Prototypical characteristics of blockbuster movie dialogue: a corpus stylistic analysis

Dan McIntyre, University of Huddersfield, UK

1. Introduction

As a cultural artefact, the blockbuster movie is often accorded the same literary value as airport fiction. As Stringer ("Introduction" 1) notes, 'Films labeled as blockbusters are frequently positioned as examples of the culturally retrograde, beneath serious consideration or analysis'. While this may be the mainstream view, it is also the case that academic interest in the blockbuster appears to be on the rise. Stringer's own edited volume, *Movie Blockbusters*, is among recent publications on the topic, alongside work by, for example, Buckland, King ("Spectacular Narrative", "New Hollywood") and Lavik. What characterises much work in this area is an acknowledgement that the blockbuster as a cultural phenomenon is an amorphous construct, lacking a single definition. Hills, for instance, argues that the blockbuster movie is 'an extra-textual, discursive construction. Texts do *not* present definitive attributes that can allow them to be classified as blockbusters, as if blockbuster status were akin to a textually identifiable film genre' (179). Stringer ("Introduction") concurs, noting

also that 'the movie blockbuster is a multifaceted phenomenon whose meanings are contingent upon the presence of a range of discourses both internal and external to Hollywood' (2). Nonetheless, Stringer does claim genre status for the blockbuster, suggesting implicitly that there must be some identifiable constitutive features of such films. While there appears to be some degree of disagreement here, I would argue that this arises from an implicit approach to classification that is Aristotelian in nature, and that taking a view that is informed by prototype theory offers a way around this seeming impasse. I will elaborate on this below.

My aim in this article is thus to contribute to our understanding of what the blockbuster movie actually is by considering character dialogue, an aspect of film that is often neglected by film theorists. To do this, I analyse a corpus of around 300,000 words of blockbuster movie screenplays, using techniques drawn from corpus linguistics (I will deal with the apparent circularity of having to define the blockbuster movie in order to study it below). I suggest that while a full understanding of the blockbuster must of necessity take into account the extract-textual aspects alluded to by Hills, there do appear to be some dialogic aspects of blockbuster screenplays that seem to be indicative of genre features. In this respect, corpus linguistics can offer quantitative support to the

arguments about blockbusters that literary and film theorists may want to make.

2. Prototypes, language and dialogue

As mentioned above, there is an apparent inconsistency to the views expressed by Hills and Stringer ("Introduction"), with the former claiming no definitive attributes for blockbusters and the latter arguing for viewing the blockbuster as a genre in its own right. I agree with Hills that classifying a film as a blockbuster is not a simple matter of identifying the presence of a number of constitutive features, though I would not go so far as to claim that blockbusters exhibit no component characteristics. Clearly there is something about such films beyond their capacity for revenue generation that allows critics to agree that, say, Spielberg's *Raiders of the Lost Ark* is a blockbuster while Nanni Moretti's *Caro Diario* is not. If this were not the case, critical discussion of blockbusters would be all but impossible. How, then, do we reconcile the views of Hills and Stringer ("Introduction")? One possible solution (the one I adopt here) is to approach the issue from the perspective of prototype theory, Rosch's influential cognitive approach to classification. According to Rosch, for any given category there are central examples (particularly good examples of members of that category), secondary examples (less good examples) and

peripheral examples (generally not very good examples). The classic exemplar of the theory concerns the category of birds. While, for British people, robins and blackbirds constitute central examples of birds (since we are likely to identify the ability to fly and the possession of feathers as characteristics central to defining a creature as a bird), ostriches and emus are secondary examples (they have feathers but are unable to fly). Still more removed, penguins belong in the peripheral set of examples, being unable to fly but able to swim. (What should also be apparent is that prototypes are to a large extent culturally defined). In this respect, categories are not discrete entities but are best described as having 'fuzzy' edges; one person's central example may be another's secondary example.

