Literator - Journal of Literary Criticism, Comparative Linguistics and Literary Studies
ISSN: (Online) 2219-8237, (Print) 0258-2279

Stylometry and characterisation in *The Big Bang Theory*

Authors:
Maryka van Zyl[1]
Yolande Botha[1]

Affiliations:
[1]School of Languages, North-West University, Potchefstroom Campus, South Africa

Corresponding author:
Yolande Botha,
lande.botha@nwu.ac.za

Dates:
Received: 02 Feb. 2016
Accepted: 06 July 2016
Published: 22 Nov. 2016

How to cite this article:
Van Zyl, M. & Botha, Y., 2016, 'Stylometry and characterisation in *The Big Bang Theory*', *Literator* 37(2), a1282. http://dx.doi.org/10.4102/lit.v37i2.1282

Dialogue is an important aspect of televisual character construction. Writers make linguistic choices on behalf of characters, and these choices can cause viewers to associate a character with a specific stereotype, subculture or social group. This study examines the linguistic construction of the character Sheldon Cooper in the CBS sitcom *The Big Bang Theory*. A cluster analysis tree of the speech of each of the five main characters in the first seven seasons (generated by the R script Stylo 0.6.0) indicated that the character of Sheldon Cooper differs from the other main characters (Leonard, Penny, Howard and Rajesh) with respect to linguistic style. These differences were further explored using corpus analysis software (WordSmith 6.0) to identify keywords and lexical bundles and to compare the use of active versus passive voice constructions. Sheldon's choice of scientific or more formal words and his relative preference for the passive voice typify his linguistic style as expository rather than colloquial.

Stilometrie en karakterisering in *The Big Bang Theory*. Dialoog is 'n belangrike aspek van televisuele karakterkonstruering. Skrywers maak talige keuses namens die karakters en hierdie keuses kan daartoe aanleiding gee dat kykers 'n karakter met 'n spesifieke stereotipe subkultuur of sosiale groep vereenselwig. Hierdie studie ondersoek die talige konstruering van die karakter Sheldon Cooper in die CBS-sitkom *The Big Bang Theory*. 'n Trosanaliseboom van die spraak van elk van die vyf hoofkarakters in die eerste sewe seisoene (gegenereer deur die *R*-skrip *Stylo* 0.6.0) dui daarop dat die karakter Sheldon Cooper van die ander hoofkarakters (Leonard, Penny, Howard en Rajesh) verskil ten opsigte van taalstyl. Hierdie verskille word verder ondersoek deur gebruik te maak van die korpusanalise program (*WordSmith* 6.0.) om sleutelwoorde en leksikale bondels te identifiseer en om die gebruik van aktiewe en passiewe werkwoordkonstruksies te vergelyk. Sheldon se keuse van wetenskaplike of meer formele woorde en sy relatiewe voorkeur vir die passiefkonstruksie tipeer sy taalstyl as verduidelikend eerder as tipies van omgangstaal.

Introduction

Co-authored by over 20 writers, the popular American comedy series *The Big Bang Theory* (Lorre & Prady 2007–present) has delivered interesting and, more importantly, rather stable character constructs, which is specifically reflected in the linguistic constructions of the main characters, especially that of Sheldon Cooper. Though the series has attracted the attention of research relating to semiotics and discourse analysis (Balirano 2013; Bednarek 2012; Ma & Jiang 2013; Shuqin 2012, 2013; Yin & Yun 2012), the construction of a character's particular idiolect also makes an important contribution to characterisation. The way in which a character speaks (the linguistic choices that the writers make on behalf of the characters, as it were) can serve to associate the character with a specific stereotype, subculture or social group, but also personal affect (Culpeper 2001:190). Social and cognitive theories within psychology and language may therefore serve as a general backdrop to the character of Sheldon, and the emergent linguistic features may readily be associated with his world view.

The series introduces four male characters (Sheldon, Leonard, Rajesh and Howard), all scientists who share a passionate interest in comic books and science fiction. Despite the fact that they largely share a social and professional environment and all belong to the scientist/nerd stereotype to varying degrees, it is evident to most viewers that Sheldon Cooper greatly differs socially from his friends. In addition, Penny is introduced as Sheldon and Leonard's neighbour. She portrays the pretty blonde stereotype and stands in contrast to the male characters on a social and intellectual level.

Sheldon is the most peculiar of the five main characters in terms of his overall social behaviour but is also the most intelligent in terms of academic qualities and achievements. Besides his

social stereotype as 'geek', he is portrayed as condescending, pedantic, egotistic and self-righteous. He often experiences difficulty with interpersonal communication (Yin & Yun 2012:1222) and is unable to disambiguate between the literal and the figurative. These traits may fit better into a cognitive or personal affect construct rather than a stereotype. In relation to his co-characters, Sheldon is rather unique and this is displayed not only in his onscreen behaviour, but also in his linguistic repertoire. An analysis of Sheldon's speech repertoire is a way of providing linguistic evidence for these intuitive observations regarding his character.

This study aims to establish whether and how Sheldon differs linguistically from his co-characters. It explores the linguistic choices of Sheldon Cooper that set him apart from the other main characters, focusing on lexis and voice. This analysis serves as an example of how idiolect can be constructed and maintained over multiple seasons of a television series with multiple writers. A stylometric analysis of the speech of each of the five main characters (Sheldon, Leonard, Howard, Rajesh and Penny) was conducted using the R script Stylo (0.6.0) (Eder & Rybicki 2011) to first establish whether there are significant quantifiable differences in the speech of the main characters. As expected, the cluster tree analyses show that Sheldon Cooper has a discernible linguistic style that differs from that of the other characters. These results are discussed in the first part of the analysis section. Having stylometrically established that Sheldon's language usage differs from that of the other characters in the series, we can ask which specific aspects of Sheldon's language constitute his idiolect. Sheldon's lexical choices as well as his tendency to use the passive voice are discussed as features that distinguish his linguistic style.

Idiolect: Language and characterisation

Balossi (2014:24) argues that language underlies characterisation in that an individual's feelings, behaviour and cognitive states are represented by and through language. Language usage reveals attitudes and beliefs and therefore serves as an important source of information into personalities or character types (Balossi 2014:24). The language of any particular character is therefore an important element of his or her general construct. Although various aspects contribute to the makeup of literary characters, such as self-presentation and mental and social representations, there is an inextricable link between language use and character portrayal (Culpeper 2001).

The association between characterisation and language usage can be quantitatively investigated. Stylometry studies take grammatical function words into account (Burrows 1987; Culpeper 2001; Kestemont 2014), as these typically are more frequent than lexical items. Function words are good indicators of authorship styles, since these closed-class words are used by all writers, are produced unconsciously and are mostly context-independent (Kestemont 2014:60). Culpeper (2001), for example, performed a keywords analysis on Shakespeare's *Romeo and Juliet*, calculating the unusual word frequencies, which may reveal salient traits of characters. Capulet's *go* often relates to an issuing of command, given his role in the household. In Juliet's case, *if* was the most significant keyword and was associated with her anxiety regarding Romeo's safety and matters related to him (Culpeper 2001:188–189).

Computer-mediated research has aided literary studies such as Balossi's (2014) stylometric analysis of Virginia Woolf's *The Waves*, which studied the differentiation of characters through various word classes – lexical as well as grammatical. Burrows (1987) examined function words in Jane Austen's novels, as well as sentence length and the use of pronouns by the various characters. Balossi (2014) used stylometric analyses to examine lexical density in language associated with gender, ideology and social status and found that 'both content as well as function words uncovered aspects of the characters' personality traits that would probably not have been predicted without this empirical study' (2014:56).

The notion of idiolect (Coulthard 2004:431) informs this study of characterisation through language. According to Coulthard, this idiolect will 'manifest itself through distinctive and idiosyncratic choices in texts' (2004:432). Coulthard (2004:432) describes these as 'linguistic impressions' created by a speaker or writer. Edwards (2009:19) maintains that a speaker or writer sets up his or her own language identity, and this identity requires 'the sameness of the individual at all times or in all circumstances'. Differences in the language identity of characters relate to a choice of specific patterns or a deviation from expectations and may involve syntactic variation and lexis (Culpeper 2001; Renkema 2004:146). Writers of a particular series or genre consciously exercise a series of these choices that contribute to character constructions (Olsson 2008:29).

While stylistics traditionally form the subject matter of literary studies (Balossi 2014; Culpeper 2001), the character as an individual (not only the literary texts as a semiotic work) can be viewed in terms of stylistic choices (Renkema 2004:147). Culpeper (2001), in his multidisciplinary approach to the process of characterisation, emphasises how different approaches, such as social and cognitive theory within linguistics, contribute to character formation and character identity and notes that audiences may base their inferences especially on textual cues. There are also studies that have explored how power relationships are portrayed through linguistic choice (Douglas & Sutton 2010; Palmeira 2015; Wakslak, Smith & Han 2014). The language of television series has also been the focus of a number of recent linguistic studies. Bednarek (2015) investigated gender in the television programmes *Weeds*, *Nurse Jackie* and *Saving Grace* and showed that female characters who used a higher degree of bad language were more readily accepted by audiences, as opposed to traditional feminine roles. Bednarek (2011:1–24) also conducted a stylistic analysis of the (in)stability of characters' linguistic variation in *Gilmore Girls* and found that

the character *Lorelai* is relatively stable across the series. Quaglio (2008:189), motivated by concerns relating to English as a second language, compares the language of the comedy *Friends* with natural conversation and shows that this series does have linguistic characteristics in common with face-to-face conversation. Rey (2001:138–156) analysed gender roles as manifested in the dialogue of *Star Trek* episodes and shows how linguistic changes reflect significant shifts in the linguistic characterisation of women and men in this series.

In a characterisation study of *The Big Bang Theory*, Bednarek (2012:199–229) investigated the character construction of Sheldon Cooper in terms of the nerd/geek social stereotype and illustrated that the male main characters are construed as stylised representations of particular social identities with traits such as intelligence, unattractiveness, an interest in fantasy-based activities, social awkwardness and physical awkwardness or no interest in sport-related activities (Bednarek 2012:203). However, Sheldon's linguistic distinctiveness cannot be ascribed to his geekiness, since Howard, Rajesh and Leonard (who are also geeks) have more in common linguistically with Penny than with Sheldon. Although Sheldon uses nerd/geek jargon, the stylometric differences between his language and that of the other geeks suggest that Sheldon's speech cannot simply be classified as nerd-speak. Sheldon's speech is not only influenced by group identity, but also by his personality traits. Culpeper (2001:190) refers to this association between language and personality as 'personal affect'.

Data and methods

The data comprise all the speech of each of the five main characters in the first seven seasons of *The Big Bang Theory* (Table 1). These were obtained by extracting the subtitles from the digital video discs into a spreadsheet (*Microsoft Excel*), in which speakers were allocated to each speech turn. Each episode was watched to confirm that speech turns were demarcated and allocated correctly and that each utterance was accurately rendered.

The spreadsheet data were sorted according to speaker so that each of the main characters' speech could be saved separately and per season in a plain text file, for example Sheldon_1, which contained all of Sheldon's speech turns across all the episodes in Season 1. Thus a 'corpus' of 35 text files was created to serve as input for the initial stylometric analysis, using each text file (all the speech turns of a character over all the episodes in a given season) as a single sample, and used to generate a cluster analysis tree in Stylo (0.6.0) (Eder & Rybicki 2011; Kestemont 2014). For the keywords analysis in WordSmith (6.0) (Scott 2012), the seven texts containing the speech turns of Sheldon were combined to form the Sheldon corpus, while the 28 texts representing the speech turns of the other four characters were combined to form a control corpus. Once the keywords analysis was completed, the speech turns of each character spanning the first seven seasons were combined to yield five 'subcorpora', one per character, for concordance (keyword in context) analysis. Part-of-speech tagged versions of these were used to investigate the use of passive and active voice.

TABLE 1: Breakdown of 'corpus'.

