[This is a pre-print version of an article accepted to be published in *Corpora* issue 15.1, 2020]

The *Sydney Corpus of Television Dialogue*: Designing and building a corpus of dialogue from US TV series

Monika Bednarek, Department of Linguistics, The University of Sydney

1. Introduction

Television dialogue – dialogue in fictional TV series – is consumed by millions of viewers worldwide, whether on their TVs or on other devices. New technologies enable viewers in countries where English is spoken as a foreign language to access versions in the original (English). The language used in TV series and films can thus become an influential model for learners (Mittmann, 2006: 575). In a questionnaire undertaken with almost 600 German university students, about 70% of respondents indicated that they watched English-language TV series in English (Bednarek, in press). From an applied linguistic perspective it is crucial to analyse the types of spoken language such learners encounter. However, the analysis of TV dialogue is also important for other sub-disciplines, such as sociolinguistics, pragmatics, stylistics, and others (see e.g. Richardson, 2010; Piazza et al., 2011). Indeed, the analysis of 'telecinematic discourse' (Piazza et al, 2011) – is currently a vibrant field of research, as evidenced by a 21-page bibliography (Bednarek & Zago, 2018).

Corpus research on television dialogue has tended to utilise either small-scale corpora or corpora that consist of dialogue from only one series or genre: The first corpus studies of television dialogue focussed on 'cult' series, namely *Star Trek* (Rey, 2001), *Will & Grace* (Baker, 2005), *Friends* (Quaglio, 2009) and *Gilmore Girls* (Bednarek, 2008, 2010). The corpora used in these studies include either the complete dialogue from the analysed series or a substantial amount of it and have provided valuable insights into these specific products, usually on the basis of fan transcripts. But it is difficult to generalise from the analysis of one series to television dialogue as a language variety. Thus, in relation to his research on *Friends*, Quaglio (2009: 14) explicitly states: 'the results of this study should not be (and are not meant to be) generalized to television dialogue overall'.

To explore lexico-pragmatic characteristics of original and dubbed television dialogue Mittmann (2006) analysed data from three TV series: *Golden Girls* (1 episode), *Dawson's Creek* (6 episodes) and *Friends* (7 episodes). Csomay & Petrovic (2012) included five episodes from one TV series (*Law and Order*) in their corpus study of technical vocabulary in films and TV shows. The *Corpus of American Television Series* (Dose 2013), compiled as a pedagogic corpus, consists of dialogue from seven episodes each from four TV series (~160,000 words). The Corpus of American Soap Operas (Davies 2012) is described on the website as containing 100 million words from over 22,000 transcripts from ten US soap operas. However, the corpus is restricted to one genre (soaps), and it is somewhat difficult to identify the accuracy of the transcripts (Bednarek, 2015). The same can be said with respect to Webb & Rodger (2009), whose dataset consists of two episodes each from 19 TV series (1990s and older). Berber Sardinha & Veirano Pinto's (2017) USTV corpus contains a wide range of different types of television texts, including 28 texts from five drama series (116, 532 words) and 28 texts from eight sitcoms (107, 533 words) as well as several texts from soap operas, mini series, animation series, programs targeted at children or teens, and many other television programs including news and reality television. As with the SOAP corpus, the texts were downloaded from websites, and their accuracy – i.e. the extent to which they

faithfully represent the on-screen dialogue that viewers encounter – remains unclear. In any case, the aim of their study is to analyse television language in general, rather than just dialogue in fictional TV series.

This overview of TV dialogue corpora was not presented here with the aim of criticising these corpora (since they may be appropriate for achieving a particular study's objective), but rather to demonstrate that most corpus research has tended to utilise either corpora that consist of dialogue from only one TV series (e.g. *Friends*) or one genre (e.g. soap opera), small-scale, purpose-built corpora, or corpora that include a fairly limited amount of *different* recent TV series. I myself have used corpora that contain dialogue ranging from one TV series to seven, ten, and twenty-seven different series. These corpus studies included analysis of a whole series (Bednarek, 2010, *Gilmore Girls*) and analysis of the whole first season of a series (Bednarek, 2012b, *The Big Bang Theory*). Using fan transcripts, I was also able to compile a corpus containing 32 episodes (in total) from seven series representing different genres such as crime, mystery, medical drama, comedy and drama (Bednarek, 2012a), and a corpus consisting of five episodes each from ten series, again representing very different genres (Bednarek, 2011). Using a mix of fan transcripts and newly transcribed data, I most recently built a corpus consisting of dialogue from one episode each of twenty-seven different TV series (Bednarek, 2014).These corpora were useful for the objectives of these studies, which mainly focused on corpus analysis of lexical and grammatical features (frequency/keyness). However, it became clear that fan transcripts are more problematic as data source for other types of linguistic analysis (see section 2.3). In other words, to allow for a more comprehensive analysis of television dialogue (not limited to lexicogrammar), new corpora are needed, which can also constitute a reference corpus for studies that focus on particular series. I designed the Sydney Corpus of Television Dialogue (SydTV) in order to fill this gap. In this paper I outline the basic corpus composition, the corpus design principles, and the processes of data collection and storage.

To give a brief overview here, SydTV is a small, specialised corpus designed with the objective of being representative of fictional US TV dialogue. TV dialogue is defined as the dialogue uttered by actors on screen as they are performing characters in fictional TV series. My primary aim was to build a specialised corpus that is as representative as possible of the language used in TV drama/comedy series, given time and funding constraints. Representativeness refers to the ability to generalize from corpus findings to the respective language variety and to the extent to which a corpus contains the variability that can be found in this variety (McEnery et al, 2006: 13). The relevant language variety in this case was defined as recent US TV dialogue, namely dialogue from fictional, non-animated TV drama and comedy 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, and has been labelled the new 'golden age of television' (Bednarek, in press). To clarify, it is the *first* season of a particular series that was initially broadcast during these years. This means that new seasons of some of the TV series included in SydTV are still being produced or broadcast at the time of writing (e.g. *Veep, The Big Bang Theory, NCIS, The Middle*) while most others are being shown as repeats or are available via services such as Netflix, Amazon, Hulu, or Itunes. I focussed on non-animated series and excluded soap operas and series targeted exclusively at children or teenagers.

2. Corpus design and data collection
2.1 Data included

SydTV contains dialogue from one first-season-episode of 66 different series, as it is not meant to be representative of a particular series but rather of 'TV dialogue' as a language variety (within the specific limits defined above: US, fictional, non-animated, drama/comedy, etc) Technically, the corpus is *SydTV 3.0* and has previous incarnations, but I refer to it simply as SydTV here. *SydTV 3.0* exists in two versions: the original version (as transcribed) and a partially standardized version, as described in the corpus manual (Bednarek, 2018). The standardized version (SydTV-Std) is useful for comparisons across corpora. For example, standardizing 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 Bednarek (2018).

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

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

2.2 Selection of data

To design the corpus, I used a mix of production and reception criteria: As far as production is concerned, the corpus aims to mimic the variability of US TV narratives. 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. The corpus thus contains dialogue from 66 TV narratives, 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). This two-fold distinction simplifies matters somewhat, since many TV series are a mix of comedy and drama, or otherwise 'hybrids' (Dunn, 2005: 138). The drama category therefore includes genre combinations 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 produced as seasons with a particular number of episodes, and 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. 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). To be clear, it is important to include pilot and final episodes in the corpus since they are part of the viewing experience, but it is crucial that the corpus is not unduly dominated by such episodes. To avoid introducing another variable that might impact on findings, only dialogue from the *first* season was included.

Reception-based criteria were also taken into account, 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 be defined in many different ways. 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 recognize either writing or the TV series overall, rather than non-dialogue related aspects. 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 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)

To find additional examples of comedy series (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 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.[1] The

[1] The lists used as sampling frames (list of potential series to include in the corpus) might include children's series, reality TV series, comedy sketch-shows, animated series, mini-series, or soap operas, but such entries were disregarded.

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. 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.

The IMDb was used systematically to ascertain the two main genre labels for each TV series, the number of episodes in the season, the typical length of episodes, and the year of first broadcast. Wikipedia was consulted for further contextualization. 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.

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 (see Sinclair, 2005), and would be important for episode-based discourse and stylistic analysis. For each chosen episode, dialogue was transcribed from scratch or on the basis of existing transcripts as explained in section 2.3 below. 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.

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	

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)

Finally, the corpus contains both network television (42 episodes), and cable television (basic cable: 11 episodes, premium cable: 13 episodes). This was not used as a 'selection' criterion during the design stage, as it would have introduced too much complexity. Rather, it is a 'descriptive' criterion, not controlled during the data collection but listed in the corpus documentation (Burnard, 2002; Love et al, 2017).[2] It is 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 of differing regulations). Further, a corpus that only includes network television series will not be representative of contemporary television, since shows by HBO (premium cable) or AMC (basic cable) are important cultural products, including series such as *The Wire* and *Breaking Bad* (both included in SydTV). SydTV contains no Netflix, Amazon or Hulu originals, since these were not as widespread during corpus design as they are now, and Netflix was not available in Australia until 2015.