The concept of the blockbuster movie, I would argue, works along similar lines. The issue, then, is not in defining for all time what the constitutive features of blockbusters are, but in identifying some of the central components of the category, acknowledging that these may well shift from film to film and from viewer to viewer. With this in mind, what I want to suggest is that blockbusters exhibit some linguistic features which can work as characteristics of this type of movie to a central, secondary or peripheral extent.

My concern with language is not one that is normally shared by film theorists or literary critics who work on film. Film, of course, is a hugely varied medium and while there are film critics whose interests are broadly narratological (see, for example, Bordwell "Narration", "Making Meaning", and Buckland), Film Studies as a subject focuses also on the analysis of *mise-en-scene*, the technical aspects of film production, audience reception and genre. Perhaps the reason why screenplays and dialogue are not usually taken account of by film critics is that, for them, the finished film constitutes the object of study. The screenplay, by contrast, is simply a guiding template and character dialogue just a means of conveying the narrative and thematic issues of the film as a whole. For the linguist though, the text is primary. Where film critics do engage with linguistic-related matters, they tend to approach them from a top-down perspective (see, for example, Cohan's analysis of narrative in *Basic Instinct*), discussing general issues rather than the minutiae of language in the screenplay itself (see McIntyre "Integrating" for an elaboration of this point). Undoubtedly this also comes down to the fact that film critics tend not to have the linguistic expertise necessary for the systematic analysis of language, just as linguists often lack the necessary skills for the analysis of the multimodal aspects of film (see McIntyre "Integrating" for an attempt to integrate

the two, though I stress that I claim no great expertise in the analysis of the non-linguistic aspects of film; the article is an attempt to raise some of the issues that need to be dealt with in order to integrate the analysis of dialogue and *mise-en-scene*). Nonetheless, there has been a recent rise in interest in the linguistic aspects of film (particularly film dialogue) evidenced by the work of Kozloff and the specifically linguistic work of Culpeper and McIntyre, McIntyre ("Dialogue"), Piazza ("Voice-over" and "Let Cinema"), Piazza, Rossi, and Bednarek, and Richardson.

What I suggest in this article is that the insights gained from corpus linguistic analysis of film dialogue can be of value in validating (or, indeed, invalidating) the qualitative (and in some cases, subjective) analyses of film and literary critics. Furthermore, such corpus linguistic work can add value to our understanding of film (and, in this article, the blockbuster particularly) by adding an extra analytical dimension that might then be integrated with the non-linguistic analyses of film critics.

3. Methodology

In order to determine the linguistic elements common to blockbuster movie dialogue I analyse a corpus of thirteen screenplays. In total, this amounts to 320,499 words (it is

perhaps surprising how short most screenplays are). While it would be preferable to have a larger data set, the pre-processing issues in preparing texts for corpus analysis are labour intensive to a degree that makes this difficult to achieve, as I will detail below. Before this, however, it is necessary to discuss briefly the problem of the apparent circularity inherent in compiling a corpus of blockbuster screenplays: how are we to know what to include if a definition of the blockbuster is so elusive?

What is clear from the literature on blockbuster movies is that, even if critics disagree about the extent to which it is possible to identify defining features of the blockbuster, the common consensus is that blockbuster status is not a purely textual phenomenon (note Hills' 2003 argument that the blockbuster is 'an extra-textual, discursive construction'). One of the extra-textual elements that goes a long way towards identifying a film as a blockbuster is its capacity for massive revenue generation. In these most basic terms, blockbusters – or 'tall revenue features' as Hall euphemistically describes them – are simply films which are 'extraordinarily successful in financial terms' (11). Hall also makes the valid point that 'the term can also be extended to refer to those films which need to be this successful in order to have a chance of returning a profit on their equally extraordinary production costs' (11).

This, then, was the primary criterion I used in constructing the corpus. The other was to choose films spanning the four decades from the 1970s to the middle of the twenty-first century. