case for all dialogue, but some instances may be very hard to understand, especially for those who do not speak the variety of the character dialogue as L1.

It is particularly difficult to transcribe every single word that is uttered in cases where characters talk at the same time (such overlap occurs e.g. in *Workaholics* and *Veep*). Words are at times swallowed in SydTV, and it is sometimes hard to tell if a sound is pronounced or not: For example, it may be unclear if a character is saying *you gonna* or *you're gonna*; *you've been* or *you been*; *do you* or *you*; *in his mind* or *his mind*. It can also sometimes be difficult to distinguish *I* and *a* from *uh* or *oh* from *ah*, etc. Another issue is to determine if an instance should be treated as sigh/groan/real laughter (etc) and therefore *not* transcribed or if it should be treated as interjection/sound/fake laughter (e.g. *ow*, *ah*, *ugh*, *he*, *ha ha*) and therefore *is* transcribed. In addition, it can be hard to determine if something is an elongated vowel merging into stutter (not transcribed) or a repetition of full word (transcribed) – especially in the case of *I*.

Needless to say, a transcript is only ever one version of on-screen dialogue. As many scholars have pointed out, transcription is not a neutral but rather a selective process of analysis, reflecting the researcher's interest and decisions, and resulting in a single, partial, reductive and fixed version (e.g. Toolan 2014: 460-461). Other transcription conventions are viable alternatives; see Bonsignori (2009) and Dose (2013) on transcribing film/TV speech. Because of time and funding constraints, many aspects of spoken discourse could not be captured in this transcription (see Du Bois 1991, Du Bois et al 1993, Clift 2016: 44-63).[10]

The vast majority of episodes were transcribed on the basis of their Itunes versions. This was the cheapest option (since individual episodes can be purchased, rather than having to acquire the whole season) while also being logistically useful (an Itunes voucher could be given to the research assistants to purchase the episodes for download). Some episodes were transcribed on the basis of their DVD versions (for example, if I already owned the relevant DVD), and a minority were transcribed on the basis of other online sources (for example, Netflix). These different versions actually mirror the range of ways in which consumers nowadays experience TV series and this variety is hence not considered problematic. However, one must be aware that minor linguistic differences may exist between original broadcast versions and subsequent incarnations such as DVD versions, including dialogue changes (Mittell 2015: 39), and censorship of swear/taboo words. For example, the Itunes version of the *Workaholics* episode bleeps swear/taboo words, while the DVD version (used here) does not. Variations between scripts, textual versions, and performances are further discussed by Adams (2013).

3.3 Transcription conventions

This section lists the transcription conventions used to transcribe SydTV episodes, including the transcription of semi-lexical features such as voiced pauses, response signals, interjections, discourse markers, phonologically reduced forms, etc. Such features are often not adequately addressed in corpus transcription guidelines, as the survey and review by Andersen (2016) shows. I have aimed to provide both the typical/major meanings and additional notes for such features where applicable.

General notes

Table 3 General notes 1

Dialogue is transcribed in Notepad on PC; encoding: ANSI (default)
Dialogue is transcribed verbatim, without correcting grammatical errors*: I seen him*, *me and him gone to the movies*, etc.
American English spelling is used (e.g. *focused, realize, physicalize, maneuvering, jewelry, whiskey, program* [but: *programmer, programmed, programmable, reprogramming, reprogrammed*]
Word abbreviations are avoided: *Fort Worth*, not *Ft. Worth*, etc.
Laughter, sighs, grunts, clearing throat, etc or background noise are generally not transcribed.
Contextual (non-speech) events (gaze, gesture, actions, etc) are not transcribed. This means that the transcript does not per se identify 'asides' and similar kinds of special utterances, for example.[11] It also does not identify whether a conversation happens in an alternate or parallel reality, in someone's memory, as a flashback or flashforward, etc, unless voice-over is used (indicated as such using (V); see below).
Audible background songs are not transcribed, unless they are lyrics sung by characters. So songs sung by onscreen characters *are* transcribed (e.g. in *Glee, New Girl, Pushing Daisies,* but not if a character sings part of the title song). However, the fact that dialogue is sung is not specified in any particular way (neither is whispering, teary voice, sing-song voice, fading voice, robot-like/computerised voice, reverb/echoing, audio speed-up which becomes unintelligible, etc).
Scene changes and shot types/sequencing are not transcribed.
The medium or channel is not transcribed – whether dialogue takes place face-to-face, on the phone, via two-way radio, etc. Only spoken text in the aural modality is included, rather than written text in the visual modality (so tweets or text messages would only be transcribed if they are read out aloud). Superimposed text (e.g. 'two weeks ago'; 'footage not found') is not transcribed.
Standard orthography is used for most words, even though they may be pronounced with a foreign accent, etc. But: certain **marked** pronunciation variants *are* imitated (see below on non-standard forms)
Punctuation is not *consistently* used to mark intonation contours, syntactic boundaries or pauses, although in some cases a comma marks a pause, while an exclamation mark may indicate shouting or emphasis. In other cases, punctuation may simply follow written norms (i.e., sentences end with a full stop, clauses are separated by a comma).
Capitalisation is also not necessarily always consistent (e.g. 'god' may be spelt *God* or *god*) and should not be used to distinguish usages without double-checking their accuracy.
Spellings of some common words: *alright* (not *all right*); *okay* (not *OK*); *mom* (not *mum*), *a while* (not *awhile*); 'anymore' is spelt *anymore* when it means 'any longer', 'nowadays' (*But you're not alone* ***anymore***; *Actually, she's not my girlfriend* ***anymore***). Otherwise it is spelt *any more* (e.g. *Could this geezer be* ***any more*** *lame? I can't listen to* ***any more*** *of your stupid bullshit words*; *I can't do* ***any more*** *than I've done*).

