Creativity & Television Drama: a *t*-score and MI Value Cut-offs Analysis of Pattern-forming Creativity in *House M.D.*

Abstract

Carter (2004) theorizes creativity in everyday common talk into two main categories: pattern-reforming and pattern-forming. This paper extends the discussion on pattern-forming creativity appeared in popular TV drama *House M.D.*, with an attempt to demonstrate how statistical devices such as *t*-score and MI value may be used to facilitate the extraction of this type of linguistic creativity. The extraction is facilitated by 2-word concgrams, mutual information (MI) and *t*-score, which are generated from the TV drama's dialogue corpus. The *t*-scores and MI values of 2-word concgrams from a total of 67 episodes are analysed through a quantitative approach. It is found that *t*-score and MI value of 2-word concgrams can be used to locate pattern-forming creativity, and their cut-offs can effectively double the percentage yield of pattern-forming creativity. This paper proposes several ways to improve efficiency in the extraction of pattern-forming creativity and provides statistical evidence for the relationship between pattern-forming creativity and MI and *t*-score.

Keywords: creativity; House M.D., TV drama; pattern-forming; mutual information; *t*-score;

1. Introduction

While linguistic creativity studies have been made simpler by the advances in computational power, the corpus research in linguistic creativity in general is not without hurdles. For corpus-driven investigation, in which “corpora are used as sources of empirical data (linguistic, socio-cultural, textual) against which intuitions about creativity are tested or preliminary findings from smaller data sets are validated” (Vo & Carter, 2010, p. 310), the challenge is even greater. Vo and Carter (2010, p. 310) list two main difficulties in pursuing corpus-driven investigations. The first is the limitations of computer software development, making “the identification and systematic extraction of linguistic creativity in both spoken and written corpora … the can of worm in the corpus linguistics – creativity nexus.” The second is the amount of manual work and the “hit-or-miss nature” of the searches, which have proven to be too “laborious, time-consuming and not always sufficiently reliable” (Vo & Carter, 2010, p. 310). Therefore, efficiency of extraction is a major concern for researchers of linguistic creativity.

One of the ways to improve efficiency of extraction of linguistic creativity is to utilise statistical devices and cut-offs on the data. Some statistics such as *t*-score and MI value have been widely adopted in corpus linguistics to assist researchers locate meaningful words and phrases, particularly in the study of collocations (Hunston, 2002; McEnery, Xiao, & Tono, 2006). Their respective cut-offs for *t*-score and MI values at 2.00 and 3.00 are often used to discriminate between collocates which are “linguistically interesting” and otherwise (Church & Hanks, 1990, p. 24; Barnbrook, 1996). However, the aforementioned statistics and cut-offs are not without disputes. Stubbs (1995, p. 9) has warned that the term “linguistically interesting” is “admittedly undefined” and that the cut-off values are based purely on “empirical analyses” and have “no strong theoretical reason” for making such selection. Cheng, Greaves and Warren (2006) have looked at the effectiveness of *t*-score, MI value and the retainability of the collocates using their respective cut-offs at 2.00 and 3.00. They conclude that, in a one-million-word sample of the Hong Kong Corpus of Spoken English (HKCSE), they are “reluctant to fully endorse the *t*-score and MI value” (Cheng, Greaves, & Warren, 2006, p. 14), as their respective cut-offs at 2.00 and 3.00 are ineffective in terms of retaining key collocates. A pair of custom cut-offs for *t*-score and MI value calculated from a dataset is likely be more desirable. Based on this notion, this paper proposes a time-saving method in achieving a balance between efficiency of extraction and collocate retention of pattern-forming creativity.

Carter (2004) hypothesizes two types of linguistic creativity in everyday common talk: pattern-reforming creativity and pattern-forming creativity. The former involves breaking patterns (such as a student commenting about a website ‘I came, I saw, I logged off,’ instead of ‘I came, I saw, I conquered,’) while the latter involves repetition (such as Tony Blair saying ‘Education, education,

education.’) (Carter, 2016). The latter is prone to collocation because of its repetitive property, and so will likely be reflected by large *t*-score and MI value. Using the 2-word concgrams, *t*-scores and MI values generated from a corpus, we can formulate custom cut-offs for the statistics which govern the appearances of pattern-forming creativity. Corpus software ConcGram v1.0 is used to compute the 2-word concgrams, *t*-scores and MI values, while the data corpus is constructed using the fan scripts of television drama *House M.D.* This data corpus, the *House M.D.* Corpus (HMDC), is a key resource for several other investigations of linguistic creativity [name deleted to maintain the integrity of the review process].

In the next section, I will first explain the choice of using fan scripts of *House M.D.* as data sources. I will then describe the steps involved in creating a corpus using the fan scripts, before defining the types of pattern-forming creativity which will be studied. Further preparation for the quantitative analysis of this study involves three data manipulating steps: 1) the calculation of internal span, 2) the manual extraction of pattern-forming creativity, and 3) the calculation of MI value cut-off and t-score cut-off.

2. Data and Methodology

2.1 *House M.D.* as Data Source

*House M.D.*is an American television medical ‘dramedy’ stretching eight seasons with a total of 177 episodes aired on the FOX Network from 16th November 2004 (ABC Medianet, 2004) to 21st May 2012 (TV By The Numbers, 2012). The series is based on the premise (which is also the title of the pilot), “Everybody lies” (Werts, 2009), a motto inscribed deep in the mind of Dr. Gregory House (Hugh Laurie), the main character who is inspired by Sir Arthur Conan Doyle’s renowned fictional detective Sherlock Holmes (Slate, 2006).

House M.D. is selected for a number of strong reasons. Firstly, it is written with creativity and language quality very much worth exploring and exploiting (Olson, 2010; Richardson, 2010). Secondly, it is a popular television program which has set 3 Guinness World Records (namely the world’s most popular TV show, the world’s most watched man on television and the world's highest-paid TV actor in a drama series) (Guinness World Record News, 2012), as well as winning 2 Golden Globes, 49 awards and 112 nominations. Bignell and Lacey (2005, p. 6) argue that “it is television’s very familiarity, and its conventional focus upon the familiar, the present time and the everyday, that opens up alternative formal and stylistic possibilities.” Bednarek (2010) echoes that popularity of television and programmes alone is worthy of study due to its significant impact on our daily lives and societies. These world records and arguments make *House M.D.* a worthy candidate for this study. Thirdly, the main character Dr. Gregory House has been the inspiration for many publications from medical science (Sanders, 2009; Holtz, 2006; 2011), medical humanities (Goodier & Arrington, 2007), philosophy (Jacoby & Irwin, 2008), psychology (Clyman, 2009; Jamieson, 2011; Cascio & Martin, 2011; Whitbourne, 2012; Li & Csikszentmihalyi, 2014) and media studies (Jackman & Laurie, 2010; Holtz, 2011; Hockley & Gardner, 2011), thereby playing a critical role in the construction of popular memory (Bignell & Lacey, 2005) and in academia. A linguistic study of House’s creativity will bridge the existing work on House from the aforementioned disciplines. Lastly, *House M.D.* is a unique creative instance in the modern television history of medical dramedy (Li & Csikszentmihalyi, 2014). It is built around one single central character, providing longitudinality in the creativeness of its repertoire and subsequently, an opportunity for the studies of creative language use to expand beyond the written form and into the scripted spoken counterpart.

2.2 Creating the HMDC

HMDC uses fan scripts – the actual transcripts from television produced by multiple ‘fans’ (Bednarek, 2010) – as the input data. The construction of the HMDC involves three major steps. Step one is the data collection of *House M.D.* fan scripts of every episode from the internet (therefore not the original screenwriters’ scripts). While fan scripts are not 100% accurate, they are selected for a number of reasons. Firstly, the finalised original scripts are inaccessible to the public. Secondly, as Bednarek (2010, p. 70) points out, fan scripts are “much more accurate than subtitles (which could be automatically extracted as alternative data source), with a much greater number of and more significant mistakes in the subtitles than in the transcripts.” Lastly, “[m]anual transcription by the researcher may in fact result in similar inaccuracies as are present in the fan transcripts (e.g. typos), and simply is not feasible for a large-scale corpus analysis" (Bednarek, 2010, p. 70). Since the *House M.D.* fan scripts used in this study are available online and have been 'peer reviewed' by other their readers – in which corrections are continuously suggested and made by the fan script readers (clinic_duty, 2007) – I have decided to adapt the fan scripts and improve their accuracies. Step two is the removal of all non-dialogue elements such as fade-ins, scene headings, action sequences, scene transitions, mood brackets, parentheticals, commercial tags and character name tags. Once the non-dialogue elements are removed, the 'pure' dialogues are stored as txt-format in 177 individual files (one file per episode) to form a raw, unscripted and unannotated version of HMDC. Step three is to improve accuracy of the transcribed dialogues in the HMDC. Every line has been manually checked against the actual lines performed by the actors in the television series after watching all episodes at least eight times. Further spell checks are performed repeatedly throughout four years of this study whenever possible and necessary. This longitudinal effort has helped to reduce the corpus impurities and improve accuracy of future calculations. The result is a 927,922-word cleaned HMDC.