Character	Tokens of speech per character across Seasons 1–7 for concordance analyses	Tokens in Sheldon corpus versus control corpus for keywords analysis
Sheldon	123 796	123 796
Leonard	68 952	202 234
Penny	51 325	
Howard	44 968	
Rajesh	36 989	

The 35 texts (or combinations of them) do not constitute a corpus in the strict sense of the word, since they are not running dialogue and are not 'natural' in the true sense of the word. They represent television dialogue for fictional characters based on what professional writers believe such characters might sound like. Yet it is argued here that the linguistic choices made on behalf of fictional characters are still informed by the linguistic knowledge of real people (the writers). With a complete 'suspension of disbelief', one can also view the characters' language as real in the (fictional) world of the text. In this sense, corpus linguistic methods typically associated with usage-based descriptions of patterns in language are suitable to investigate the linguistic patterns in *The Big Bang Theory* data.

Sheldon has more speech turns than any other individual character. The speech of the characters of Bernadette and Amy (later Howard and Sheldon's girlfriends, respectively) were not included in the compilation of the corpus, because these characters were not present from the start of the series. For example, Bernadette is featured a few times in Season 3, only to disappear until the middle of Season 4. Amy is only introduced at the end of the Season 3. A 'corpus' including all the characters and spanning the whole series could potentially be used to investigate the role of gender in language. The focus of this study, however, is on the linguistic style of Sheldon, as central character.

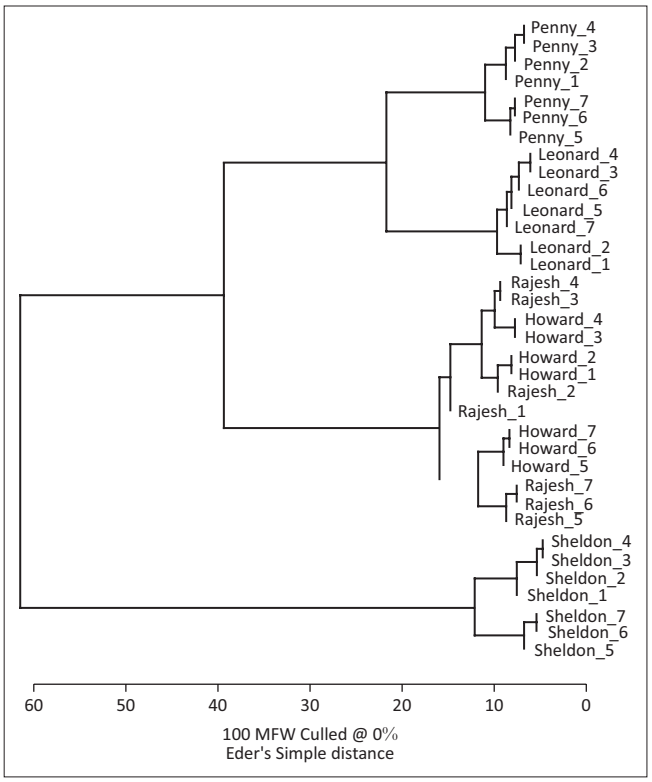
Stylo (0.6.0) was used to establish whether Sheldon has a discernible idiolect maintained across the first seven seasons. This R script was recently developed for the purposes of authorship attribution, genre recognition and style development by Eder, Rybicki and Kestemont (2016).

The cluster analysis in Figure 1 is based on the 100 most frequent words (both grammatical and lexical) and uses Eder's simple algorithm given below (Eder & Rybicki 2011).

$$\delta(AB) = \sum_{i=1}^{n} \sqrt{\mathrm{A}_i} - \sqrt{\mathrm{B}_i} \qquad \text{[Eqn 1]}$$

This algorithm calculates word frequencies independent of corpus sizes. This was important due to the large discrepancy in the quantity of Sheldon's output compared to that of other individual characters.

The keywords function in WordSmith 6.0 (Scott 2012) was used to answer the question: how is Sheldon linguistically



MFW, most frequent words.

FIGURE 1: Cluster analysis for *The Big Bang Theory*, Seasons 1–7.

different? The log-likelihood test was selected as measure of keyness with the *p*-value set at < 0.000001. Keywords are words that occur (statistically) significantly more (or less) frequently in a corpus than would be expected in terms of a larger control corpus (Stubbs 2010:25). According to Bondi (2010:1), keywords (both positive and negative keywords) are those lexical items that play a role in identifying important elements of a text. Scott (2010:43) maintains that keyness is a property possessed by words, word clusters and phrasal units. Because keyness is a purely quantificational property, both lexical and grammatical words can be keywords. Lexical keywords are often indicative of contextual differences between texts and include words that relate to a specific topic or a specific setting. When grammatical words are keywords they are indicative of stylistic differences [see in this regard Kestemont's (2014) work on function words in authorship attribution].

Analysis and findings

The results of the stylometric analysis are presented first ('Linguistic style'), followed by a discussion of Sheldon's lexical choices ('Keywords and clusters') and preference for the passive voice ('Passive voice').

Linguistic style

The linguistic outputs of each character (per season) were compared using Stylo (0.6.0) to determine whether there were stylistic differences and similarities in their language. The following output was generated.

The tree cluster (Figure 1) shows that the speech of each of the main characters is relatively consistent throughout the various seasons, with the outputs of each of the characters generally grouping closely together. The cluster should be read in terms of branching and not vertically. The longer branches indicate greater distance between characters. The first branching shows an overall two-way distinction – Sheldon on the one side and the other four main characters (henceforth referred to as 'the Rest') on the other side. The Rest branch further splits into two branches – the Penny–Leonard branch and the Howard–Rajesh branch. For each character's output, there is bifurcation between Seasons 1–4 and Seasons 5–7.

Sheldon very distinctly branches off from the other characters. It could be expected that the male characters who share Sheldon's work and hobby context would belong to the same branch. However, these characters are grouped with Penny. The cluster tree shows not only that Sheldon is linguistically different, but that there are linguistic similarities among the other characters. The social dynamics between the characters are clearly reflected. Leonard and Penny are romantically involved and Howard and Rajesh are best friends. Although both the Penny–Leonard and Rajesh–Howard clusters are roughly equally distanced from Sheldon, Penny and Leonard are linguistically distinguishable from each other – as can be seen from Penny's and Leonard's respective output clusters on the sub-branch.

A possible explanation for the mixed Rajesh–Howard may be due to the paucity of Rajesh's speech turns overall. He says very little in Season 1, but his speech turns increase as the series progresses.