Table 4 General notes 2

Speaker (character): CAPITALS are used for the names of speaking characters, with a colon attached;[12] unknown or minor characters are referred to as MAN/WOMAN or by their role and are numbered if there are more than one. In the original transcripts, speaker names are tagged using angle brackets (e.g. <MAN:>), but for ease of reading, the tags have been removed from examples in this manual.	REV TIM TOM: MRS ZIMBERG: POOL GUY: MAN: WOMAN 1: WOMAN 2:
Voice-over by character: (V) is used for characters speaking in voice-over	JACKIE (V): Ah morning. Such calm. I'm so quiet and so peaceful.
Voice-over by narrator: VOICE is used for voice-over that is not uttered by a character	VOICE: Chuck continued to keep the secret ingredient of her pies secret.
Audible dialogue emanating from media (e.g. radio, television) is transcribed	RADIO: What does it mean to accept Jesus as your personal Savior?
Non-English utterances are predominantly transcribed in cases where English sub-titles are provided: only the English subtitles are then transcribed and (S) is added to the character name. This means that in cases where (whole) non-English utterances are not subtitled, there is no speaker name or dialogue present in the transcript (and hence no indication that non-English turns have occurred). [*The Shield* and *The Wire* have Spanish mixed with English, not subtitled. Utterances spoken entirely in Spanish are omitted, but where English words occur with Spanish ones alongside each other within an utterance, both are transcribed.]	1 JIN (S): What's going on between you and him? SUN(S): Who? JIN(S): Michael. 2 CLAUDETTE: Manuel Ruiz! Ha visto este hombre? There. DUTCH: Hey, hey hey you. You, amigo. Amigo.
+BLEEP+ is used for bleeps.	GEORGE SENIOR: Oh, I've got the worst +BLEEP+ attorneys.
Reading and quoting passages (character reads something verbatim or clearly quotes) are marked with double quotation marks " ".	TERRELL: Finally, he just opens up his mouth and says, "I guess you gonna have to tear my ass apart, homes". CHARLIE: Oh man, so well what did you do? TERRELL: Took him to the back alley, and I tore his ass apart! MAC: I love that guy. "I took him to the back alley"? Who does that shit? It's like a movie. REESE: With Stevie "the wheelie" Kenarban?
Titles are used without punctuation	*Mr* *Mrs* *Miss* *Doctor, doctor* *Sir* *Ma'am*
No symbols or accents are used	*per cent*, not *%*; *dollars*, *cents*, not *$* *usee*, not *usée* *and*, not *&*
Proper nouns (names, departments, organisations, etc) and the pronoun *I* (and acronyms; see below) are capitalised; the beginnings of turns are capitalised	Jim, I KID1: Use more ketchup. JIM: Hey, Dana!

Turn-taking

Table 5 Turn-taking

Lines are used to keep speakers separate	LACEY: Alright, I'll give it a shot. CHARLIE: No no no no no, no, that's not Bobo, that's Ed.
Two or more speakers, in unison (saying the same): repeated twice, with each speaker separately identified	RANDY: Drink, drink, drink, drink, drink, drink. MAN 1: Drink, drink, drink, drink, drink, drink.
Two or more speakers (overlap): written in separate lines (overlap not explicitly identified)	MAC: You're not making this very easy. Stop being a dick. CHARLIE: You're trying to impress Terrell with a couple of black friends. I'm not being a dick!
Dot points are used in two situations: 1 When one person's utterance is cut off by another's, the interrupted utterance receives an ellipsis. 2 When an utterance becomes audible mid-sentence, dot points are attached to the beginning of the first intelligible word. Self-interruptions are either marked by commas or dot points.	1a ELI GOLD: If I could be so bold... ALICIA: No, you can't. 1b JOHN: Liv, put the gun... OLIVIA: Stop lying to me, John. 2 TV COMMENTARY: ...Kenny Powers a career that once showed so much promise, you can feel the sun going down on it right here.
Full repetitions of a word or phrase are transcribed. Partial words tend not to be transcribed (e.g. if a character wants to say *she never answered*, but is interrupted after *she never an-*, this is transcribed as *she never…*; similarly, if they stutter, e.g. *w-what* is transcribed as *what*; *y-yeah* is transcribed as *yeah*.	CHARLIE: It's not all your fault. I, I probably haven't been the best son. LEVI: Oh no I'm, I'm I'm I'm very busy as you can see, come on!

Filled/voiced pauses, discourse markers, backchannel/listening/response cues or signals, hesitation markers/word searches, (dis)agreement markers, exclamations, interjections, attention seekers, etc

Table 6 Filled pauses, etc

Transcription	**Typical/major meanings**	**Notes (where applicable)**
uh huh	agreement/yes	open-mouthed
mm-hmmm	agreement/yes	bilabial
mmm	agreement/comprehension/contemplation	bilabial
uh-uh	disagreement/no	open-mouthed
uh-uh-uh	disagreement/no (chiding)	open-mouthed
mm-mmm; *mm-mm-mm*	disagreement/no	bilabial
hmm	thinking, waiting, questioning	with varying length and intonation
uh	filler, word search	with varying length
umm	filler, word search	with varying length
ugh	revulsion/disgust/exasperation	This can be pronounced in different ways and no distinction is made between variants
oh	surprise/comprehension/discourse marker…	dipthong [oʊ], with varying length
ah	e.g. yelling, realisation, satisfaction	[ɑː] or [ɑ] vowel sound with varying length but may be difficult to distinguish from [oʊ] or [ɔ] or [ɔː] (e.g. in some occurrences of *oh my god*)
oh-oh	(usually negative) surprise	first vowel monophtong, second vowel diphtong
ooh	usually excitement/curiosity	high back vowel with varying length, often long [uː]
heh	dry/fake laugh (but not normal laughter)	pronounced as [he] or variants thereof

hootie hoo	greeting (‘hello’)	
toodle-oo	greeting (‘bye’)	
huh	usually questioning (often as tag) or surprise	
whoa	surprise/dissatisfaction (‘stop’)	
woah	sound to halt horses	
wow	amazement	
aw	‘how cute/sweet’ or expression of sympathy	
mwa	‘kissing’ sound (but not normal kissing)	
sh	shushing (‘silence’)	
psst	shushing (‘silence’)	
yo	seeking attention (‘hey’)	
oi	seeking attention (‘hey’)	
booyah	triumph/joy	
shazam	triumph/achievement	
tada	elation/joy/achievement	
yay	pleasure/joy/cheer	
yippee	joy/other positive emotion	
woo	joy/excitement	
yee-ha	excitement	
attaboy	approval/admiration	
ha ha; *ha*, *ha*; *ha ha ha*	[fake] amusement, enjoyment (but not normal laughter)	
ha	triumph; dismissal; other emotion	
pfft	dismissal	
ow	expression of pain	
ew	revulsion/disgust	

Some other sounds/interjections are also transcribed as heard (but not all sounds are transcribed, e.g. someone imitating a buzzer sound, sighs, groans, etc, as noted above):

- aha;
- eh;
- ey;
- ho;
- ho-ho;
- woo-ho-hoo;
- oh-ho;
- oh-ho-ho-ho;
- blah blah blah;
- ba-ba ba-ba, ba-ba-ba;
- lalalalalalala;
- nanananana;
- blah dee bloo dee blah blah;
- giggle, giggle, choochee, choochee;
- la-di-da

Contractions, phonologically reduced forms, colloquialisms

Table 7 Contractions, etc

Contractions of *is*, *am*, *are*, *had*, *have*, *do*, *would* are transcribed as such. Some examples are on the right. Note that it is often difficult to tell if *have* is contracted (‘*ve*) or not (*have*), as there are many in-between cases where instances are not clearly either. One option was chosen depending on the markedness of the contraction.	*I’d*, *I’m*, *I’ll*, *she’s*, *she’ll*, *he’s*, *they’re*, etc *‘ve* for ‘have’ (e.g. *I’ve*, *they’ve*, *could’ve*, *must’ve*, *should’ve*, *might’ve*, *couldn’t’ve*) *‘d* for ‘did/do/would’ (e.g. *how’d*, *what’d*, *why’d*, *where’d*, *that’d*, *there’d*)

Shortened, assimilated and otherwise marked pronunciation variants or colloquial forms are transcribed as shown on the right[13]	*oughta* – ‘ought to’ *woulda* – ‘would have’ *coulda* – ‘could have’ *kinda* – ‘kind of’ *gonna* – ‘going to’ *gotta* – ‘got to’ *wanna* – ‘want to’ *lotta* – ‘lot of’ *sorta* – ‘sort of’ *c’mon* – ‘come on’ *lemme* – ‘let me’ *gimme* – ‘give me’ *whatcha* – ‘what (do/did) you’ *gotcha* – ‘got you’ *y’all* – ‘you all’ *‘cause* – (‘because’ – includes potential variants like kɔːz, kɔz, kɒz, kəz, kʌz, etc) *‘em* – ‘them’ *nothin’*; *goin’*; *holdin’*; *gettin’* (obvious/marked instances of alveolar variant) *youse* – pluralised ‘you’ *cuz* – truncated ‘cousin’ *nope* – ‘no’ *yep* – ‘yes’ *yeah* – ‘yes’ *‘til* – ‘until’ *‘round* – ‘around’ *outta* – ‘out of’ *bro*(*s*) – ‘brother(s)’ (*little bros before big bros*) *hos* – ‘whores’ (*bros before hos*)

Acronyms, abbreviations, letters as variables

Table 8 Acronyms, etc

Capitals denote the spelling out of each letter, as in acronyms such as *FBI*.	*FBI* – ‘ef bee I’ *CIA* – ‘see I ay’ *CHAMAID* – ‘see aitch ay em ay I dee’ *LEO* – ‘el ee oh’ vs. *leo* – ‘leo’ *ID* – ‘ai dee’ *X* – ‘ecs’ [slang for *ecstacy*] *ALS-ing* [‘ay el es ing’]

Numbers

Table 9 Numbers

All numbers are fully spelled out as words	*one* – 1 *two* – 2 (etc) *three-D* – ‘3-D’ *o* – 0 (zero) spoken as ‘oh’ *point* – decimal point (.) *nineteen sixty-four* *BE five years old*, *ten months old*, etc *his eighteen-month-old sister*; *a six-year-old*, etc

Hyphenation and compounds

The hyphen character (-) is used for hyphenation only, never for a break between clauses or the introduction of a new phrase.

Table 10 Hyphenation and compounds

Consistency was the aim, but given the large number of different compounds, human error is a possibility. The potential inconsistency in the spelling of compounds should be kept in mind when searching for word forms or interpreting word lists.	Some words spelt with hyphen are presented on the right	*low-carb* *low-rent* *wishy-washy* *vo-tech* *A-lister* *a run-of-the-mill procedure* *level four-A* *Miss Fluff-and-Fold* *J-to-the-Bieber* *you're cha-cha-cha-ing* *the risks of plus-one-ing me* *Social-Skills-athon* *long-lensed* *high-yield* (adj) *one-night stand*; *every-day stand* *four-wheel drive* *force-feed* *off-limits* *double-check* *vice-president* *much-needed* *good-bye*; *bye-bye* *mix-up* (N) *break-in* (N) *well-liked* *a one-room shack* *the three-session deadline* *last-minute* (but: *at the last minute*) *door-to-door* *play-by-play* *face-to-face* *grown-up/grown-ups* (N) *mother-daughter* + NOUN (e.g. *mother-daughter situation*)
	Some examples of two or more words spelt without hyphen are presented on the right	*jet ski* *soul mate* *blood work* *baby backs* *band member* *neck brace* *tox screen* *ska band* *drum roll* *punch line* *rain check* *gift wrap* *time bomb* *gen Xer* *meth head* *beach house* *cop killer*; *dog killer*; *psycho killers*; *Son of Sam killer*; *serial killer/s*; *ice truck killer* *girl scout/s*; *boy scout*; *major league scout*; *TV scouts* *high school* *cat fight/s* *the White House* *fairy tale* (N)

		cream cheese *ice cream* *hand cream* *credit card/s* *game show* *phone call/s* *pay phone* *hip hop* *real estate* *deal breaker* *foot rub* (N) *eye pads* *board member* *crime fighters* *neighborhood watch* *Long Island iced tea* *a magic eight ball* *a walnut marble top vanity* *rush hour gridlock* *per cent* *the men's room*; *the ladies' room* *board member* *belly button* *gift wrap* (V) *all natural* (adj) *show time* *bubble gum flavored* *good night* *parents' day* *a make out sesh* ('session') *up front* (adv) *in between* *super hot*; *super busy*; *super lucky* (etc) *some fleabag motel* *a million to one shot*
	Some words spelt as one word are presented on the right	*ballroom*; *bathroom*; *bedroom*; *classroom*; *courtroom*; *showroom* (but: *chat room*; *conference room*; *dining room*; *dressing room*; *home ec room*; *emergency room*; *family room*; *guest room*; *hospital room*; *hotel room*; *laundry room*; *living room*; *locker room*; *meeting room*; *motel room*; *screening room*; *waiting room*) *boyfriend*, *girlfriend* (but: *lady friend*; *dyke friend*) *email* (not *e-mail*) *webcam* *manscape* *motorboating* *bowhunting* *rejigger* *spokesmodel* *crustless* *handicapable* *powwow* *unboyfriendable*; *untutorable* *pornalicious* *misjoke* *crimelords* *supervillains* *unbag*; *unhear*; *unspell* *reread* *midday* *setup* (N) *breakup* (N) *overthink*; *overinvest*

		hungover *mainstreaming* *coordinate* *metrosexual* *buzzkill* *birthdate* *backworthy* *freestyling* *roommate* *carjacking* *eyewitness* *weekend/s* *striptease* *wheelchair/s* *stepsister* *cheerlead* (V); *cheerleader* *homework* *headmaster* *granddaughter*; *grandmother, etc* *earring/s* *nosebleed* *madman* *dollhouse/s* *setback/s* *kickoff* (N) *amazeballs* *Superman* *ballbuster* *multiunits*

Swear/taboo words, euphemisms

Table 11 Swear/taboo words, euphemisms, abusives, etc

Spelling used	*jeez* (not *geez*) *coochie* *toosh* *frigging, fricking, freaking*
One word	*goddamn* (for *God damn*) *goddamnit* (for *God damn it*) *damnit* (for *damn it*) *motherfucker/ing*; *asshole*; *blowjob*; *cocksucking*; *fuckable*; *shitheads* *jackass*; *badass* (but*: ... with your bad ass*) *fucktard*; *dickwad*; *dickweed*; *dickhead*; *dipshit; fuckwad*; *dingwit* *bullshit*
hyphenated	*god-awful* *big-ass*; *bitch-ass*; *dumb-ass*; *smart-ass*; *kick-ass*; *tight-ass*; *lame-ass*; *short-ass*; *ugly-ass*; *smooth-ass*, etc (but: *this movie is going to kick ass*; *tell her to get her skinny ass in here*); *ass-fucked*; *ass-kicking*; *ass-clown*, etc (hyphenated, but: *ass ache*; *ass burger*) *shit-faced*; *shit-ass*; *pencil-fucked*; *ass-fucked*; *butt-fuck*; *fuck-up*; *fucked-up* (pre-noun; but: *that will fuck a kid up*; *you fucked up*, etc) *limp-dick*; *dick-wagging* *a screw-up*; *screwed-up* (pre-noun, but: *I screwed up*, etc) *piss-pants* *poo-poo*; *poo-head* *jerk-off* (noun, but *jerk off* [verb])
two words	BE *pissed off* *holy shit*; *holy crap*; *holy moly* *douche bag*

Other (spelling of other words, e.g. slang, nicknames, abbreviations, idioms, cultural products, innovations, other)

Table 12 Other

minuscule
smush; *smushed*
brunet (m); *brunette* (f)
blond (m); *blonde* (f)
prenup; *postnup*
blondie; *indie*
nana; *grandmama* ('grandmother')
play hooky; *go in halvesies with you*
Mrs Ladypant
perps; *perv*
doozy; *skeevy*; *icky*; *smirky*; *pissy*; *hinky*; *shrimpy*; *talky*; *archy*; *queeny*; *insidey*; *abortiony*
awol; *veep*; *potus*; *home ec*; *special ed*; *Chud*
Ponzi scheme
rhino [for *rhinoplasty*]; *lipo* [for *liposuction*]; *boty* (for *botox*); *syph* [for *syphilis*]; *hep C* [for *hepatitis c*]
barbie [for *barbecue*]; *intel* [for *intelligence*]; *comm* [for *communication*]; *decomp* [for *decomposition*]; *pic* [for *picture*]; *ref* [for *referee*]; *mic* [for *microphone*]; *hon* [for *honey*]; *plex* [for *plexitheater/multiplex*]; *sesh* [for *session*]; *nugs* [for *nuggets*]; *benzos* [for *benzodiazepines*]; *haps* [for *happenings*]; *rezzy* (for *reservation*]; *vic/s* [for *victim/s*]
sicko
preggers
skank
meanie
phat
nunchakus; *bubkes*
jeggings
jit
ghostbuster
anesthetic
dictagram
eastview
podule
niblet
get vaped out; *vapetron*
bieberites; *belieber*; *belieb*; *Bieberhole*
conkle
Trailerville
spankitude
alrighty
Mad libs
Kettle corn
'*Bucks* [for *Starbucks*]
Nascar moms
Scratch 'n sniff

4. Producing a standardised version of SydTV: SydTV-Std

To produce a partially standardised version of SydTV I used Wordsmith's Text Converter function to standardise the lexicogrammatical items listed in Table 13. I focused on items that I knew from pilot studies to have a significant influence on the calculation of keyness. Most of the variants were automatically standardised using a conversion file (settings: within file conversion, whole word, not case sensitive), with some changes made separately (using the 'just one change' setting). A test file was converted first and results checked, before SydTV was converted. In addition, I manually changed every instance where the alveolar form [In] was used in words ending in <ing> (e.g. *somethin'*, *goin'*, *fuckin'*) to *-ing* (e.g. *something*, *going*, *fucking*). I did not standardise any contractions, since I was interested in their occurrence and since they are also present in most other corpora of spoken English. The partially standardised version of SydTV is known as SydTV-Std.

Table 13 Variants and their standardised forms

Variant	Standardised form
gonna	*going to*
wanna	*want to*
gotta	*got to*
oughta	*ought to*
coulda	*could have*
kinda	*kind of*
woulda	*would have*
shoulda	*should have*
lotta	*lot of*
sorta	*sort of*
c'mon	*come on*
lemme	*let me*
gimme	*give me*
gotcha	*got you*
'em	*them*
y'all	*you all*
'til	*until*
'round	*around*
outta; outa	*out of*

5. Versions and access to other scholars (via CQP web)

The corpus that I have described in this manual is *SydTV Version 3.0*, although I referred to it simply as SydTV. SydTV Version 1.0 was used in an Honours thesis (Price 2015), while SydTV Version 2.0 was used by Bednarek (2017). These versions had not yet been proof-read by me, nor had they been corrected by the fourth research assistant (see note 8). Further, Version 1.0 contained episodes from two series later excluded because they were erroneously identified as US-American (*Wilfred*, an Australian comedy, and *Rome*, a British-American-Italian drama). Version 2.0 accidentally included the final episode of the *third* season of the sitcom *Two Broke Girls*, which was replaced with the final episode of the first season.

Version 3.0 was used in Bednarek (2018) and other publications (see syd-tv.com) and is the basis for the frequency and keyness lists available at www.syd-tv.com. The version that is available to other scholars via CQPweb is *SydTV Version 4.0*, which differs very slightly from Version 3.0, as described in the appendix below. As noted above, this version is lemmatised and part-of-speech tagged and uses XML-compatible tags for speakers (i.e. <u who="JACKIE"> Hey. </u>). Both the original and the standardised version of the corpus are in fact available in two different formats: One format uses the CLAWS part-of-speech tagger and USAS semantic tagging, while the other format uses Treetagger's part of speech tagger, with one modification ('IN/that' for *that* as complementiser was replaced with 'CJT' from the CLAWS5 tagset, because of problems associated with the slash in the tag). The treetagger version does *not* include semantic tags. It is recommended that users access the CLAWS versions to make use of all functionalities (see Table 14). It is important to be aware that part-of-speech tags (and semantic tags) were assigned by an automatic software tool and are not always correct. Regardless, these CQPweb versions allow users to incorporate lemmas, part-of-speech tags, and semantic tags into their corpus searches and analyses.

Table 14 QCPweb SydTV 4.0 versions

Corpus name	**Details**
Sydney Corpus of Television Dialogue (SydTV) CLAWS	Original version; processed using the CLAWS POS tagger; Oxford Simplified Tagset also used; includes semantic tagging using the UCREL semantic analysis system (USAS)
Standardised Sydney Corpus of Television Dialogue (SydTV-Std) CLAWS	Standardised version; processed using the CLAWS POS tagger; Oxford Simplified Tagset also used; includes semantic tagging using the UCREL semantic analysis system (USAS)
Sydney Corpus of Television Dialogue	Original version; processed using TreeTagger (lemmatisation, POS tagging); no semantic tagging, no Oxford Simplified Tags
Standardised Sydney Corpus of Television Dialogue	Standardised version; processed using TreeTagger (lemmatisation, POS tagging); no semantic tagging, no Oxford Simplified Tags

The Simple Query Syntax provided at the CQPweb interface works for all of the four corpora, except for functionalities that were not included in the relevant corpus. When searching for a specific POS tag, the appropriate tagset needs to be used in the query syntax, depending on the version that is accessed (TreeTagger or CLAWS). The only XML tag that is used is the tag for speakers/turns, which can be used in any search. For example, searching for *<u> hey* retrieves all instances of *hey* that occur at the beginning of a speaker's turn. Searching for *alright *</u>* retrieves instances of *alright* at the end of a speaker's turn.