2.3 Transcription

The vast majority of episodes (48 episodes) were transcribed from scratch by research assistants under my direction (mainly on the basis of iTunes, DVD or streaming versions). In addition, two scripts and sixteen fan transcripts, available online on various sites, were used as starting points by the research assistants who corrected these texts by checking them against the audio-visual file of the respective episode. Because of the expensive (time-consuming) nature of detailed transcription, transcription was mainly orthographic, although marked pronunciation variants, contractions, discourse markers, hesitation markers, listening cues, dis/agreement markers, and interjections were transcribed. The transcription conventions are provided in the corpus manual (Bednarek, 2018). A number of measures were used to increase the accuracy of the transcriptions. For example, transcripts were proofread and transcribers were asked to double-check transcripts against videos. Nevertheless, human error is still a possibility and minor inconsistencies remain (Bednarek, 2018).

The most obvious limitation of the corpus is its small size, a result of the fact that most dialogue was transcribed from scratch. I did consider some alternatives in order to create a bigger corpus, namely using online scripts, subtitles or fan transcripts. However, these have several disadvantages, as discussed in Bednarek (2015). Of the three options, fan transcripts are the closest to on-screen dialogue, but uncorrected fan transcripts are not completely accurate, and unsuitable for analysis of informality, colloquiality, discourse phenomena, performance features, etc. In addition, fan transcripts as well as subtitles or scripts may not be available for all of the episodes to be included in a corpus, which makes it difficult to achieve balance in terms of the variables of textual time, 'quality' vs. 'mainstream,' and drama vs. comedy.

For all of these reasons, transcription was used for 70 per cent of the corpus, as explained above. 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).

[2] Burnard (2002: 57-58) distinguishes selection criteria, where 'the domain of values for this category was predefined, and a target population identified for each' from descriptive criteria, where 'no particular target was set for the proportion of material of a particular type' but 'other things being equal, attempts would be made to maximize variability within such categories'. I use the term *descriptive criteria* to refer to criteria that are not used to select data for inclusion in a corpus but that are used to describe aspects of variability in a corpus in relevant corpus documentation materials.

2.4 Storing the data

All SydTV files are plain text files (.txt). The corpus is predominantly 'raw' text, although speakers were identified as such. For the version I used in Bednarek (in press), speaker names were simply marked by angle brackets, i.e. <JACKIE:>. In the version available via an online interface, the tags are XML-compatible: <u who="JACKIE"> Hey. </u>. Information on access to SydTV is provided at www.syd-tv.com.

3. Conclusion

This article has introduced SydTV as a new corpus of US TV dialogue for studies that either aim to investigate TV dialogue *as a language variety* or to find out how similar/different dialogue from a *particular* series is to TV dialogue in general (using SydTV as reference corpus). SydTV could be potentially useful for pragmatic, stylistic, sociolinguistic and other research that does not require large corpora. I have explained how the corpus was designed and built, using a mix of production- and reception-based criteria. The corpus construction was guided by the aim to mimic the variability of US TV narratives, that is, to include dialogue from many different TV series and to include a balance of comedy and drama, mainstream and quality, and different moments of textual time. 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 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 American TV dialogue.

In Bednarek (in press), SydTV is assessed explicitly in terms of its affordances and limitations. Clearly, the corpus cannot be used for all types of linguistic analysis, including those that require larger amounts of data. It is my hope that the corpus can be used as a starting point to identify interesting scenes for analysis, which can then be transcribed in more detail by researchers by accessing the original video. In this way, corpus linguistics could be combined with qualitative approaches to discourse that require and use more detailed transcription methods. SydTV can also be used as a reference corpus in linguistic studies of one particular series, which is often the case in stylistics, pragmatics, and sociolinguistics. To enable other scholars to use SydTV as a reference corpus or to investigate TV dialogue more generally, I have made frequency lists publicly available and also offer free access to an online interface. Information on both of these resources is provided at the companion website to SydTV: www.syd-tv.com.

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*, pp. 232-5. New York/London: Routledge.

Baker, P. 2005. *Public Discourses of Gay Men*. London: Routledge.

Bednarek, M. 2008. '"What the hell is wrong with you?" A corpus perspective on evaluation and emotion in contemporary American pop culture' in A. Mahboob and N. Knight (eds.) *Questioning Linguistics*, pp.95-126. Newcastle: Cambridge Scholars Press.

Bednarek, M. 2010. *The Language of Fictional Television: Drama and Identity*. London/New York: Continuum.