2.3 *Defining Pattern-Forming Creativity*

Pattern-forming creativity is closely related to verbal repetition in conversations (Carter, 2016). According to Tannen ([1989] 2007, p. 101), verbal repetition is "a resource by which conversationalists together create a discourse, a relationship, and a world. It is the central linguistic meaning-making strategy, a limitless resource for individual creativity and interpersonal involvement." In film or in a TV drama such as *House M.D.*, the consistent use of verbal repetition by a character is a character trait – also known as a motif, which is central to the viewers' familiarization and identification of characters (Bordwell & Thompson, [1990] 2008).

Pattern-forming creativity occurs in co-constructed as well as non-co-constructed, self-repeated forms. The former "is more likely to grow out of dialogic interaction" and the latter "can occur in monologues and in the context of a transmission of information" (Carter, 2004, p. 139). Concurrently, pattern-forming creativity can also be in the form of synchronic repetition: "repeating one's own or another's words within a discourse", or diachronic repetition:

“repeating words from a discourse distant in time.” (Tannen, [1989] 2007, p. 102). Therefore, a total of four combinations of pattern-reforming creativity is studied: non-co-constructed, self-repetition (synchronic), non-co-constructed, self-repetition (diachronic), co-constructed repetition (synchronic) and co-constructed repetition (diachronic).

Due to the episodic design of television series, the connection between one instance of pattern-forming creativity and another within a certain span of words in the same episode will be strong. However, this kind of repetitive process rarely crosses from one episode to the next; in other words, there is negligible connection between one instance in one episode and another instance in a different episode. Therefore, in order to avoid the inclusion of pattern-forming creativity across episodes, the extraction of pattern-forming creativity must be performed on a per-episode basis. In the extraction of pattern-forming creativity, ConcGram v1.0 is selected for its capability to find all permutations of word co-occurrences within certain span (Greaves, 2009). An alternative program to ConcGram v1.0 is WordSmith Tools 6.0’s

WSConcgram, but ConcGram v1.0 is chosen for its superiority in processing speed.

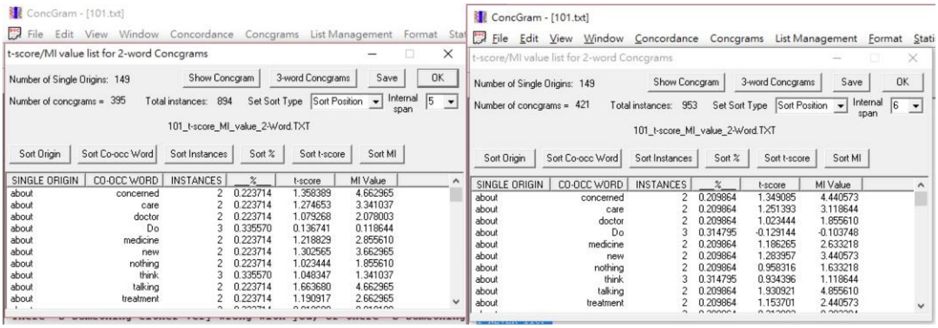
2.4 Calculating internal span

Internal span is one of the key components required by ConcGram v1.0 in the creation of a 2-word concgram list and *t*-score / MI value lists for the extraction of pattern-forming creativity. By the definition given in the Concgram Manual, setting an internal span of 2 refers to the display of all concgram permutations up to two intervening words (i.e. AB, A*B and A**B) (Greaves, 2009). While selecting the maximum possible internal span allowed by the software (max = 10) does provide a wider possible coverage of pattern-forming creativity, it will also lower the percentage of creativity hit rate due to the increase in non-creativity-bearing concgrams, which will eventually lead to the waste of time in the process of pattern-forming creativity extraction. Therefore, in order to achieve a balance between creativity hit rate and time efficiency, there is a need to find the optimal word span for the computation of concgrams. The approach is to calculate the required internal span based on an overall mean value of the averages of words per sentence, or sentence span, in every episode of the TV series.

Tabl**e** 1: A sim*plified table of average number of words per sentence by episode*

Episode	Average words per sentence	Episode	Average words per sentence
101	7.8	501	7
102	7.6	502	6.8
103	7.3	503	6.8
⋮	⋮	⋮	⋮
415	6	821	6.7
416	5.9	822	7.4
Std. Dev.	0.61654754	Mean	6.893103448

Table 1 shows the average number of words per sentence from episode 1 of season 1 to episode 22 of season 8 in HMDC, each one of them obtained using Microsoft Word’s *Word Count* function. The last row of the table shows the mean of all averages of 6.893 words per sentence and the standard deviation of 0.617. Taking one standard deviation above mean and a sentence span of 7.507 ($= 6.893 + 0.617$) is obtained. At this point, taking both the ceiling and floor of this value may be reasonable as 7.507 lies between 7 and 8, but because a difference of 1 in sentence span will result in a difference of around a hundred instances of concgram, as shown in an example in Figure 1,



Number of Single Origins: 149

Number of concgrams = 395 Total instances: 894 Set Sort Type: Sort Position Internal span: 5

101_t-score_MI_value_2-Word.TXT

SINGLE ORIGIN	CO-OCC WORD	INSTANCES	%	t-score	MI Value
about	concerned	2	0.223714	1.358389	4.662965
about	care	2	0.223714	1.274653	3.341037
about	doctor	2	0.223714	1.079268	2.078003
about	Do	3	0.335570	0.136741	0.118644
about	medicine	2	0.223714	1.218829	2.855610
about	new	2	0.223714	1.302565	3.662965
about	nothing	2	0.223714	1.023444	1.855610
about	think	3	0.335570	1.048347	1.341037
about	talking	2	0.223714	1.663680	4.662965
about	treatment	2	0.223714	1.190917	2.662965

Number of Single Origins: 149

Number of concgrams = 421 Total instances: 953 Set Sort Type: Sort Position Internal span: 6

101_t-score_MI_value_2-Word.TXT

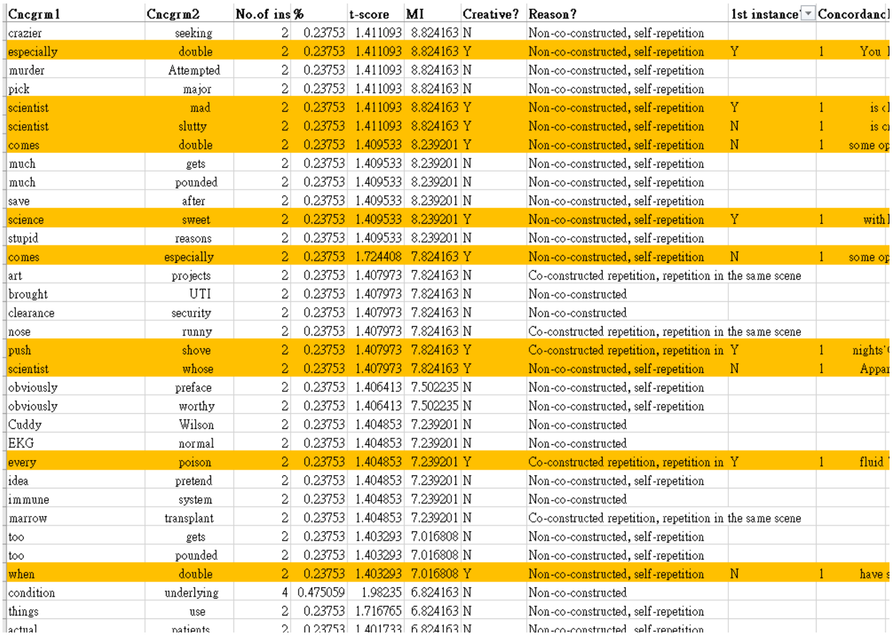
SINGLE ORIGIN	CO-OCC WORD	INSTANCES	%	t-score	MI Value
about	concerned	2	0.209864	1.349085	4.440573
about	care	2	0.209864	1.251393	3.118644
about	doctor	2	0.209864	1.023444	1.855610
about	Do	3	0.314795	-0.129144	-0.103748
about	medicine	2	0.209864	1.186265	2.633218
about	new	2	0.209864	1.283957	3.440573
about	nothing	2	0.209864	0.958316	1.633218
about	think	3	0.314795	0.934396	1.118644
about	talking	2	0.209864	1.930921	4.855610
about	treatment	2	0.209864	1.153701	2.440573

Figure 1*: Difference in the number of concgrams with internal span 5 (left) vs 6 (right)*

taking the floor of 7.507 (= 7) should provide adequate coverage for sentences of average word length while maintaining sufficient balance between creativity hit rate and time required for the extraction of pattern-forming creativity. Since internal span is the “intervening words between the centre word and the outer co-occurring word in a concgram” (Greaves, 2009, p. 35), a sentence span of 7 will equate to an internal span of 5 (= sentence span – centre word – outer co-occurring word), hence the choice of internal span for the computation of concgrams. Since statistical devices available in ConcGram v1.0 such as *t*-score and MI are only available with 2-word concgram, only 2-word concgram lists are generated using internal span = 5. The result is a *t*-score/MI value list for 2-word congrams, which can be sorted according to needs.