Note that there is a known display issue with CQPweb, where word forms in a frequency list are sometimes presented with initial lower case and sometimes with initial upper case. This is purely a display issue and does not affect the frequency counts which are not case-sensitive unless specified. For example, counts for 'IT' and 'My' in a frequency list include all instances of the forms (*it, It*; *my, My*). **Beware also the specific token definition used within CQPweb (see above).**

The availability of corpus metadata allows user to search only particular texts in the corpus (e.g. 'censored' episodes), using the 'restricted query' option. Genres are not included in the restricted query option, but users can still use the IMDb genre categorization to build sub-corpora. For example, if users want to build a corpus that includes all texts that are categorized as comedy (either as genre 1 or as genre 2; see above), they can follow these instructions:

1. Go to **Create/edit subcorpora**
2. Choose to **Define new subcorpus via: Scan text metadata** and click on **Go!**
3. In **Which metadata field do you want to search?** choose IMDB Category 1
4. In **Search for texts where this metadata field .... contains** write the category (Action, Adventure, Comedy, Crime, Drama) and click on **Get list of texts**
5. Write the name for the new subcorpus, click on **include all texts** and click on **Add texts**
6. If you want to include further texts in the subcorpus, do the same for IMDB Category 2 (categories: Adventure, Comedy, Crime, Drama, Fantasy, Mystery, Romance, Scifi, Sport), and add the text(s) to the new subcorpus you created in step 5.
7. Compile Frequency list
8. On the **Standard query page**, in the **Restriction** box, select the subcorpus

A screencast of these steps is available here.
(Warning: The IMDb genre categorization is not without faults.)

Other corpus metadata can also be used to build sub-corpora (for keywords analysis) or in restricted queries (without the need of building sub-corpora first): censorship, year of first broadcast, TV show, textual time (beginning, middle, end; pilot, final), type (mainstream/quality). A more comprehensive video can be found on the CQPweb Youtube channel, which explains how to create subcorpora (https://www.youtube.com/watch?v=huYhoS64fQQ).

Information on how to access the corpora via CQPweb is provided at www.syd-tv.com

Appendix

Changes made to transcripts included in SydTV Version 4.0:

1. The tagging of speaker names is different, as explained above.
2. *NCIS*: corrected *your pathetic* to *you're pathetic*
3. *Hot in Cleveland*:
   Corrected turn assignation:

   <VICTORIA:> Ah so the ladies love Max, huh? Especially Agnes Bratford, or should I say fat-ass hag-ford.
   →
   <VICTORIA:> Ah so the ladies love Max, huh?
   <ELKA:> Especially Agnes Bratford, or should I say fat-ass hag-ford.

   Specified quoted thought via double quotation marks:

   <VICTORIA:> Yeah sometimes we'd pull out of our driveways at the same time and, as our electronic gates were opening, I'd think I don't even know their names, and they don't know mine, and the world was good.
   →
   <VICTORIA:> Yeah sometimes we'd pull out of our driveways at the same time and, as our electronic gates were opening, I'd think **"**I don't even know their names, and they don't know mine**"**, and the world was good.

   More appropriate representation of a non-lexical item:

   <MELANIE:> Elka why, uh, I mean, why?
   →
   <MELANIE:> Elka why, **huh**, I mean, why?

4. *Weeds:* specified that speakers are shown on TV screen (a video recording that Celia is watching)

   <ASIAN:> Oh, Dean.
   <DEAN:> Goddamnit! Come on!
   <ASIAN:> Dean!
   <DEAN:> Put it in, Helen! Put it in! Oh, my God!
   <QUINN:> Fuck you.
   <CELIA:> That little cunt.
   <QUINN:> Fuck off.
   <CELIA:> I should have had an abortion.
   →
   <ASIAN **on TV screen**:> Oh, Dean.
   <DEAN **on TV screen**:> Goddamnit! Come on!
   <ASIAN **on TV screen**:> Dean!
   <DEAN **on TV screen**:> Put it in, Helen! Put it in! Oh, my God!
   <QUINN **on TV screen**:> Fuck you.
   <CELIA:> That little cunt.
   <QUINN **on TV screen**:> Fuck off.
   <CELIA:> I should have had an abortion.

5. *Pushing Daisies*: clarified that the speaker interrupts himself rather than abbreviating the adjective *ridiculous* to *ridic*:
   Changed this:

   <NAPOLEON LENEZ:> This is ridic! How dare! You planted that sock! I am not going to stand here and be accused. I think it's best if you both leave.
   →
   <NAPOLEON LENEZ:> This is ridic**...!** How dare! You planted that sock! I am not going to stand here and be accused. I think it's best if you both leave.

6. *The Office*: specified via double quotation marks that characters are imaginatively quoting (mimicking) other characters:

   <JIM:> Okay, sh. Stop, whatever you’re doing ‘cause this is going to be good. Hi, my name’s Dwight Schrute and I would like to buy a purse from you. Good Lord! Look at these purses. This is something special. Oh, my God. Is this Salvatore de Chiniasta?
   <PAM:> Oh definitely, definitely step in and out of it like that.
   <JIM:> Yes.
   <PAM:> Yeah. You put it on...
   <JIM:> Well, I want to stress test it, you know, in case anything happens. Oh!
   <PAM:> Oh!
   <JIM:> That was really, this is necessary to do to really give it a good workout. This is the, oh! This is the prettiest one of all. I’m going to be the prettiest girl in the ball. Oh, how much?
   →
   <JIM:> Okay, sh. Stop, whatever you’re doing ‘cause this is going to be good. **“**Hi, my name’s Dwight Schrute and I would like to buy a purse from you. Good Lord! Look at these purses. This is something special. Oh, my God. Is this Salvatore de Chiniasta?**”**
   <PAM:> **“**Oh definitely, definitely step in and out of it like that.**”**
   <JIM:> **“**Yes.**”**
   <PAM:> **“**Yeah. You put it on...**”**
   <JIM:> **“**Well, I want to stress test it, you know, in case anything happens. Oh!**”**
   <PAM:> **“**Oh!**”**
   <JIM:> **“**That was really, this is necessary to do to really give it a good workout. This is the, oh! This is the prettiest one of all. I’m going to be the prettiest girl in the ball. Oh, how much?**”**

7. *The Wire:*
   <BUBBLES:> I ain’t, I didn’t asked for nothin’.