Bednarek, M. 2011. 'The language of fictional television: a case study of the "dramedy" *Gilmore Girls*', *English Text Construction* 4(1), pp 54-83.
Bednarek, M. 2012a. '"Get us the hell out of here": Key words and trigrams in fictional television series', *International Journal of Corpus Linguistics* 17(1), pp 35-63.
Bednarek, M. 2012b. 'Construing "nerdiness": characterisation in *The Big Bang Theory*', *Multilingua* 31, pp 199-229.
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*, pp. 49-70. London/New York: Bloomsbury Academic.
Bednarek, M. 2015. 'Corpus-assisted multimodal discourse analysis of television and film narratives' in P. Baker and T. McEnery (eds.) *Corpora and Discourse Studies*, pp. 63-87. Basingstoke/New York: Palgrave Macmillan.
Bednarek, M. 2018. *Guide to the Sydney Corpus of Television Dialogue (SydTV)*. Available at www.syd-tv.com.
Bednarek, M. in press. *Language and Television Series: A Linguistic Approach to TV Dialogue*. Cambridge: Cambridge University Press.
Bednarek, M. and R. Zago. 2018. *Bibliography of linguistic research on fictional (narrative, scripted) television series and films/movies, version 2 (February 2018)*. Retrieved from http://unipv.academia.edu/RaffaeleZago.
Berber Sardinha, T. and M. Veirano Pinto. 2017. 'American television and off-screen registers: a corpus-based comparison', *Corpora* 12(1), pp 85-114.
Burnard, L. 2002. 'Where did we go wrong? A retrospective look at the British National Corpus' in B. Kettemann and G. Markus (eds.) *Teaching and Learning by Doing Corpus Analysis*, pp. 51-71. Amsterdam: Rodopi.
Csomay, E. and M. Petrović. 2012. '"Yes, your Honor!": A corpus-based study of technical vocabulary in discipline-related movies and TV shows', *System* 40(2), pp 305-15.
Davies, M. 2012. *Corpus of American Soap operas.* Available from http://corpus.byu.edu/soap/.
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.
Dunn, A. 2005. 'The genres of television' in H. Fulton, R. Huisman, J. Murphet and A. Dunn, *Narrative and Media* pp. 125-39. Cambridge: Cambridge University Press.
Love, R., Dembry, C., Hardie, A., Brezina, V. and T. McEnery. 2017. 'The Spoken BNC2014: designing and building a spoken corpus of everyday conversations', *International Journal of Corpus Linguistics* 22(3), pp 319-344.
McEnery, T., Xiao, R. and Y. Tono 2006. *Corpus-Based Language Studies. An Advanced Resource Book*. London/New York: Routledge.
Mittell, J. 2015. *Complex TV. The Poetics of Contemporary Television Storytelling*. New York/London: New York University Press.
Mittmann, B. 2006. 'With a little help from *Friends* (and others): Lexico-pragmatic characteristics of original and dubbed film dialogue' in C. Houswitschka, G. Knappe and A. Müller (eds.) *Anglistentag 2005, Bamberg – Proceedings*, pp.573-85. Trier: WVT.

Piazza, R., Bednarek, M. and F. Rossi. (eds.). 2011. *Telecinematic Discourse: Approaches to the Language of Films and Television Series*. Amsterdam/Philadelphia: John Benjamins.
Quaglio, P. 2009. *Television Dialogue. The Sitcom* Friends *vs. Natural Conversation*. Amsterdam/Philadelphia: John Benjamins.
Queen, R. 2015. *Vox Popular: The Surprising Life of Language in the Media*. Malden/Oxford: Wiley-Blackwell.
Rey, J. M. 2001. 'Changing gender roles in popular culture: Dialogue in *Star Trek* episodes from 1966 to 1993' in D. Biber and S. Conrad (eds.) *Variation in English*: *Multi-dimensional Studies*, pp. 138-56. London: Longman.
Richardson, K. 2010. *Television Dramatic Dialogue. A Sociolinguistic Study*. Oxford: Oxford University Press.
Sinclair, J. M. 2005. 'Corpus and text: basic principles' in M. Wynne (ed.) *Developing Linguistic Corpora. A Guide to Good Practice*, pp. 1-16. Oxford: Oxbow Books/Arts and Humanities Data Service.
Thompson, K. 2003. *Storytelling in Film and Television*. Cambridge, M.A./London: Harvard University Press.
Toolan, M. 2014. 'Stylistics and film' in M. Burke (eds.) *The Routledge Handbook of Stylistics*, pp. 455-70. Oxon/New York: Routledge.
Webb, S. and M.P.H. Rodgers. 2009. 'Vocabulary demands of television programs', *Language Learning* 59(2), pp 335-66.