2.5 Manual extraction of pattern-forming creativity

The extraction of pattern-forming creativity is facilitated by the *t*-score/MI value list for 2-word concgrams. The list generated by ConcGram v1.0 from each episode is first exported as an Excel spreadsheet and sorted by MI / *t*-score value as shown in Figure 2. Each concgram and their instances are then manually checked against their original video source, dialogues and context for the presence or absence of pattern-forming creativity. Results are then recorded under column 'Reason?' and marked under column 'Creative?' as 'Y' for yes if they are present and 'N' for no if there are absent.



Cncgrm1	Cncgrm2	No.of ins	%	t-score	MI	Creative?	Reason?	1st instance	Concordanc
crazier	seeking	2	0.23753	1.411093	8.824163	N	Non-co-constructed, self-repetition		
especially	double	2	0.23753	1.411093	8.824163	Y	Non-co-constructed, self-repetition	Y	1 You
murder	Attempted	2	0.23753	1.411093	8.824163	N	Non-co-constructed, self-repetition		
pick	major	2	0.23753	1.411093	8.824163	N	Non-co-constructed, self-repetition		
scientist	mad	2	0.23753	1.411093	8.824163	Y	Non-co-constructed, self-repetition	Y	1 is
scientist	slutty	2	0.23753	1.411093	8.824163	Y	Non-co-constructed, self-repetition	N	1 is
comes	double	2	0.23753	1.409533	8.239201	Y	Non-co-constructed, self-repetition	N	1 some op
much	gets	2	0.23753	1.409533	8.239201	N	Non-co-constructed, self-repetition		
much	pounded	2	0.23753	1.409533	8.239201	N	Non-co-constructed, self-repetition		
save	after	2	0.23753	1.409533	8.239201	N	Non-co-constructed, self-repetition		
science	sweet	2	0.23753	1.409533	8.239201	Y	Non-co-constructed, self-repetition	Y	1 with
stupid	reasons	2	0.23753	1.409533	8.239201	N	Non-co-constructed, self-repetition		
comes	especially	2	0.23753	1.724408	7.824163	Y	Non-co-constructed, self-repetition	N	1 some op
art	projects	2	0.23753	1.407973	7.824163	N	Co-constructed repetition, repetition in the same scene		
brought	UTI	2	0.23753	1.407973	7.824163	N	Non-co-constructed		
clearance	security	2	0.23753	1.407973	7.824163	N	Non-co-constructed		
nose	runny	2	0.23753	1.407973	7.824163	N	Co-constructed repetition, repetition in the same scene		
push	shove	2	0.23753	1.407973	7.824163	Y	Co-constructed repetition, repetition in	Y	1 nights'
scientist	whose	2	0.23753	1.407973	7.824163	Y	Non-co-constructed, self-repetition	N	1 Appar
obviously	preface	2	0.23753	1.406413	7.502235	N	Non-co-constructed, self-repetition		
obviously	worthy	2	0.23753	1.406413	7.502235	N	Non-co-constructed, self-repetition		
Cuddy	Wilson	2	0.23753	1.404853	7.239201	N	Non-co-constructed		
EKG	normal	2	0.23753	1.404853	7.239201	N	Non-co-constructed		
every	poison	2	0.23753	1.404853	7.239201	Y	Co-constructed repetition, repetition in	Y	1 fluid
idea	pretend	2	0.23753	1.404853	7.239201	N	Non-co-constructed, self-repetition		
immune	system	2	0.23753	1.404853	7.239201	N	Non-co-constructed		
marrow	transplant	2	0.23753	1.404853	7.239201	N	Co-constructed repetition, repetition in the same scene		
too	gets	2	0.23753	1.403293	7.016808	N	Non-co-constructed, self-repetition		
too	pounded	2	0.23753	1.403293	7.016808	N	Non-co-constructed, self-repetition		
when	double	2	0.23753	1.403293	7.016808	Y	Non-co-constructed, self-repetition	N	1 have
condition	underlying	4	0.475059	1.98235	6.824163	N	Non-co-constructed		
things	use	2	0.23753	1.716765	6.824163	N	Non-co-constructed, self-repetition		
actual	patients	2	0.23753	1.401733	6.824163	N	Non-co-constructed, self-repetition		

Figure 2: *2-word concgram on Excel spreadsheet, sorted by MI, highlighting pattern-forming creativity*

Table 2 shows how descriptions of pattern-forming creativity under column 'Reason?' are categorized. The descriptions fall into two main categories: absent/undetected and present. If instances of a concgram indicate presence of pattern-forming creativity, they are classified into 'Non-co-constructed, self-repetition' – for pattern-forming creative instances of a concgram showing repetitions produced by one speaker, and 'Co-constructed repetition' – for pattern-forming creative instances of a concgram showing repetitions produced by two or more speakers. These two categories are then further supplemented by '…in the same scene / synchronic repetition' or '…across scenes / diachronic repetition' to represent the complete scenarios (Tannen, [1989] 2007, p. 102). Otherwise, if instances of a concgram indicate absence of pattern-forming creativity or that such creativity has not been detected, an additional description type 'Non-co-constructed' may apply to the four aforementioned.

Table 2*: Combinations of 'Reasons' for pattern-forming creativity*

Presence / absence of pattern-forming creativity	Type	Repetition in scene(s)		
		(-)	…in the same scene / synchronic repetition	…across scenes / diachronic repetition
Absent/undetected	Non-co-constructed	✓	N.A.	N.A.
Present or absent/undetected	Non-co-constructed, self-repetition	N.A.	✓	✓
	Co-constructed repetition	N.A.	✓	✓

The following examples demonstrate how each scenario is categorized. The concgrams recognised by the software are underlined.

1 Babbled like a baby. Present deterioration of mental status. See that? They all assume I 'm a patient

1 minutes later and she did just fine. The altered mental status is intermittent, just like the verbal

Pattern-forming creativity is absent/undetected. Instances of concgram do not show any signs of self-repetition or co-construction with no direct reference to the same idea. ‘Non-co-constructed’ is displayed.

1 you ever seen a worm under an x-ray, a regular old no contrast 100-year-old technology x-ray? They

2 an x-ray, a regular old no contrast 100-year-old technology x-ray? They light up like shotgun

Pattern-forming creativity is absent/undetected. Instances of concgram show a non-constructed, repetition use by one single speaker in the same scene. ‘Non-co-constructed, self-repetition in the same scene’ is displayed.

1 ca n't trust people. She probably knew she was allergic to gadolinium, figured it was an easy way to get

2 It 'll just be another minute. She 's having an allergic reaction to gadolinium. She 'll be dead in two

Pattern-forming creativity is absent/undetected. Instances of concgram show a non-constructed, repetition use by one single speaker in two separate scenes regarding the same idea. ‘Non-co-constructed, self-repetition across scenes’ is displayed.

1 the inflammation. The more often this happens... What? “The more often this happens...”What??

2 this happens... What? “The more often this happens...”What?? Forget it. If you do n't trust

Pattern-forming creativity is absent/undetected. Instances of concgram show a co-constructed repetition by 2 or more speakers in the same scenes regarding the same idea. ‘Co-constructed repetition, repetition in the same scene’ is displayed.

1 Because you guys were right. He did n't have two conditions at the exact same time. First, he got a cough.

2 Tell the family House 's theory? Two odd conditions striking completely coincidentally at the exact

Pattern-forming creativity is absent/undetected. Instances of concgram show a co-constructed repetition by 2 or more speakers in two separate scenes regarding the same idea. ‘Co-constructed repetition, repetition across scenes’ is displayed.

1 of the medicine, too. She probably weighed that danger against the danger of not breathing. Oxygen is so

2 She probably weighed that danger against the danger of not breathing. Oxygen is so important during

Pattern-forming creativity is present. Instances of concgram show a non-constructed, repetition use by one single speaker in the same scene. ‘Non-co-constructed, self-repetition in the same scene’ is displayed.

1 's cave. Car 's clean. Did you just see a blond guy with a pretentious accent? Ca n't see an

2 episodes and a heart attack. Do you see a blond guy who still has peach fuzz standing up there?

Pattern-forming creativity is present. Instances of concgram show a non-constructed, repetition use by one single speaker in two separate scenes regarding the same idea. ‘Non-co-constructed, self-repetition across scenes’ is displayed.

1 country doctor. Brain tumors at her age are highly unlikely. She 's 29. Whatever she 's got is

2 unlikely. She 's 29. Whatever she 's got is highly unlikely. Protein markers for the three most

Pattern-forming creativity is present. Instances of concgram show a co-constructed repetition by 2 or more speakers in the same scenes regarding the same idea. ‘Co-constructed repetition, repetition in the same scene’ is displayed.

1 you to do your job. Well, like the philosopher Jagger once said, “You ca n't always get what you want.”

2 Oh, I looked into that philosopher you quoted, Jagger, and you 're right, “You ca n't always get what

Pattern-forming creativity is present. Instances of concgram show a co-constructed repetition by 2 or more speakers in two separate scenes regarding the same idea. ‘Co-constructed repetition, repetition across scenes’ is displayed.

2.6 Calculating MI value cut-off and t-score cut-off

As the 2-word concgram lists contain instances of non-pattern-forming creativity, ideally it is best to perform manual search and extraction of pattern-forming creativity from each 2-word concgram in every list generated for every single episode; however, taking Season 1 Episode 1 as example, if every episode generates at least 395 concgrams then there will be 395 x 177 episodes = 69,915 congrams and a minimum of 69,915 x 2 = 139,830 instances to be manually checked. Time-wise, it is highly impractical. It is thus necessary to determine a reasonable cut-off value which reduces the total number of concgrams to the minimum, maximises the hit rate of pattern-forming creativity and saves time on manual checking. As mentioned previously, ConcGram v1.0 uses a default MI cut-off value at 3.000000 and *t*-score cut-off value at 2.000000. However, celebrated these values are, whether empirical or theoretical, using these default cut-off values may not provide the optimal threshold that meets the specificity of HMDC. Therefore, it is preferable to establish a custom MI cut-off and *t*-score cut-off from the data instead.

An approach of small sample's averages is used. First, 2-word concgram lists of two selected episodes (Season 1 Episode 1 *Pilot: Everybody lies* for it is the beginning of the show, Season 4 Episode 11 *Frozen* for it is near the middle of the entire series and also the episode with the highest U.S. viewers of the entire series (Seidman, 2008)) are generated and exported as Excel spreadsheets. After manual extraction of pattern-forming creativity has been performed, all extracted concgrams of pattern-forming creativity are further manually checked to determine if they are the first instance to appear in this list. On the 2-word concgram on Excel spreadsheet as shown in Figure 2, they are marked ‘Y’ under column ‘1st instance?’ with the row highlighted if the instance is the first appearance and ‘N’ if the instance has appeared earlier on in the list. This step is performed when it is sorted by MI value and repeated when sorted by *t*-score as shown. The final 'Y' of the column, that is the final first appearance of a pattern-forming creativity in a MI or *t*-score sorted concgram list, gives the cut-off value with which the data is sorted.

Table 3 below shows the cut-offs for MI and *t*-score with respect to the selected episodes:

***Table 3**: Calculation of MI and t-score cut-offs, with choice of values highlighted*

	Season 1 Episode 1 Pilot: Everybody lies	Season 4 Episode 11 *Frozen*	Cut-off average
Number of concgrams	395	373	N.A.
Total instances	894	861	N.A.
MI value cut-off	4.740968	2.792806	3.766887
t-score cut-off	1.361327	1.268443	1.314885
Number of concgrams after cut-offs	201	155	N.A.
Number of concgram instance after cut-offs	437	359	N.A.
Percentage of concgram instances removed after cut-offs	51.12%	58.30%	N.A.

Taking the average of MI (= (4.740968 + 2.792806) / 2) and *t*-score (= (1.361327 + 1.268443) / 2) from the two episodes, the MI cut-off of 3.766887 and *t*-score cut-off of 1.314885 are obtained. Both the MI value and *t*-score cut offs are used simultaneously as filtering criteria of the 2-word concgrams as suggested by Stubbs (1995). Using such averages as cut-offs is by no mean perfect, as some instances of pattern-forming creativity would be excluded. A more accurate cut-off can be calculated if more episodes are considered. However, it is worth noting that cut-offs are designed to maximise hit-rates within a minimal amount of time, not designed to ensure 100 percent selection of instances. As Stubbs (1995, p. 13) points out that,

> *"The important thing is that we have a replicable procedure for filtering out cases which might be entirely due to chance. The cases which survive the filters provide a set of words, based on solid quantitative evidence, for further human interpretation."*

Given that the two cut-offs trim more than 50 percent of the non-creative-bearing concgrams while retaining most of those creative-bearing ones, this cut-off calculation and the MI and *t*-score cut-offs produced are therefore adopted. The analysis of these cut-offs in relation to pattern-forming creativity is carried out in the next section.

3. Cut-off Analysis

3.1 Introduction

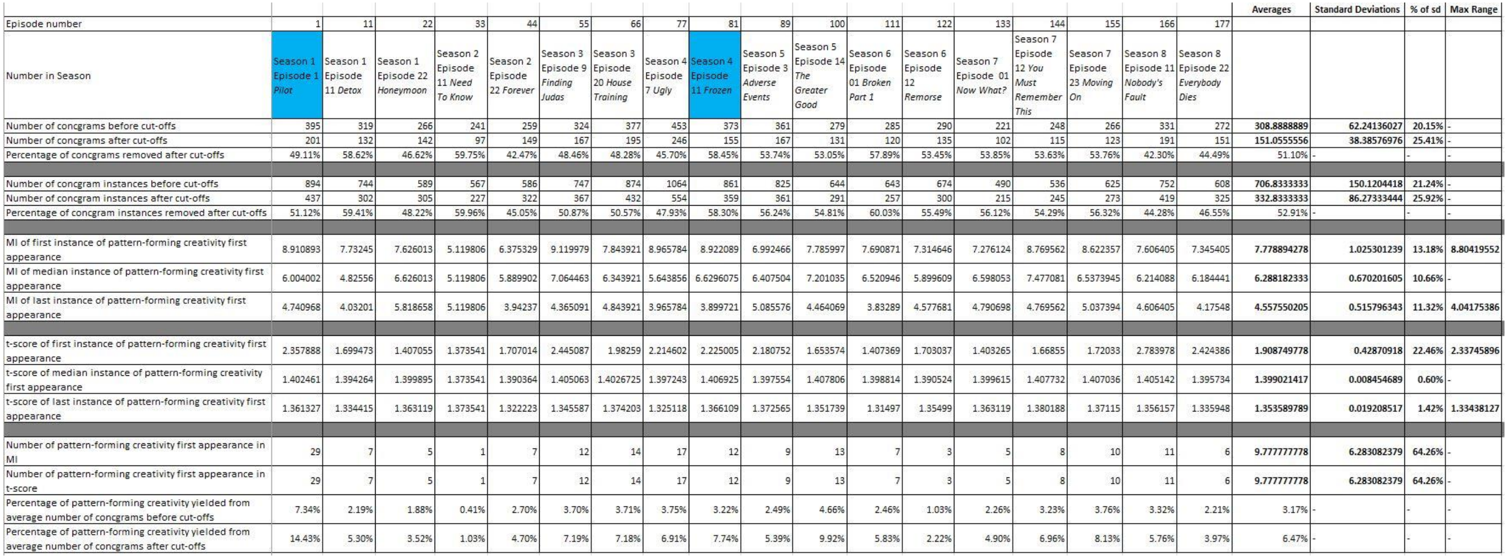
According to McEnery, Xiao and Tono (2006, pp. 56-57) quoting Hunston (2002), an MI value greater or equals to 3 can be considered "as evidence that two items are collocates", while a *t*-score greater or equal to 2 is "normally considered to be statistically significant". However, despite the fact that pattern-forming creativity falls under the consideration of collocations and statistical significance, the MI and *t*-score cut-offs produced in the above section have evidently shown that the MI and *t*-score cut-offs may not be the best options (Cheng, Greaves, & Warren, 2006). In order to provide a clearer and more detailed picture as to how pattern-forming creativity may be governed by MI and *t*-score, a cut-off analysis is carried out in the hope to fill some of the niche of the MI and *t*-score cut-offs by-default which Stubbs (1995) has criticised.

With the aforementioned aim, three Excel sheets with an extension of the table similar to Table 3 are created for this analysis: Excel sheet ‘every 10 episodes’, ‘every 5 episodes’, and ‘every 3 episodes’. These Excel sheets include statistical results from the extraction of pattern-forming creativity performed on concgram lists from the episodes selected specified in the name of the sheets, i.e. Excel sheet ‘every 10 episodes’ selects roughly one episode from every ten episodes, and so on. The extended version of Table 3 includes more statistical requirements as shown in Table 4. Some of the most important additions include the percentage of concgrams and of concgram instances removed after cut-offs are applied (which is 100 percent minus the ratio of the number of concgrams/concgram instances after cut-offs to the number of concgrams/concgram instances before cut-offs), averages and percentages of the lower, median and upper bound (and hence maximum range governed by one standard deviation from the lower and upper limit) of MI and *t*-score from the first appearances of pattern-forming creativity in each episode, their corresponding averages and their standard deviations as well as the numbers of pattern-forming creativity yielded from the number of concgrams before and after MI and *t*-score cut-offs are applied.

__Table 4:__ Extended version of table for cut-off analysis

	Averages	Standard Deviations	% of sd	Max Range
Episode number	╱	╱	╱	╱
Number in Season	╱	╱	╱	╱
Number of concgrams before cut-offs				╱
Number of concgrams after cut-offs				╱
Percentage of concgrams removed after cut-offs		╱	╱	╱
Number of concgram instances before cut-offs				╱
Number of concgram instances after cut-offs				╱
Percentage of concgram instances removed after cut-offs		╱	╱	╱
MI of first instance of pattern-forming creativity first appearance				
MI of median instance of pattern-forming creativity first appearance				╱
MI of last instance of pattern-forming creativity first appearance				
t-score of first instance of pattern-forming creativity first appearance				
t-score of median instance of pattern-forming creativity first appearance				╱
t-score of last instance of pattern-forming creativity first appearance				
Number of pattern-forming creativity first appearance in MI				╱
Number of pattern-forming creativity first appearance in t-score				╱
Percentage of pattern-forming creativity yielded from average number of concgrams before cut-offs		╱	╱	╱
Percentage of pattern-forming creativity yielded from average number of concgrams after cut-offs		╱	╱	╱

3.2 *‘Every 10 episodes’ Analysis*



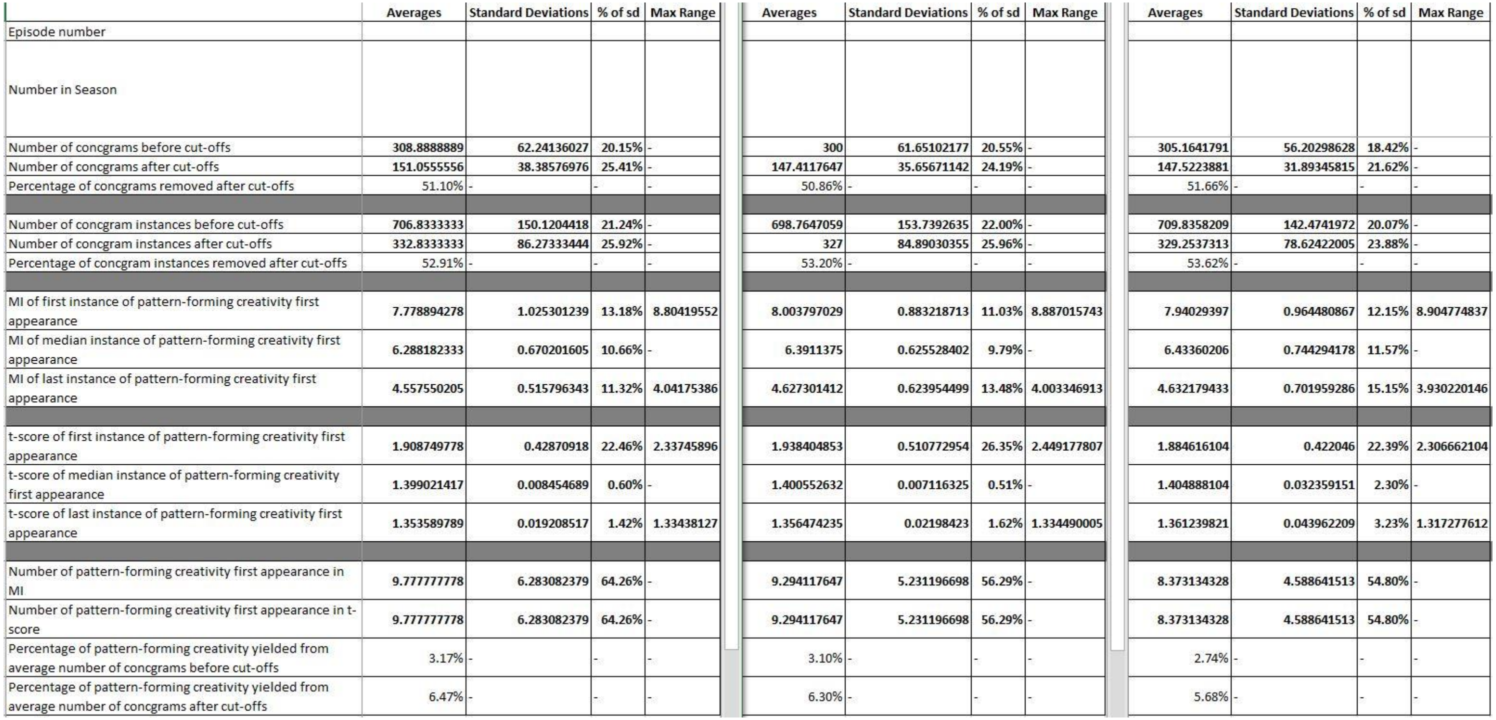
																			Averages	Standard Deviations	% of sd	Max Range
Episode number	1	11	22	33	44	55	66	77	81	89	100	111	122	133	144	155	166	177				
Number in Season	Season 1 Episode 1 *Pilot*	Season 1 Episode 11 *Detox*	Season 1 Episode 22 *Honeymoon*	Season 2 Episode 11 *Need To Know*	Season 2 Episode 22 *Forever*	Season 3 Episode 9 *Finding Judas*	Season 3 Episode 20 *House Training*	Season 4 Episode 7 *Ugly*	Season 4 Episode 11 *Frozen*	Season 5 Episode 3 *Adverse Events*	Season 5 Episode 14 *The Greater Good*	Season 6 Episode 01 *Broken Part 1*	Season 6 Episode 12 *Remorse*	Season 7 Episode 01 *Now What?*	Season 7 Episode 12 *You Must Remember This*	Season 7 Episode 23 *Moving On*	Season 8 Episode 11 *Nobody's Fault*	Season 8 Episode 22 *Everybody Dies*				
Number of concgrams before cut-offs	395	319	266	241	259	324	377	453	373	361	279	285	290	221	248	266	331	272	308.8888889	62.24136027	20.15%	-
Number of concgrams after cut-offs	201	132	142	97	149	167	195	246	155	167	131	120	135	102	115	123	191	151	151.0555556	38.38576976	25.41%	-
Percentage of concgrams removed after cut-offs	49.11%	58.62%	46.62%	59.75%	42.47%	48.46%	48.28%	45.70%	58.45%	53.74%	53.05%	57.89%	53.45%	53.85%	53.63%	53.76%	42.30%	44.49%	51.10%	-	-	-
Number of concgram instances before cut-offs	894	744	589	567	586	747	874	1064	861	825	644	643	674	490	536	625	752	608	706.8333333	150.1204418	21.24%	-
Number of concgram instances after cut-offs	437	302	305	227	322	367	432	554	359	361	291	257	300	215	245	273	419	325	332.8333333	86.27333444	25.92%	-
Percentage of concgram instances removed after cut-offs	51.12%	59.41%	48.22%	59.96%	45.05%	50.87%	50.57%	47.93%	58.30%	56.24%	54.81%	60.03%	55.49%	56.12%	54.29%	56.32%	44.28%	46.55%	52.91%	-	-	-
MI of first instance of pattern-forming creativity first appearance	8.910893	7.73245	7.626013	5.119806	6.375329	9.119979	7.843921	8.965784	8.922089	6.992466	7.785997	7.690871	7.314646	7.276124	8.769562	8.622357	7.606405	7.345405	7.778894278	1.025301239	13.18%	8.80419552
MI of median instance of pattern-forming creativity first appearance	6.004002	4.82556	6.626013	5.119806	5.889902	7.064463	6.343921	5.643856	6.6296075	6.407504	7.201035	6.520946	5.899609	6.598053	7.477081	6.5373945	6.214088	6.184441	6.288182333	0.670201605	10.66%	-
MI of last instance of pattern-forming creativity first appearance	4.740968	4.03201	5.818658	5.119806	3.94237	4.365091	4.843921	3.965784	3.899721	5.085576	4.464069	3.83289	4.577681	4.790698	4.769562	5.037394	4.606405	4.17548	4.557550205	0.515796343	11.32%	4.04175386
t-score of first instance of pattern-forming creativity first appearance	2.357888	1.699473	1.407055	1.373541	1.707014	2.445087	1.98259	2.214602	2.225005	2.180752	1.653574	1.407369	1.703037	1.403265	1.66855	1.72033	2.783978	2.424386	1.908749778	0.42870918	22.46%	2.33745896
t-score of median instance of pattern-forming creativity first appearance	1.402461	1.394264	1.399895	1.373541	1.390364	1.405063	1.4026725	1.397243	1.406925	1.397554	1.407806	1.398814	1.390524	1.399615	1.407732	1.407036	1.405142	1.395734	1.399021417	0.008454689	0.60%	-
t-score of last instance of pattern-forming creativity first appearance	1.361327	1.334415	1.363119	1.373541	1.322223	1.345587	1.374203	1.325118	1.366109	1.372565	1.351739	1.31497	1.35499	1.363119	1.380188	1.37115	1.356157	1.335948	1.353589789	0.019208517	1.42%	1.33438127
Number of pattern-forming creativity first appearance in MI	29	7	5	1	7	12	14	17	12	9	13	7	3	5	8	10	11	6	9.777777778	6.283082379	64.26%	-
Number of pattern-forming creativity first appearance in t-score	29	7	5	1	7	12	14	17	12	9	13	7	3	5	8	10	11	6	9.777777778	6.283082379	64.26%	-
Percentage of pattern-forming creativity yielded from average number of concgrams before cut-offs	7.34%	2.19%	1.88%	0.41%	2.70%	3.70%	3.71%	3.75%	3.22%	2.49%	4.66%	2.46%	1.03%	2.26%	3.23%	3.76%	3.32%	2.21%	3.17%	-	-	-
Percentage of pattern-forming creativity yielded from average number of concgrams after cut-offs	14.43%	5.30%	3.52%	1.03%	4.70%	7.19%	7.18%	6.91%	7.74%	5.39%	9.92%	5.83%	2.22%	4.90%	6.96%	8.13%	5.76%	3.97%	6.47%	-	-	-

Figure 3*:* *Excel sheet 'every 10 episodes'*

For the sampling of episodes, the spread and the inclusion of the cut-off-generating episodes are the only concerns. For example, Figure 3 shows an Excel sheet ‘every 10 episodes’ with Table 4, including episode number 1, 11, 22, 33, 44, 55, 66, 77, 81, 89, 100, 111, 122, 133, 144, 155, 166 and 177, a total of 18 episodes, with episode number 1 and 81 being the two episodes used to calculate the MI cut-off (3.766887) and *t*-score cut-off (1.314885), hence blue-highlighted cells. Using the above selection criteria and cut-offs, it can be seen that a level of consistency has been achieved. First, after cut-offs are applied, the percentage of concgrams and of concgram instances removed in every episode are consistently above 42 percent and 44 percent respectively, giving an overall average of 51.10 percent and 52.91 percent. A huge narrowing of standard deviation in the number of concgrams and concgram instances after cut-offs is also observed, converging from 62.24 to 38.39 and 150.12 to 86.27 respectively. Percentage of pattern-forming creativity yielded from the number of concgrams after cut-offs in each episode has mostly doubled when compared to the percentage yield before cut-offs, helping an overall increase of yield from 3.17 percent to 6.47 percent in the sample. Such numbers support that the use of MI and *t*-score cut-offs have effectively increased the density of pattern-forming creativity in the concgram lists.

The sample also produced interesting results in the first instance of pattern-forming creative concgram analysis. First, standard deviations of MI and *t*-score of the first instance, median instance and last instance of pattern-forming creativity first appearance are not far off from their respective means. The standard deviations of MI of the first instance, median instance and last instance range are 1.025, 0.516 and 0.670 respectively, which correspond to 13.18 percent, 11.32 percent and 10.66 percent of their numerical averages. These standard deviations are around 1.0 in numerical values and around 10 percent, which are not low but are close to one another enough to provide a reasonable range (4.042 – 8.804) at a distance of one standard deviation (lower limit = 4.558 - 0.516, upper limit = 7.779 + 1.025). Whereas the standard deviations of *t*-score of the first instance, median instance and last instance range are 0.429, 0.008 and 0.019 respectively, which correspond to 22.46 percent, 0.60 percent and 1.42 percent of their numerical averages. While the first of the three standard deviations of *t*-score offers a larger percentage difference like that of MI's, it is worth noting that the MI cut-off (3.766887) and *t*-score cut-off (1.314885) have in fact helped produce tighter lower limits, which otherwise could have been wider than they are presented here. Having presented that, the standard deviations of the *t*-score of the median instance and last instance are of considerably low in numerical values and percentages. In summary, considering this sample alone, *t*-score’s maximum range (1.334 – 2.337) would give a more accurate lower limit (= 1.354 - 0.019 = 1.334) and median (= 1.399 ± 0.008) but a larger upper limit (= 1.909 + 0.429 = 2.337) than MI’s maximum range, whereas MI’s maximum range is more consistent across all three standard deviations. A synergy of both MI and *t*-score maximum ranges can further increase the hit rate of concgrams of pattern-forming creativity. Lastly, with an improved yield of the overall pattern-forming creative concgrams from 3.17 percent to 6.47 percent, the cut-offs have not only doubled the effectiveness but also halved the time required to process every single concgram of every episode. Even though the hit-rate of creativity-bearing concgrams is still low, a synergetic application of both MI and *t*-score maximum ranges can be used to narrow the search and increase efficiency even further.

***Figure 4:** Excel sheet 'every 10 episodes' (left), 'every 5 episodes' (middle) and 'every 3 episodes' (right)*



	Averages	Standard Deviations	% of sd	Max Range	Averages	Standard Deviations	% of sd	Max Range	Averages	Standard Deviations	% of sd	Max Range
Episode number												
Number in Season												
Number of concgrams before cut-offs	308.8888889	62.24136027	20.15%	-	300	61.65102177	20.55%	-	305.1641791	56.20298628	18.42%	-
Number of concgrams after cut-offs	151.0555556	38.38576976	25.41%	-	147.4117647	35.65671142	24.19%	-	147.5223881	31.89345815	21.62%	-
Percentage of concgrams removed after cut-offs	51.10%	-	-	-	50.86%	-	-	-	51.66%	-	-	-
Number of concgram instances before cut-offs	706.8333333	150.1204418	21.24%	-	698.7647059	153.7392635	22.00%	-	709.8358209	142.4741972	20.07%	-
Number of concgram instances after cut-offs	332.8333333	86.27333444	25.92%	-	327	84.89030355	25.96%	-	329.2537313	78.62422005	23.88%	-
Percentage of concgram instances removed after cut-offs	52.91%	-	-	-	53.20%	-	-	-	53.62%	-	-	-
MI of first instance of pattern-forming creativity first appearance	7.778894278	1.025301239	13.18%	8.80419552	8.003797029	0.883218713	11.03%	8.887015743	7.94029397	0.964480867	12.15%	8.904774837
MI of median instance of pattern-forming creativity first appearance	6.288182333	0.670201605	10.66%	-	6.3911375	0.625528402	9.79%	-	6.43360206	0.744294178	11.57%	-
MI of last instance of pattern-forming creativity first appearance	4.557550205	0.515796343	11.32%	4.04175386	4.627301412	0.623954499	13.48%	4.003346913	4.632179433	0.701959286	15.15%	3.930220146
t-score of first instance of pattern-forming creativity first appearance	1.908749778	0.42870918	22.46%	2.33745896	1.938404853	0.510772954	26.35%	2.449177807	1.884616104	0.422046	22.39%	2.306662104
t-score of median instance of pattern-forming creativity first appearance	1.399021417	0.008454689	0.60%	-	1.400552632	0.007116325	0.51%	-	1.404888104	0.032359151	2.30%	-
t-score of last instance of pattern-forming creativity first appearance	1.353589789	0.019208517	1.42%	1.33438127	1.356474235	0.02198423	1.62%	1.334490005	1.361239821	0.043962209	3.23%	1.317277612
Number of pattern-forming creativity first appearance in MI	9.777777778	6.283082379	64.26%	-	9.294117647	5.231196698	56.29%	-	8.373134328	4.588641513	54.80%	-
Number of pattern-forming creativity first appearance in t-score	9.777777778	6.283082379	64.26%	-	9.294117647	5.231196698	56.29%	-	8.373134328	4.588641513	54.80%	-
Percentage of pattern-forming creativity yielded from average number of concgrams before cut-offs	3.17%	-	-	-	3.10%	-	-	-	2.74%	-	-	-
Percentage of pattern-forming creativity yielded from average number of concgrams after cut-offs	6.47%	-	-	-	6.30%	-	-	-	5.68%	-	-	-

3.2 ‘Every 10, 5, 3 episodes’ Analysis

Figure 4 shows a screen capture of the final four columns of the Excel sheets ‘every 10 episodes’ (left), ‘every 5 episodes’ (middle) and ‘every 3 episodes’ (right). Excel sheet ‘every 10 episodes’, ‘every 5 episodes’ and ‘every 3 episodes’ consists of 18, 34 and 67 episodes respectively. When comparing all three Excel sheets side-by-side, trends become more apparent.

Firstly, the percentage of concgrams and of concgrams instances removed after cut-offs remain relatively constant around 50 percent across all three Excel sheets regardless of the number of episodes included, which shows that the MI (3.766887) and *t*-score (1.314885) cut-offs are able to provide a consistent level of trimming despite the fact that each episode produces different number of concgrams and concgram instances. This provide a good evidence to support the effectiveness of calculating a custom MI and *t*-score cut-offs from a small sample of a specific data set rather than using the commonly accepted MI (= 3.0) and *t*-score (= 2.0) cut-offs.

Secondly, MI of the first, median and last instance of pattern-forming creativity first appearance also maintained consistency in numbers and percentages as the number of episodes increased. The first began with 7.779 in ‘every 10 episodes’ to 8.004 in ‘every 5 episodes’ to 7.940 in ‘every 3 episodes’, all within a range of 0.225. The respective standard deviations (and percentages in brackets) are 1.025 (13.18 percent), 0.883 (11.03 percent) and 0.964 (12.15 percent), all within a range of 0.142. The median has an average value that changes from 6.288 in ‘every 10 episodes’ to 6.391 in ‘every 5 episodes’ to 6.434 in ‘every 3 episodes’, spanning a range of 0.146. The respective standard deviations (and percentages in brackets) are 0.670 (10.66 percent), 0.626 (9.79 percent) and 0.744 (11.57 percent), spanning a difference of 0.118. The last sees a slight increase from 4.558 in ‘every 10 episodes’ to 4.627 in ‘every 5 episodes’ to 4.632 in ‘every 3 episodes’, spanning a range of 0.074. The respective standard deviations (and percentages in brackets) are 0.516 (11.32 percent), 0.624 (13.48 percent) and 0.702 (15.15 percent), spanning a range of 0.186. Maximum range governed by one standard deviation from the lower and upper limit widens gradually as the number of episodes accounted for almost quadrupled from 18 episodes to 67 episodes, that is from 4.042—8.804 in ‘every 10 episodes’ to 4.003 – 8.887 in ‘every 5 episodes’ to 3.930 – 8.905 in ‘every 3 episodes’, representing a widening of 0.112 (= 4.042 - 3.930) at the lower limit and 0.101 (= 8.905 - 8.804) at the upper limit that equates to 0.213 (= 0.112 + 0.101) or 4.28 percent (= (8.905 - 3.930) / 0.213) of the maximum range in ‘every 3 episodes’. This relatively minor widening (<5 percent) provides evidence that most concgrams of pattern-forming creativity first appearance in *House M.D.* could be found within the maximum range of MI, given that the calculated MI and *t*-score cut-offs are used. The maximum range also has about the same percentage of standard deviations at its lower, median and upper limit, which shows the stability of MI maximum range.

Thirdly, *t*-score of the first, median and last instance of pattern-forming creativity first appearance see a great fluctuation in numbers and percentages as the number of episodes increased. The first began with 1.909 in ‘every 10 episodes’ to 1.938 in ‘every 5 episodes’ to 1.885 in ‘every 3 episodes’, all within a range of 0.024. The respective standard deviations (and percentages in brackets) are 0.429 (22.46 percent), 0.511 (26.35 percent) and 0.422 (22.39

percent), all within a range of 0.089. The median has an average value that changes from 1.399 in ‘every 10 episodes’ to 1.401 in ‘every 5 episodes’ to 1.405 in ‘every 3 episodes’, spanning a range of 0.006. The respective standard deviations (and percentages in brackets) are 0.008 (0.60 percent), 0.007 (0.51 percent) and 0.032 (2.30 percent), spanning a difference of 0.025. The last sees a slight increase from 1.354 in ‘every 10 episodes’ to 1.356 in ‘every 5 episodes’ to 1.361 in ‘every 3 episodes’, spanning a range of 0.007. The respective standard deviations (and percentages in brackets) are 0.019 (1.42 percent), 0.022 (1.62 percent) and 0.044 (3.23 percent), spanning a range of 0.025. Maximum range governed by one standard deviation from the lower and upper limit widens gradually as the number of episodes accounted for almost quadrupled from 18 episodes to 67 episodes, that is from 1.334—2.337 in ‘every 10 episodes’ to 1.334 – 2.449 in ‘every 5 episodes’ to 1.317 – 2.307 in ‘every 3 episodes’, representing a widening of 0.017 (= 1.334 - 1.317) at the lower limit and a narrowing of -0.142 (= 2.307 - 2.449) at the upper limit that equates to -0.125 (= 0.017 - 0.142) or -7.92 percent (= (2.307 - 1.317) / -0.125) of the maximum range in ‘every 3 episodes’. This relatively significant narrowing contributed mainly by the narrowing at the upper limit indicates that using *t*-score to locate concgrams of pattern-forming creativity first appearance in *House M.D.* at the upper end may not be desirable, given the high percentage of standard deviation and the rather significant fluctuation at the upper limit (>5 percent). However, using *t*-score at the lower end and at the median have statistically shown to be reliable (<5 percent), given that the calculated MI and *t*-score cut-offs are used. Analysis of trends over 3 Excel sheets confirms that a synergetic application of both MI and *t*-score maximum ranges can be used to narrow the search for congrams of pattern-forming creativity first appearance and increase efficiency.

Fourthly, standard deviation of various numbers such as number of concgrams and concgram instances before and after cut-offs as well as the number of pattern-forming creativity first appearance (in both MI and *t*-score) have seen a general downtrend from ‘every 10 episodes’ to ‘every 5 episodes’ to ‘every 5 episodes’ to ‘every 3 episodes’ as the total number of episodes considered increases from 18 to 34 to 67. This implies that the data has become gradually less dispersed and is likely to continue if all episodes are considered.

Lastly, the percentage of pattern-forming creativity yielded from average number of concgrams before cut-offs and after cut-offs, which is calculated using the number of patter-forming creativity first appearance in MI/*t*-score divided by number of concgrams before cut-offs and after cut-offs respectively, saw their highest at 3.17 percent and 6.47 percent in ‘every 10 episodes’ and lowest at 2.74 percent and 5.68 percent in ‘every 3 episodes’ respectively. Such slight decrease in percentages is contributed mainly by the fall of 1.405 (= 9.778 – 8.373) in the numerator, a relatively significant value compared to the minor decrease in the large denominators (from 308.889 to 305.164 for number of concgrams before cut-offs and 151.056 to 147.522 for number of concgrams after cut-offs). Overall, judging by the percentage of pattern-forming creativity yielded from average number of concgrams after cut-offs from each episode in the ‘every 3 episodes’ Excel sheet, only 5 of the 67 episodes managed to reach more than 10 percent. Therefore, even when all episodes are considered, it is expected that the average yield to remain no higher than 10 percent using the calculated MI and *t*-score cut-offs alone. However, should maximum range be used in the cut-off process, the number of concgrams after cut-offs can be reduced further and possibly increase the yield of pattern-forming creativity.

4. Conclusion

This paper has proposed and demonstrated several steps to improve efficiency in the process of identification and extraction of pattern-forming creativity from fan scripts of television drama *House M.D.*. It has included a discussion on the calculation of internal span which helps to limit the number of words included in the concgram search and consequently the number of concgrams generated. The paper has extended Carter's (2004) definition of pattern-forming creativity to include creative patterns produced in a non-co-constructed, self-repeated manner and provided examples for each subcategories of pattern-forming creativity proposed.