   As already mentioned above, when I played this example as audio file to an audience of speakers of American English, there was no clear consensus about whether Bubbles says *I ain’t, I didn’t* or *I didn’t, I ain’t* (both with reduced forms of *didn’t*) or *I ain’t, I ain’t* or *I ain’t even*. However, there did seem to be consensus that the character says *ask* rather than *asked* (transcribed), so I changed *asked* to *ask* in the original and standardised SydTV version 4.0:
   <BUBBLES:> I ain’t, I didn’t **ask** for nothin’.
   <BUBBLES:> I ain’t, I didn’t **ask** for nothing.

8. *Dollhouse*:
   The episode features a television program at the beginning, with ‘vox pops’ interspersed throughout the episode. In the transcript for version 3.0 this was not specified:

   <REPORTER:> When you hear the words dollhouse you probably think of little girls playing tea party, but for some people in Los Angeles those words have a different meaning, a darker meaning.
   <MAN 1:> Yeah, everyone knows that. They got people programmed to do whatever. They could be for sex or, you know, kill a guy. They’re out there. Dolls.

   In version 4.0, I added ‘on TV program’ after the speaker names throughout the episode to specify that these utterances are vox pox and part of a television programme, e.g.:

   <MAN 1 **on TV program**:> Yeah, everyone knows that. They got people programmed to do whatever. They could be for sex or, you know, kill a guy. They’re out there. Dolls.

   In addition, the episode features a conversation between two characters (Paul and Mellie) being watched on a computer screen by other characters. This was also specified in version 4.0:

<MELLIE **on computer screen**:> Joel Mynor from Bouncy the Rat? He was on the cover of Wired.
<PAUL **on computer screen**:> You read Wired?
<MELLIE **on computer screen**:> You can see the cover in stores. Brownish hair, pudgy, kinda cute?
<PAUL **on computer screen**:> I don’t remember him as cute.

9. *True Blood*. After the analysis for Bednarek (2018) had long been completed I watched the *True Blood* episode again and noticed that some of the utterances were thoughts in voice-over. When I checked these against the transcript I noticed that two utterances had not been identified explicitly as voice-over. I then decided to check the whole manuscript against the episode again and made several other changes to the transcript included in SydTV Version 4.0. This does not necessarily mean that the transcript in SydTV Version 3.0 is ‘incorrect’, as some of the changes simply reflect different linguistic experiences or the subjectivity of transcription. Note for example that the transcriber only transcribed obvious/marked instances of the alveolar variant as *‘in* rather than *ing* (e.g. *somethin’* / *something*), and one can disagree about what counts as ‘marked’. In certain cases it is also difficult to tell if a speaker says *we’re gonna* or *we gonna*; *I’m gettin*’ or ‘*gettin*’, *this is* or *this*. Changes were made, as appropriate, to both the original and partially standardised version of the corpus.

   Even though these modifications would only result in minor changes (for example with respect to the frequency of particular words) rather than major changes to the tendencies and trends described in Bednarek (2018), I have listed here all of the changes made to the episode transcript, for the sake of transparency:

   Lexical changes:

   - So how about holy water. → So **what** about holy water**?**
   - Her tiny little legs look so darn smooth → Her tiny little legs**, flexible and smooth.**
   - No hair anywhere on her body, all mine. → No hair anywhere **about** her body, **oh my**.
   - But, here I am, having just had → But, here I am, **I mean**, just had
   - you all jacked up on caffeine → you all jacked up on **the** caffeine
   - I just need to → I just **uh** need to
   - Goddamn, son of a bitch and → Goddam, son of a **bitchin’**
   - in that big of a hurry → in that **big a** hurry
   - Uh, Sam, I, I’m sorry → Sam, **uh,** I’m sorry
   - Tryin’ to borrow money → Tryin’ to borrow **some** money
   - And, I think → And, I **mean**
   - You go on ahead hooker → you go ahead on cookin’
   - I know what you’re looking for → **Sit with me.** I know what you’re looking for. **Come on**.
   - weren’t for little Stackhouse bitch → weren’t for **that** little Stackhouse bitch

   Pronunciation changes:

   - Three instances of *gram* changed to *gran*.
   - 24 instances of **ing* changed to **in’* (alveolar variant)
   - Two instances of *out of* changed to *outta*
   - One instance of *might’ve* changed to *mighta*
   - One instance of *kind of* changed to *kinda*
   - Two instances of *we’re gonna* changed to *we gonna*
   - One instance of *them* changed to *‘em*
   - You won’t gimme → you won’t **give me**
   - I’m gettin’ eaten alive → **Gettin’** eaten alive
   - This is a family place →**This a** family place
   - The setting is crucial → **Setting** is crucial.

Speaker changes:

<UNCLE BARTLETT:> → <BARTLETT:>

<ROYCE:> Well, that ain't right, him comin' in here like that! Ain't right them things even exist.
<GUY:> Well, it is a full moon tonight.
→

<ROYCE:> Well, that ain't right, him comin' in here like that!
<GUY 1:> Ain't right them things even exist.
<GUY 2:> Well, it is a full moon tonight.

<TARA:> Lure it out? With a bunch of rocks?
<MISS JEANETTE:> Uh huh.
<TARA:> Don't you need a Ouija board and some chicken bones?
→
<TARA:> Lure it out? With a bunch of rocks? Uh huh. Don't you need a Ouija board and some chicken bones?

<AMY:> No, see, the V adapts. It wants to be in us. We honor Gaia, and seek the deepest relationship to her.
→
<AMY:> No, see, the V adapts. It wants to be in us.
<JASON:> Huh.
<AMY:> We honor Gaia, and seek the deepest relationship to her.

<SHERIFF:> It was probably arson. We can all see that. Now we know one way to get rid of 'em. Excuse me.
<POLICE 1:> Special of the day, country-fried vampire.
<POLICE 2:> This'll take the heat off of them having to find out who's killing those women.
<SOOKIE:> Is Bill in there?
<POLICE 1:> No way of knowin'. They're awful messy.
<ANDY BELLEFLEUR:> But there was four of 'em.
→
<SHERIFF:> It was probably arson. We can all see that. Now we know one way to get rid of 'em.
<OFFICIAL:> Excuse me. Stay out of the way here, ma'am.
<SHERIFF:> Excuse me.
<POLICE 1:> Special of the day, country-fried vampire.
<POLICE 2:> This'll take the heat off of them having to find out who's killing those women.
<POLICE 1:> Oh yeah.
<SOOKIE:> Is Bill in there?
<SHERIFF:> No way of knowin'. They're awful messy.
<ANDY BELLEFLEUR:> But there was four of 'em.

Other changes:

In two instances I clarified that an utterance is voice-over, by adding (V) after the speaker name.