This stylometric analysis confirms that there is internal consistency in the speech of each of the characters across the first seven seasons of the series and that there is a discernible linguistic style associated with three of the five main characters. More importantly, it confirms that Sheldon's word choice is significantly different from those of the other characters.

Keywords and clusters

The Keywords function in WordSmith (6.0) was used to statistically identify positive and negative keywords for Sheldon compared to the Rest. A list of 135 keywords (both lexical and grammatical) was yielded. The first 54 words were positive keywords (significantly more frequent than the reference corpus), whereas the remaining 81 were negative (significantly less frequent). The keywords list can be viewed as an Appendix A. In the discussion of selected keywords below the (log-likelihood-based) keyness value yielded by WordSmith is given in brackets.

Keywords are often context-dependent, as is the case with the proper nouns *Cooper* (39.31), *Leonard* (107.71) and *Penny* (59.58), which are indicators of the people in the immediate

surroundings of Sheldon. The keywords *Cooper* and *my* (27.47) indicate significant self-reference (although it is interesting that the first-person pronouns *I* and *me* are not listed as keywords).

As far as nouns go, lexical items such as *string theory* (29.40), *universe* (30.30), *Nobel* (29.27) (as in *Nobel prize*) and *flags* (33.44) are strong indicators of Sheldon's interests. Sheldon's key nouns are very specific and sometimes more formal, for example, *beverage* (28.95), *coitus* (25.83) (as opposed to *sex* [-48.41], which is a negative keyword), *moment* (24.15), *interest* (33.88) and *result* (30.95). The non-specific, colloquial noun *stuff* (-58.89) is a negative keyword for Sheldon, illustrating Sheldon's preference for clear and specific terms over vague references. A concordance of the word *result* (30.95) shows how Sheldon imposes scientific observation on mundane circumstances, for example:

1. [After cleaning Penny's apartment while she was sleeping]
   Sheldon: Granted, my methods may have been somewhat unorthodox, but I think the end *result* will be a measurable enhancement of Penny's quality of life. (Season1: Episode2)
2. Leonard: Sheldon, this date is probably my one chance with Penny. What happens if I blow it?
   Sheldon: Well, if we accept your premise, and also accept the highly improbable assumption that Penny is the only woman in the world for you, then we can logically conclude that the *result* of blowing it would be that you end up a lonely, bitter old man with no progeny. The image of any number of evil lighthouse keepers from Scooby-Doo cartoons comes to mind. (Season1: Episode2)
3. Rajesh: We could eat after the movie.
   Sheldon: Unacceptable, the delay would *result* in tomorrow morning's bowel movement occurring at work. (Season2: Episode14)
4. Sheldon: This seat is ideally located both in relation to the heat source in the winter and a cross breeze in the summer. It also faces the television at a direct angle, allowing me to immerse myself in entertainment or gameplay without being subjected to conversation. As a *result*, I've placed it in a state of eternal dibs.(Season3: Episode22)

The examples above illustrate that Sheldon does not have separate registers for different conversational topics. He talks to his friends about everyday activities in a manner that closely agrees with academic discourse. It would also seem that the writers of the series carefully employ these register choices in the construction of Sheldon's character.

Sheldon tends to choose more specific, less typical evaluative adjectives over adjectives that are frequent in everyday speech. The adjectives *remarkable* (35.5) and *appropriate* (27.31) are positive keywords, while *nice* (-31.35), *cool* (-95.53) and *great* (-77.53) are negative keywords for Sheldon. Given Sheldon's frequent references to norms of human behaviour, the appearance of the adjective *social* (32.68) among the positive keywords is expected. The collocates of this adjective in Sheldon's speech (namely *protocol*, *convention*, *pressure*, *obligation* and *contract*), however, are unexpected in conversation and are typically associated with academic discourse. Sheldon's choice of nouns and adjectives is indicative of the detached, scientific manner in which he views his surroundings and the people in his life. They also portray his observance of structure, rules and norms, which is in line with his attempts to avoid vagueness and ambiguity.

The verb *begin* (40.80) is typically used by Sheldon in a context of condescending explanation or logical argumentation – often about everyday matters unrelated to his profession, as the following examples illustrate:

5. [Sheldon does not like Penny's idea for an app]
   Penny: Why not?
   Sheldon: Oh, Penny, where do I *begin*? The simple-mindedness of your idea is exceeded only by its crass consumerism and banality (Season4: Episode12)
6. [Giving advice to Penny about a business idea]
   Sheldon: Good. Let's *begin* with the premise that everything you've done up to this point is wrong. (Season2: Episode18)
7. Stuart: What's wrong with Christmas?
   Sheldon: Oh, where to *begin*? Trees indoors. Overuse of the words 'tis and 'twas. And the absurd custom of one stocking. Everyone knows socks belong in pairs. Who uses one sock? (Season6: Episode111)

In many respects, Sheldon's speech lacks the linguistic features that are characteristic of conversation and Biber, Conrad and Leech (2002:429) associate certain social and situational circumstances with the grammatical features of conversation. Some of the characteristics of the language of conversation described by them include expression of stance, avoidance of elaboration, elicitation of responses, the use of attention-signalling forms to manage interactions and discourse markers such as interjections or inserts (Biber *et al*. 2002:431–433). When comparing Sheldon's language with that of the other characters, words that potentially express stance appear as both positive and negative keywords.

Among the positive keywords for Sheldon, there are a number of words that potentially express epistemic stance or his attitude or position regarding the information in a proposition (see Biber *et al*. 2002:382 on stance expressions), for instance the verbs *suppose* (24.44) and *suggest* (26.11) and the modal auxiliaries *may* (43.25) and *will* (41.30). The epistemic adverb *perhaps* (69.43) is a positive keyword for Sheldon, whereas its synonym *maybe* (-93.47) is a negative keyword. Both *maybe* and *perhaps* can be used to express possibility or uncertainty. Using the *Corpus of Contemporary American English*, Lindquist (2009:60–61) found that *maybe* occurs about twice as many times as *perhaps* in spoken language (198 per million vs. 398 per million), whereas *perhaps* is used nearly 10 times more frequently than *maybe* in academic writing (262 per million vs. 28 per million). Lindquist (2009:61) also observed a decline in the use of *perhaps* with a concomitant incline in the use of *maybe* over the last three decades in the *Time* magazine corpus.