The calculation of custom *t*-score and MI value cut-offs has also been discussed and a detail cut-off analysis is presented. After manual extraction of pattern-forming creativity has been performed on 67 episodes of *House M.D.*, it is found through the cut-off analysis that the use of MI and *t*-score cut-offs has effectively doubled the percentage yield of pattern-forming creativity in the concgram lists. Statistical figures obtained from 3 separate concgram list analysis consisting of 18, 34 and 67 episodes have been compared and results have shown consistency in the percentage yield of pattern-forming creativity through the use of custom *t*-score and MI cut-offs. Analysis has also shown that *t*-score and MI maximum range will likely improve efficiency further while retaining a reasonable hit rate in the extraction of pattern-forming creativity when used iteratively.

Above all, it is hope that this paper can draw more researchers' attention to creativity studies and have provided certain clues to enhancing automated extraction of pattern-forming creativity in computational creativity.

While these statistical elements and tools are powerful devices which will help in the reduction of time cost in the linguistic creativity extraction, one must acknowledge that none of the statistical devices are perfect in their design and suitable for all linguistic situations. Stubbs (1995) points out that, "[a] result may not reach "significance", as defined by such a test, due to a bias or to natural variability in the data: and it is obvious to corpus linguists that language is highly variable.” The fact that *t*-score is a more suitable test for lexical items than it is for grammatical ones (Stubbs, 1995), or that MI tends to suffer from overestimation in extreme cases of collocations (Gries & Stefanowitsch, 2004), or even the presumption that association measures (AMs) such as Mutual Information (MI) and *t*-score are symmetric / bidirectional in nature (Gries, 2015), are some examples of the limitations of their statistical devices and a reflection of English as a highly variable language. Therefore, in short, the results are as good as the corpus itself. Any results obtained by these statistical devices are limited to the dataset of *House M.D.*. They should not and cannot be compared to results obtained using another corpus of TV drama or a combination of several ones.

5. References

ABC Medianet. (2004, November 23). *ABC Medianet*. Retrieved November 26, 2017, from https://web.archive.org/web/20081221172650/http://abcmedianet.com/web/dnr/dispDNR.aspx?id=112304_07

Barnbrook, G. (1996). *Language and Computers: A Practical Introduction to the Computer Analysis of Language.* Edinburgh: Edinburgh University Press.

Bednarek, M. (2010). *The Language of Fictional Television: Drama and Identity.* London: Continuum International Publishing Group.

Bignell, J., & Lacey, S. (2005). *Popular television drama : critical perspectives.* (J. Bignell, & S. Lacey, Eds.) Manchester: Manchester University Press.

Bordwell, D., & Thompson, K. ([1990] 2008). *Film Art: An Introduction.* New York: McGraw-Hill.

Carter, R. (2004). *Language and Creativity: The Art of Common Talk.* London: Routledge.

Carter, R. (2016, Aug 23). *Creativity in everyday language*. Retrieved January 5, 2018, from OpenLearn - Open

University: http://www.open.edu/openlearn/languages/language-and-creativity/content-section-2.2

Cascio, T., & Martin, L. L. (Eds.). (2011). *House and Psychology: Humanity Is Overrated.* Wiley.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics, 11*(4), 411-433.

Church, K., & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics, 16*(1), 22-29.

clinic_duty. (2007, October 2). *House MD - 4.01 Alone - House Transcripts*. Retrieved November 28, 2017, from https://clinic-duty.livejournal.com/21422.html

Clyman, J. (2009, June 22). *Inside the Therapy TV Show You Need to Watch*. Retrieved April 20, 2013, from Psychology Today: http://www.psychologytoday.com/blog/reel-therapy/200906/inside-the-therapy-tv-show-you-need-watch

Cross, J., & Papp, S. (2008). Creativity in the use of verb + noun combinations by Chinese learners of English. In G. Gilquin,

S. Papp, & M. Belén Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 57-84). BRILL.

Furiassi, C., & Hofland, K. (2007). The retrieval of false anglicisms in newspaper texts. In R. Facchinetti (Ed.), *Corpus Linguistics 25 Years on* (pp. 348-364). Amsterdam - New York: Rodopi.

Goodier, B. C., & Arrington, M. I. (2007). Physicians, patients, and medical dialogue in the NYPD Blue prostate cancer story. *Journal of Medical Humanities, 28*(1), 45-58.

Greaves, C. (2009, February 5). ConcGram 1.0: A phraseological search engine – user manual. Hong Kong: The Hong Kong Polytechnic University. Retrieved from http://www.benjamins.com/jbp/series/CLS/1/manual.pdf (also onConcGram CD)

Gries, S. T. (2015). Some Current Quantitative Problems in Corpus Linguistics and a Sketch of Some Solutions. *Language and Linguistics, 16*(1), 93-117.

Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on ‘alternations’*. *International Journal of Corpus Linguistics*, 97-129.

Guinness World Record News. (2012, May 22). *Record-Holding TV Show 'House' Comes to an End*. Retrieved November 23, 2015, from Guinness World Records: http://www.guinnessworldrecords.com/news/2012/5/record-holding-tv-show-house-comes-to-an-end-42046/

Hockley, L., & Gardner, L. (Eds.). (2011). *House: The Wounded Healer on Television: Jungian and Post-Jungian Reflections.* Hove: Routledge.

Holtz, A. (2006). *The Medical Science of House M.D.* . New York: Berkley Boulevard.

Holtz, A. (2011). *House M.D. vs. Reality: Fact and Fiction in the Hit Television Series.* Berkley Trade.

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Jackman, I., & Laurie, H. (2010). *House, M.D.: The Official Guide to the Hit Medical Drama.* It Books.

Jacoby, H., & Irwin, W. (Eds.). (2008). *House and Philosophy: Everybody Lies.* John Wiley & Sons.

Jamieson, D. (2011, September). *Does TV accurately portray psychology?* Retrieved April 20, 2013, from American

Psychology Association: http://www.apa.org/gradpsych/2011/09/psychology-shows.aspx

Jordanous, A. (2010). Defining Creativity: Finding Keywords for Creativity Using Corpus Linguistics Techniques. *First International Conference on Computational Creativity (ICCCX)*, (pp. 278-287). Lisbon, Portugal.

Li, Q., & Csikszentmihalyi, M. (2014). Moral Creativity and Creative Morality. In S. Moran, D. Cropley, & J. Kaufman (Eds.), *The Ethics of Creativity* (pp. 75-91). New York: Palgrave Macmillan.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book.* New York: Routledge.

Olson, C. (2010, April 30). *Digital Convergence Episode 15: Gale Tattersall – Cinematographer, Director of Photography for FOX's House MD*. Retrieved May 3, 2013, from Digital Film.tv: http://digitalfilm.tv/digital-convergence-episode-15-gale-tattersall-cinematographer-director-of-photography-for-foxs-house-md

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

Sanders, L. (2009). *Every Patient Tells a Story: Medical Mysteries and the Art of Diagnosis.* Kindle.

Seidman, R. (2008, February 5). *Broadcast Nielsen Ratings w/e Feb 3: Fox Breaks Records*. Retrieved from TV by the Numbers Zap2it.com: http://tvbythenumbers.zap2it.com/2008/02/05/broadcast-nielsen-ratings-we-feb-3fox-breaks-records/2559/

Slate, L. (2006, April 17). *Hugh Laurie and Cast Make a House Call*. Retrieved July 7, 2014, from Academy of Television Arts & Sciences: http://web.archive.org/web/20131227083156/http://www.emmys.tv/events/2009/hugh-laurie-and-cast-make-house-call

Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative methods. *Functions of Language, 2*(1), 23-55.

Tannen, D. ([1989] 2007). *Talking Voices: Repetition, Dialogue and Imagery in Conversational Discourse.* Cambridge: Cambridge University Press.

TV By The Numbers. (2012, May 22). *Monday Final Ratings: 'DWTS', 'AGT', and 'House' Retrospective Adjusted Up; 'Clash of the Commercials' Adjusted Down*. Retrieved November 26, 2017, from TV By The Numbers by zap2it.com: http://tvbythenumbers.zap2it.com/sdsdskdh279882992z1/monday-final-ratings-dwts-agt-and-house-retrospective-adjusted-up-clash-of-the-commercials-adjusted-down/135303/

Vo, T. A., & Carter, R. (2010). What can a corpus tell us about creativity? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 302-315). Abingdon: Routledge.

Werts, D. (2009, January 29). *Fox's medical marvel stays on top*. Retrieved July 8, 2014, from Variety: http://variety.com/2009/scene/news/fox-s-medical-marvel-stays-on-top-1117999278/

Whitbourne, S. K. (2012, March 10). *We're Not Psychologists but We Play Them on Television*. Retrieved April 20, 2013, from Psychology Today:

http://www.psychologytoday.com/blog/fulfillment-any-age/201203/we-re-not-psychologists-we-play-them-television