In two cases, I used double quotation marks to specify that an utterance is quoted:

<TARA:> Better? Or you want me to call? Hi sailor, it's me, the girl you been fuckin', mind if I drop by to interrupt your cussing spell, say hi to you and your cute little dog?
→
<TARA:> Better? Or you want me to call? **"**Hi sailor, it's me, the girl you been fuckin', mind if I drop by to interrupt your cussing spell, say hi to you and your cute little dog?**"**

<AMY:> No, it's talking with your teeth clenched together so you don't get lines in your face. Amy, please tell me you're not having sex with that disgusting man.
→
<AMY:> No, it's talking with your teeth clenched together so you don't get lines in your face. **"**Amy, please tell me you're not having sex with that disgusting man.**"**

Table 15 List of all episodes included in SydTV

TV series	Year	Episode Nº	Episode name
According to Jim	2001	4	Anniversary
Anger Management	2012	1	Charlie goes back to therapy
Arrested Development	2003	22	Let ‘em eat cake
Baby Daddy	2012	1	Pilot
The Big Bang Theory	2007	16	The peanut reaction
Birds of Prey	2002	1	Pilot
Bones	2005	20	The graft in the girl
Breaking Bad	2008	3	And the bag’s in the river
Castle	2009	2	Nanny McDead
Community	2009	1	Pilot
Desperate Housewives	2004	19	Live alone and like it
Dexter	2006	12	Born free
Dollhouse	2009	6	Man on the street
Drop Dead Diva	2009	12	Dead model walking
Eastbound & Down	2009	1	Chapter 1
Enlightened	2011	1	Pilot
Entourage	2004	7	The scene
Fringe	2008	13	The transformation
Gilmore Girls	2000	11	Paris is burning
Girls	2012	3	All adventurous women do
Glee	2009	9	Wheels
Gossip Girl	2007	17	Woman on the verge
Grey’s Anatomy	2005	9	Who’s zoomin’ who?
Hot in Cleveland	2010	5	Good neighbors
House	2004	18	Babies and birthwater
How I Met Your Mother	2005	12	The wedding
Human Target	2010	11	Victoria
In Treatment	2008	13	Sophie: week three
It’s Always Sunny in Philadelphia	2005	1	The gang gets racist
Jericho	2006	14	Heart of winter
Legend of the Seeker	2008	14	Hartland
Lost	2004	17	… in translation
Malcolm in the Middle	2000	1	Pilot
Mike & Molly	2010	3	First kiss
Modern Family	2009	15	My funky valentine
My Name is Earl	2005	21	The bounty hunter
NCIS	2003	1	Yankee white
New Girl	2011	1	Pilot
Nurse Jackie	2009	3	Chicken soup
Outsourced	2010	22	Rajiv ties the baraat, pt 2
Parks & Recreation	2009	6	Rock show
Prison Break	2005	16	Brother’s keeper
Pushing Daisies	2007	7	Smell of success
Royal Pains	2009	12	Wonderland
Southland	2009	2	Mozambique
Suits	2011	10	The shelf life

Supernatural	2005	19	Provenance
Teen Wolf	2011	12	Code breaker
The Big C	2010	1	Pilot
The Good Wife	2009	21	Unplugged
The Middle	2009	23	Signals
The New Adventures of Old Christine	2006	1	Pilot
The Office	2005	6	Hot girl
The Shield	2002	4	Dawg days
The Vampire Diaries	2009	22	Founder’s day
The Wire	2002	9	Game day
Thirty Rock	2006	11	The head and the hair
Tru Calling	2003	15	The getaway
True Blood	2008	7	Burning house of love
Two and a Half Men	2003	6	Did you check with the captain of the flying monkeys?
Two Broke Girls	2011	24	And Martha Stewart have a ball, pt 2
United States of Tara	2009	8	Abundance
Veep	2012	1	Fundraiser
Weeds	2005	1	You can’t miss the bear
Workaholics	2011	8	To friend a predator
24	2001	20	7.00 - 8.00 pm

References

Adams, M. (2013). Vignette 13b. Working with scripted data. Variations among scripts, texts, and performances. In: C. Mallison, B. Childs and G. van Herk, eds., *Data Collection in Sociolinguistics. Methods and Applications*. New York/London: Routledge, pp. 232-5.

Andersen, G. (2016). Semi-lexical features in corpus transcription. Consistency, comparability, standardisation. *International Journal of Corpus Linguistics* 21(3): 323-347.

Baron, A. & Rayson, P. (2008). VARD 2: a tool for dealing with spelling variation in historical corpora. In: *Proceedings of the Postgraduate Conference in Corpus Linguistics*, 22 May 2008, Aston University, Birmingham, UK.

Bednarek, M. (2014). ‘Who are you and why are you following us?’ Wh-questions and communicative context in television dialogue. In J. Flowerdew, ed., *Discourse in Context*. (*Contemporary Applied Linguistics* 3). London/New York: Bloomsbury. [formerly Continuum], pp. 49-70.

Bednarek, M. (2017). The role of dialogue in fiction. In M. Locher and A. H. Jucker, eds., *Pragmatics of Fiction*. Berlin/New York: de Gruyter Mouton, pp. 129-58.

Bednarek (2018). *Language and Television Series: A Linguistic Approach to TV Dialogue*. Cambridge: Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243-257.

Bonsignori, V. (2009). Transcribing film dialogue: From orthographic to prosodic transcription. In M. Freddi and M. Pavesi, eds., *Analysing Audiovisual Dialogue. Linguistic and Translational Insights*. Bologna: Clueb, pp. 185-200.

Clift, R. (2016). *Conversation Analysis*, Cambridge: Cambridge University Press.

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 13 (*Corpus Linguistics and Variation in English: Focus on Non-Native Englishes*). Retrieved from www.helsinki.fi/varieng/series/volumes/13/dose/, 15 February 2017.

Douglas, P. (2011). *Writing the TV Drama Series: How to Succeed as a Professional Writer in TV*, 3rd edn, Studio City, CA: Michael Wiese Productions.

Du Bois, J. W. (1991). Transcription design principles for spoken discourse research. *Pragmatics*, 1(1), 71-106.

Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., & Paolino, D. (1993). Outline of discourse transcription. In J. A. Edwards and M. D. Lampert, eds., *Talking Data: Transcription and coding in discourse research*. Hillsdale/London: Lawrence Erlbaum, pp. 45-87.

Dunn, A. (2005). The genres of television. In H. Fulton, R. Huisman, J. Murphet and A. Dunn, *Narrative and Media*. Cambridge: Cambridge University Press, pp. 125-39.

Hassler-Forest, D. (2014). *The Walking Dead*. Quality television, transmedia serialization and zombies. In R. Allen and T. van den Berg, eds., *Serialization in Popular Culture*. New York/London: Routledge, pp. 91-105.

McCabe, J. & Akass, K. (eds.). (2007). *Quality TV. Contemporary American Television and Beyond*, London/New York: I.B. Tauris.

Mittell, J. (2015). *Complex TV. The Poetics of Contemporary Television Storytelling*, New York/London: New York University Press.

Price, J. (2015). *'Oh Jesus Christ!' The use of bad language in contemporary American television series* (Unpublished Honours thesis). University of Sydney, Australia.

Richardson, K. (2010). *Television Dramatic Dialogue. A Sociolinguistic Study*, Oxford: Oxford University Press.