Rather than general language change, this is attributed to the inclusion of more fiction and reported speech, or to the language of the magazine generally becoming more informal and more like spoken language, in line with the phenomenon called 'the colloquialisation of the language' (Lindquist 2009:62). Sheldon's preference for *perhaps* over *maybe* contributes to the bookishness of his linguistic style.

Sheldon often uses the word *which* (77.87) to introduce a relative clause. In the Longman Spoken and Written English (LSWE) corpus, relative clauses introduced by *which* occur nearly 10 times more frequently in academic writing than in conversation (Biber *et al*. 2002:285–286):

8. I've taken the liberty of drafting these workflow charts *which* outline our various duties. (Season4: Episode12)
9. Perhaps I'll spend some time developing a unified theory of comedy, *which* will allow me to elicit laughter from anyone at any time. (Season7: Episode12)

Sheldon's use of relative clauses lends an expository air to his speech, which is unexpected in conversation. In conversation, syntactic elaboration through noun modification is rare compared to expository written registers (Biber *et al*. 2002: 430–431). This elaborative tendency in Sheldon's speech is also reflected by the fact that *of* (183.38), *as* (138.14) and *by* (80.40) are very strong positive keywords for Sheldon. The positive keyness of the conjunction *although* (25.40) also illustrates Sheldon's expository style.

Whereas lexical words such as *theory* can be expected as positive keywords for Sheldon, it is noteworthy that frequently occurring grammatical function words are also positive keywords for Sheldon. Grammatical keywords are indicative of style rather than context. The words *the* and *of* are usually the most frequent words in any corpus of English, but they seem to be used with significantly more frequency by Sheldon compared to the other characters. The keyness of *the* (182.44) indicates more definite references in Sheldon's speech.

Keywords can be studied through their typical co-occurrence with other lexicosemantic units (Bondi 2010:3). In other words, a word as a single unit need not be the sole focus but rather the clusters in which it often occurs (Stubbs 1996:35). Biber *et al*. (2002:435) refer to the seemingly prefabricated sequences of words that are frequent in conversation, such as *I don't know*, as lexical bundles. These lexical bundles do not usually correspond to a complete grammatical unit (Biber *et al*. 2002:445). It is worth investigating keywords with a grammatical function, rather than lexical content, in terms of structural patterns and not as single units, since function words delineate structural relationships with other words or phrases. The concordance function in WordSmith (6.0) is used to identify frequent word clusters that contain the keywords *of* and *as*. The preposition *of* is the most frequently used preposition in English, whereas the word *as* can be used as a preposition or subordinator or as part of more complex relational expressions, viz *such as* and *as well as* (Biber *et al*. 2002).

Comparing the different function word classes in use, Biber *et al*. (2002:32) indicate the striking difference in the LSWE corpus between academic and conversation registers with regard to prepositions, coordinators and determiners. Prepositions occur approximately 60 000 times per million words in conversation, compared to roughly 150 000 times per million words in academic writing, and determiners occur 100 000 times per million words in academic writing as opposed to 40 000 times in conversation (Biber *et al*. 2002:32).

The keyness *of* the word of (183.38) can partly be explained by the very frequent use of the phrase *of course*, which Sheldon uses in explanations and arguments, and the use of *of* in partitions that denote 'kind' or 'quantity', for example: *some sort of* (13×), *a series of* (11×). The most frequent cluster in a concordance of the word *of* in Sheldon's speech is *one of the X* (39×), where X is often a noun modifier or superlative, for example: *best* Thanksgivings, *classic* American routes, *fantasy* worlds, *few* mammals, *few* forms, *few* benefits, *good* one, *great* minds (3×), *great* joys, *great* American trains, *great* challenges, *greatest* intellects, *most* beguiling, *most* intelligent people, *most* effective techniques, *other great* minds. These examples illustrate Sheldon's use of premodifiers as elaborative device, relying heavily on superlative and comparative forms. This illustrates something of a tendency to avoid vagueness and be explicit and clear on general matters, but also illustrates the way in which Sheldon habitually evaluates and compares entities.

Looking at three-word clusters in the concordances of *of* and *as* that occur more than five times, we found clusters used frequently by all of the characters such as *a lot of* and *as long as*. However, there are also a number of clusters that are only frequent for Sheldon, namely: *understanding of the* (9×), *one of my* (8×), *an example of* (8×), *great minds of* (7×), *might as well* (11×), *as you know* (10×).

It is also interesting to note the number of discourse markers among the negative keywords for Sheldon and more frequent for his co-characters: *wow* (-58.04), *sorry* (-31.32), *thanks* (-46.68), *fine* (-27.82), *okay* (-542.58), *yeah* (-342.27), *oh* (-48.64), *hi* (-65.45), *uh* (-88.93), *damn* (-28.30), *hell* (-30.32). This does not imply that Sheldon never uses interjections, but he chooses unusual ones like *bazinga* (46.42). The word *bazinga* is almost always used in situations where Sheldon makes his co-characters believe that he is serious when he is actually attempting to make a joke, then expresses his delight at catching them off-guard.

Sheldon's choice of nouns, verbs and adjectives with very specific meanings over vague, more general words; the way in which he expresses epistemic stance; his use of noun modifiers; his relatively frequent use of subordinators and prepositions and the scarcity of discourse markers all indicate that Sheldon's linguistic style is more closely aligned with that of academic discourse rather than conversation.

TABLE 2: Passive voice results.

Character	Past participle form of lexical content verbs	Passive usages	All lexical content verbs	Lexical content verbs not used to expressive passive voice
Sheldon	1589	795	15 843	15 048
Leonard	586	221	9862	9641
Penny	395	118	7463	7345
Howard	420	159	6460	6301
Rajesh	332	129	4909	4780

The features discussed above contribute to constructing a linguistic style for Sheldon that is more expository than colloquial.