Scripted Series Report. 2010/2011 season. *Médiamétrie*. http://www.mediametrie.com/eurodatatv/.

Sinclair, J. M. (2005). Corpus and text: basic principles. In M. Wynne, ed., *Developing Linguistic Corpora. A Guide to Good Practice*. Oxford: Oxbow Books/Arts and Humanities Data Service, pp. 1-16.

Thompson, K. (2003). *Storytelling in Film and Television*, Cambridge, M.A./London: Harvard University Press.

Toolan, M. (2014). Stylistics and film. In M. Burke, ed., *The Routledge Handbook of Stylistics*. Oxon/New York: Routledge, pp. 455-70.

Notes

[1] No series that started in 2013 and beyond were collected because the external criterion of critical acclaim through awards – outlined below – could not be reliably applied, as these series were simply too new/recent (when the corpus was designed) to have any award nominations or wins.

[2] Comedy-drama hybrids in SydTV include *Enlightened*, *Nurse Jackie*, *United States of Tara*, *The Big C*, *Weeds*, *Entourage*, *Desperate Housewives*, *Grey's Anatomy*, *Glee*, *Gilmore Girls*, *Girls*, *Royal Pains*, *Suits*, and others. In an example of metafictionality, the following dialogue lines comment explicitly on this hybridity of contemporary television:

> Leonard: And we weren't even watching TV! We were watching Netflix like the kids do!
> Penny: Yeah! Is it a comedy? Is it a drama? Nobody knows!
> (*The Big Bang Theory*, season 8, episode 17)

[3] An episode was classified as a 'beginning' episode if it occurred in about the first third or 30 per cent of a season. The first episode was labelled separately as *pilot*. Further, an episode was classified as a 'middle' episode if it occurred between 41 per cent (e.g. episode 9 of 22) and 75 per cent (e.g. episode 15 of 20). Finally, an episode was classified as an 'end' episode if it occurred between 78 per cent (e.g. episode 7 of 9) and 96 per cent (e.g. episode 23 of 24). The final episode (season finale) was identified as such.

[4] An alternative approach would have been to mimic the circulation/reception of texts, which would have meant including more episodes from one popular genre (e.g. crime) or one popular series. The

result of such an approach might have been a corpus consisting mainly of episodes from *Game of Thrones*, *House of Cards*, *Breaking Bad*, or *The Big Bang Theory*. However, the corpus would then not include the variability of contemporary American TV dialogue.

[5] ADVICe: http://lknol.com/advice.php, last accessed 29/12/2015; MICASE: http://quod.lib.umich.edu/m/micase/, last accessed 29/12/2015.

[6] http://childes.psy.cmu.edu/, last accessed 29/12/2015.

[7] https://www.isip.piconepress.com/projects/switchboard/, last accessed 29/12/2015.

[8] Because the compilation of the corpus took several years, different research assistants were involved in the creation of SydTV. The first research assistant collected the fan transcripts and scripts, revised some fan transcripts, and transcribed twenty-three episodes from scratch. The second research assistant corrected two scripts, revised some fan transcripts, and transcribed twenty-five episodes from scratch. At this stage, a third research assistant used the software VARD (Baron & Rayson 2008) to identify and correct some spellings (e.g. *aright* [*alright*], *stor* [*story*], *Arii* [*Ari*], *Gosseling* [*Gosling*], *meand* [*me and*], etc). Non-standard spellings such as *wanna* (‘want to’) were retained as appropriate. I then proofread print-outs of the dialogue for all 66 episodes and where it seemed that there was an error I located the relevant scene in the audio-visual file and checked it against the dialogue, correcting it if necessary. Interestingly, much of what appeared not to make sense when reading the transcript was resolved through the audio-visual context of situation (for instance, scene changes, being on the phone, addressing another speaker, and other aspects not captured in the transcription). Through the proofreading process I also identified some transcripts with more errors than others – I hence asked a fourth research assistant to check 47 of the transcripts again by watching the whole episode, after which I reviewed her suggested changes.

[9] The L1 varieties of this audience included Australian English (3x), American English (3x), British English (1x), Chinese/Mandarin/Cantonese (3x), Adi (Tibeto-Burman) (1x), Urdu (1x), French (1x), and Japanese (1x). Participants were asked to fill in the slots in these two examples:

(i) I ___________________ for nothin’
(ii) I ___________________ anything all day, like you said

Suggestions for the slot in (i): *I ain’t here* (2x); *I ain’t there* (2x); *I ain’t in*; *I ain’t unt*; *I ain’t enough*; *I ain’t that*; *I ain’t got*; *I ain’t dead*; *I ain’ ?*; *I ain’t …*; *I dint ask*; *I’d net*; *I want out*; *I amount*.

Suggestions for the slot in (ii): *I ain’t eat* (4x); *I ain’t eaten*; *I didn’t eat*; *I’d eat*; *I’ll eat*; *I ain’t did*; *I ain’t/didn’t*; *I didn’t*; *I didn’t do*; *I didn’t think*.

[10] For example, intonation units and contours, accent, timing (e.g. tempo, pause, lengthening), nonverbal noises (e.g. laughter, in-/exhalation, throat-clearing), voice quality, gaze, gestures, mode (e.g. face-to-face/phone), relevant non-linguistic events or actions (e.g. phone ringing, thunder, food being served), etc.

[11] Consider this extract from SydTV (*How I Met Your Mother*):

> BARNEY: Oh, God, I’m so sorry. That’s just, that’s, two vodka cranberries please.
> CLAUDIA: You remembered I drink vodka cranberries.
> BARNEY: Remember? When it comes to you, how can I forget? They all drink vodka cranberries. So, is there anything else you need, sweetie?

This conversation takes place at a bar, and the lines *two vodka cranberries please* and *They all drink voda cranberries* are addressed to the bartender, rather than Claudia.

[12] The internet movie database or subtitles or other online information was used to identify the characters’ names.

[13] At times, the transcription will be subjective, for instance when the pronunciation falls somewhere between *y’all* and *you all*. Not all variants could be transcribed – for example, the following have not been identified in the transcription: *a’ight* (*alright*); *‘ere* (*here*); *ol’* (*old*), *coupla* (*couple/couple of*), *dunno* (*don’t know*), *fella* (*feller*), *nigga* (*nigger*), *s’posed* (*supposed*), *‘fraid not* (*afraid not*), *g’night* (*good night*), *dawg* (*dog*), *‘sides* (*besides*), *ya* (*you*), *nah* (*no*), ‘*bout* (*about*), *biatch* (*bitch*), *Imma/I’m*(*m*)*a* (*I’m gonna*), *wha*’ (*what*), *jus’* (*just*), *wan’t* (*wasn’t*), etc. *Well* is very

often swallowed to some degree, but simply transcribed as *well*. I would advise those interested in pronunciation variants to use the corpus as a starting point and undertake detailed transcription of relevant scenes by watching the original episode (all SydTV episodes are publicly available at low cost via distributors such as Itunes).