Passive voice

The previous section dealt with the ways in which Sheldon's lexical choices contribute to make his utterances generally sound quasi-academic rather than colloquial. One of the most obvious areas to explore whether the observations that apply to lexical choice can also apply at the level of grammatical choice is voice. The passive voice is characteristic of academic discourse and is comparatively rare in conversation. Using a part-of-speech tagged version of each of the character's output, a concordance was drawn of all lexical verb forms that potentially occur in the passive, that is verbs tagged as VVN (past participle) by CLAWS (Garside & Smith 1997). Because the CLAWS tagger tags all instances of the lemmas *be*, *have* and *do* separately, these were excluded from the analysis. Only verbs that cannot also function as auxiliaries were considered. The VVN concordances were subsequently checked line by line to isolate instances of the passive voice (as opposed to perfective or adjectival uses of the participle). In order to quantitatively compare the passive and active voice across the speech of the five characters, the passive verb phrases needed to be regarded as a proportion of all verb phrases used by each character. The latter were determined with the search string VV*, which finds all lexical verbs (excluding the lemmas *be*, *have* and *do*) regardless of morphological form. Table 2 below provides the raw frequencies yielded by concordance searches of the tagged data.

Another R script, Coll.analysis 3.5, developed by Gries (2007) to perform collostructional analysis, was used to compare the (active vs. passive) voice preferences of each character based on a one-tailed Fisher–Yates exact test. The results are summarised in Table 3 below.

Collostructional strengths above 3 indicate a *p*-value of at least 0.001, while a collocational strength above 2 indicates a *p*-value of at least 0.01. These results show not only that Sheldon uses the passive voice more frequently than the other characters, but that the verb phrases in his speech are passive more frequently than they are active. This is a clear indication that the stylistic differences of Sheldon's language surpasses lexical choice and is also evidenced at the level of grammatical style. Sheldon's preference for the passive voice is in line with the other linguistic choices that typify his discourse style as quasi-academic and expository rather than colloquial.

TABLE 3: Voice preference based on collostructional strength.

Lexical content verbs of	Preferred construction	Collocational strength
Sheldon	Passive	56.1
Leonard	Non-passive	9.83
Howard	Non-passive	3.95
Rajesh	Non-passive	2.08
Penny	Non-passive	20.47

Conclusion

This study started by asking whether viewers' intuition that the character of Sheldon Cooper's language usage is different from that of the other characters in the series *The Big Bang Theory* could be quantitatively supported. In the stylometric analysis it was established that Sheldon is indeed linguistically distinguishable from the other characters, even from his fellow nerds. The texts of the other characters' speech also clustered quite neatly in the cluster tree analysis, demonstrating that the writers of the series manage to maintain a clear voice for each of the main characters throughout the series. The results of the stylometric analysis confirmed that Sheldon's linguistic style is worth further investigation.

An analysis of Sheldon's positive and negative keywords (compared to the speech of the other four main characters) showed that Sheldon's speech has much in common with academic or expository discourse and does not display as many of the typical features of conversation. The observations about Sheldon's lexical choices were corroborated in a grammatical analysis of passive versus active verb constructions. What sets Sheldon's language use apart from that of the other characters is that Sheldon's speech is not colloquial. This study illustrates how linguistic style is employed as a characterisation tool.

Given the closeness of the Leonard and Penny texts in the cluster analysis, the distances between the language of Sheldon and Amy, and of Howard and Bernadette, need further investigation. Sheldon's preference for the passive voice is evidence that his language is also grammatically distinct from that of the other characters and suggests that further investigation into Sheldon's grammatical choices could lead to additional insights into linguistic choice as a characterisation tool.

Acknowledgements

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Authors' contributions

M.v.Z. was responsible for the analyses. Both authors, M.v.Z. and Y.B. contributed equally in the interpretation of the data and the description of the findings presented in this article.

References

Balirano, G., 2013, 'The strange case of *The Big Bang Theory* and its extra-ordinary Italian audiovisual translation: A multimodal corpus-based analysis', *Perspectives* 21(4), 563–576. http://dx.doi.org/10.1080/0907676X.2013.831922

Balossi, G., 2014, *A corpus linguistic approach to literary language and characterization: Virginia Woolf's The Waves*, John Benjamins, Amsterdam.

Bednarek, M., 2011, 'The stability of the televisual character: A corpus stylistic case study', in R. Piazza, M. Bednarek & F. Rossi (eds.), *Telecinematic discourse: Approaches to the language of films and television series*, pp. 1–24, John Benjamins, Amsterdam.

Bednarek, M., 2012, 'Constructing nerdiness: Characterisation in *The Big Bang Theory*', *Multilingua* 31, 199–229. http://dx.doi.org/10.1515/multi-2012-0010

Bednarek, M., 2015, '"Wicked" women in contemporary pop culture: "Bad" language in Weeds, Nurse Jackie, and Saving Grace', *Text and Talk* 35(4), 431–451. http://dx.doi.org/10.1515/text-2015-0011

Biber, D., Conrad, S. & Leech, G., 2002, *Student grammar of spoken and written English*, Pearson Education, Essex.

Bondi, M., 2010, 'Perspectives on keywords and keyness: An introduction', in M. Bondi & M. Scott (eds.), *Keyness in texts*, pp. 1–18, John Benjamins, Amsterdam.

Burrows, J.F., 1987, *Computation into criticism: A study of Jane Austen's novels and an experiment in method*, Clarendon, Oxford.

Coulthard, M., 2004, 'Authorship identification, idiolect, and linguistic uniqueness', *Applied Linguistics* 25(4), 431–477. http://dx.doi.org/10.1093/applin/25.4.431

Culpeper, J., 2001, *Language and characterisation: People in plays and other texts*, Routledge, New York.

Douglas, K.M. & Sutton, R.M., 2010, 'By their words ye shall know them: Language abstraction and the likeability of describers', *European Journal of Social Psychology* 40(2), 366–374.

Eder, M. & Rybicki, J., 2011, *Stylo version 0.6.0*, computer software, University of Kraków, Poland.

Eder, M., Rybicki, J. & Kestemont, M., 2016, 'Stylo: A package for computational text analysis', *R Journal* 16(1), 1–15, viewed 15 April 2016, from https://journal.r-project.org/archive/accepted/eder-rybicki-kestemont.pdf

Edwards, J., 2009, *Language and identity*, Cambridge University, New York.

Garside, R. & Smith, N., 1997, 'A hybrid grammatical tagger: CLAWS4', in R. Garside, G. Leech & A. McEnery (eds.), *Corpus annotation: Linguistic information from computer text corpora*, pp. 102–121, Longman, London.

Gries, S., 2007, Coll.analysis 3.2a. A script for R to perform collostructional analysis.

Kestemont, M., 2014, 'Function words in authorship attribution: From black magic to theory?', in *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*, 27 April 2014,Sweden.

Lindquist, H., 2009, *Corpus linguistics and the description of English*, Edinburgh University Press, Edinburgh.

Lorre, C. & Prady, B., 2007–2013, *The Big Bang Theory*, DVD, Warner Bros. Entertainment, USA.

Ma, Z. & Jiang, M., 2013, 'Interpretation of verbal humor in the sitcom *The Big Bang Theory* from the perspective of adaptation-relevance theory', *Theory and Practice in Language Studies* 3(12), 2220–2226. http://dx.doi.org/10.4304/tpls.3.12.2220-2226

Olsson, J., 2008, *Forensic linguistics*, Continuum International Publishing Group, London.

Palmeira, M., 2015, 'Abstract language signals power, but also lack of action orientation', *Journal of Experimental Social Psychology* 61, 59–63. http://dx.doi.org/10.1016/j.jesp.2015.07.003

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

Renkema, J., 2004, *Introduction to discourse studies*, John Benjamins, Amsterdam.

Rey, J.M., 2001, 'Changing gender roles in popular culture: Dialogue in Star Trek episodes from 1966–1993', in S. Conrad & D. Biber (eds.), *Variation in English multi-dimensional studies*, pp. 138–156. Pearson Education, Essex.

Scott, M., 2010, 'Problems in investigating keyness, or clearing the undergrowth and marking out trails', in M. Bondi & M. Scott (eds.), *Keyness in texts*, pp. 43–57, John Benjamins, Amsterdam.

Scott, M., 2012, *WordSmith Tools version 6,* computer software, Lexical analysis software, Liverpool.

Shuqin, H., 2012, 'An analysis of humor in *The Big Bang Theory* from pragmatic perspectives', *Theory and Practice in Language Studies* 2(6), 1185–1190, viewed 03 October 2015, from http://www.academypublication.com/issues/past/tpls/vol02/06/13.pdf

Shuqin, H., 2013, 'A relevance theoretic analysis of verbal humor in *The Big Bang Theory*', *Studies in Literature and Language* 7(1), 10–14, viewed 03 October 2015, from http://www.cscanada.net/index.php/sll/article/download/j.sll.1923156320130701.2549/5034

Stubbs, M., 1996, *Text and corpus analysis*, Blackwell, Oxford.

Stubbs, M., 2010, 'Three concepts of keywords', in M. Bondi & M. Scott (eds.), *Keyness in texts*, pp. 21–42, John Benjamins, Amsterdam.

Wakslak, C.J., Smith, P.K. & Han, A., 2014, 'Using abstract language signals power', *Journal of Personality and Social Psychology* 107(1), 41–55, http://dx.doi.org/10.1037/a0036626

Yin, Z. & Yun, M., 2012, 'Rhetorical devices in dialogues of *The Big Bang Theory*', *Sino-US English Teaching* 9(6), 1220–1229, viewed 03 October 2015, from http://www.davidpublishing.com/davidpublishing/upfile/7/23/2012/2012072382010977.pdf

Appendix starts on the next page →

Appendix 1

TABLE 1-A1: Sheldon versus the rest of the keywords list (*p*-value at 0.000001) generated with WordSmith 6.0.

N	Keyword	Frequency	%	Texts	Reference corpus frequency	Reference corpus (%)	Keyness
1	OF	2159	1.74	7	2352	1.16	183.38
2	THE	4144	3.34	7	5106	2.53	182.44
3	AS	467	0.38	7	330	0.16	138.14
4	LEONARD	537	0.43	7	452	0.22	107.71
5	BY	304	0.25	7	227	0.11	80.40
6	AN	474	0.38	7	428	0.21	78.42
7	WHICH	179	0.14	7	100	0.05	77.87
8	LORD	50	0.04	7	4	-	72.04
9	PERHAPS	57	0.05	7	8	-	69.43
10	IS	1446	1.17	7	1790	0.89	60.52
11	PENNY	409	0.33	7	385	0.19	59.58
12	DEAR	52	0.04	6	10		55.38
13	IN	1477	1.19	7	1891	0.94	48.39
14	BAZINGA	28	0.02	4	1	-	46.42
15	WIL [*sic*]	50	0.04	5	13	-	45.02
16	MAY	127	0.10	7	83	0.04	43.25
17	WILL	263	0.21	7	241	0.12	41.60
18	BEGIN	30	0.02	7	3	-	40.80
19	COOPER	69	0.06	7	31	0.02	39.31
20	WHEATON	38	0.03	5	8	-	38.65
21	HELLO	129	0.10	7	91	0.05	38.21
22	AGREEMENT	44	0.04	6	13	-	36.34
23	ITS	58	0.05	7	24	0.01	36.01
24	COURSE	135	0.11	7	101	0.05	35.52
25	OUR	230	0.19	7	213	0.11	35.22
26	REMARKABLE	18	0.01	6	0	-	34.82
27	INTEREST	28	0.02	7	4	-	33.88
28	FOWLER	26	0.02	5	3	-	33.87
29	FLAGS	21	0.02	3	1	-	33.44
30	SOCIAL	40	0.03	7	12	-	32.68
31	GIVEN	45	0.04	7	16	-	32.15
32	AMY	178	0.14	5	157	0.08	31.43
33	WOULD	375	0.30	7	409	0.20	31.31
34	RESULT	16	0.01	6	0	-	30.95
35	UNIVERSE	41	0.03	7	14	-	30.30
36	THEORY	39	0.03	7	13	-	29.40
37	NOBEL	27	0.02	7	5	-	29.27
38	BEVERAGE	21	0.02	6	2	-	28.95
39	SPOCK	34	0.03	4	10	-	28.17
40	MY	1258	1.01	7	1687	0.83	27.47
41	QUITE	32	0.03	7	9	-	27.35
42	FARRAH	22	0.02	3	3	-	27.08
43	KRIPKE	33	0.03	5	10	-	26.76
44	STRING	33	0.03	7	10	-	26.76
45	ALSO	77	0.06	7	50	0.02	26.51
46	HAS	205	0.17	7	200	0.10	26.50
47	SUGGEST	25	0.02	7	5	-	26.11
48	COITUS	28	0.02	7	7	-	25.83
49	NOW	419	0.34	7	486	0.24	25.82
50	ALTHOUGH	41	0.03	7	17	-	25.40
51	SUPPOSE	34	0.03	7	12	-	24.44
52	ROOMMATE	48	0.04	6	24	0.01	24.15
53	MOMENT	48	0.04	7	24	0.01	24.15
54	APPROPRIATE	22	0.02	7	4	-	24.06
55	CUTE	5	-	3	50	0.02	-24.00
56	DIDN'T	137	0.11	7	361	0.18	-24.46
57	STAY	14	0.01	5	80	0.04	-24.49
58	CAN	363	0.29	7	806	0.40	-24.83

Appendix Table 1-A continues on the next page →

TABLE 1-A1 (continues…): Sheldon versus the rest of the keywords list (*p*-value at 0.000001) generated with WordSmith 6.0.

N	Keyword	Frequency	%	Texts	Reference corpus frequency	Reference corpus (%)	Keyness
59	SHE	230	0.19	7	550	0.27	-25.01
60	HUH	23	0.02	6	107	0.05	-25.51
61	BAD	38	0.03	7	146	0.07	-25.75
62	WEIRD	6	-	4	56	0.03	-25.76
63	TOO	109	0.09	7	307	0.15	-26.02
64	LISTEN	28	0.02	7	122	0.06	-26.47
65	ASS	5	-	3	54	0.03	-27.09
66	FINE	91	0.07	7	273	0.14	-27.82
67	DAMN	7	-	4	63	0.03	-28.30
68	TALKING	42	0.03	7	162	0.08	-28.78
69	OOH	16	0.01	6	95	0.05	-30.28
70	HELL	18	0.01	4	101	0.05	-30.32
71	DEAL	11	-	5	80	0.04	-30.71
72	RAJ	64	0.05	7	219	0.11	-30.74
73	SORRY	147	0.12	7	403	0.20	-31.32
74	NICE	55	0.04	7	199	0.10	-31.35
75	LIKE	432	0.35	7	968	0.48	-31.47
76	HOWARD	94	0.08	7	290	0.14	-31.85
77	ARE	574	0.46	7	1238	0.61	-31.85
78	HOW	315	0.25	7	750	0.37	-33.44
79	LOVE	64	0.05	7	229	0.11	-35.29
80	HANG	28	0.02	7	139	0.07	-36.12
81	DON'T	615	0.50	7	1348	0.67	-38.48
82	HONEY	4	-	3	64	0.03	-38.54
83	HE'S	119	0.10	7	363	0.18	-38.71
84	MA	4	-	2	66	0.03	-40.22
85	KIDDING	10	-	6	91	0.05	-41.18
86	TELL	120	0.10	7	372	0.18	-41.38
87	OUT	355	0.29	7	866	0.43	-43.17
88	DOING	85	0.07	7	299	0.15	-44.51
89	THANKS	31	0.03	7	165	0.08	-46.68
90	SHELDON'S	4	-	2	74	0.04	-46.98
91	SEX	29	0.02	6	162	0.08	-48.41
92	IT	1371	1.11	7	2800	1.39	-48.43
93	OH	691	0.56	7	1540	0.76	-48.64
94	GET	324	0.26	7	845	0.42	-55.23
95	WOW	16	0.01	7	135	0.07	-58.04
96	HER	260	0.21	7	721	0.36	-58.25
97	STUFF	10	-	6	114	0.06	-58.89
98	COOL	13	0.01	6	126	0.06	-59.35
99	BERNADETTE	22	0.02	5	158	0.08	-60.05
100	LITTLE	138	0.11	7	459	0.23	-60.56
101	PRETTY	23	0.02	7	164	0.08	-61.96
102	TALK	67	0.05	7	290	0.14	-62.33
103	DO	583	0.47	7	1392	0.69	-63.06
104	LOOK	155	0.13	7	507	0.25	-64.38
105	SURE	78	0.06	7	323	0.16	-64.78
106	HI	23	0.02	6	169	0.08	-65.45
107	HIM	183	0.15	7	576	0.29	-66.66
108	GIRL	20	0.02	7	164	0.08	-69.09
109	KIND	45	0.04	7	243	0.12	-69.91
110	GOD	24	0.02	6	184	0.09	-73.70
111	MEAN	74	0.06	7	333	0.16	-75.82
112	UP	300	0.24	7	857	0.42	-76.02
113	GREAT	85	0.07	7	365	0.18	-77.53
114	GUY	23	0.02	7	196	0.10	-84.87
115	WANT	222	0.18	7	710	0.35	-85.51
116	UH	118	0.10	7	471	0.23	-88.93
117	WE	566	0.46	7	1452	0.72	-89.36
118	WHAT'S	76	0.06	7	363	0.18	-89.78
119	ABOUT	341	0.28	7	990	0.49	-92.20

Appendix Table 1-A continues on the next page →

TABLE 1-A1 (continues…): Sheldon versus the rest of the keywords list (*p*-value at 0.000001) generated with WordSmith 6.0.

N	Keyword	Frequency	%	Texts	Reference corpus frequency	Reference corpus (%)	Keyness
120	MAYBE	49	0.04	7	292	0.14	-93.47
121	GO	295	0.24	7	905	0.45	-98.18
122	KNOW	509	0.41	7	1397	0.69	-109.42
123	COME	150	0.12	7	608	0.30	-117.86
124	REALLY	151	0.12	7	640	0.32	-133.34
125	SO	440	0.35	7	1324	0.66	-136.61
126	GOT	134	0.11	7	652	0.32	-165.40
127	WHAT	665	0.54	7	2044	1.01	-223.84
128	JUST	408	0.33	7	1467	0.73	-229.30
129	YOU	3488	2.81	7	7728	3.82	-242.30
130	GONNA	38	0.03	7	500	0.25	-277.50
131	GUYS	5	-	3	386	0.19	-325.66
132	YEAH	200	0.16	7	1129	0.56	-342.27
133	SHELDON	114	0.09	7	1019	0.50	-456.89
134	HEY	36	0.03	7	745	0.37	-491.33
135	OKAY	166	0.13	7	1316	0.65	-542.58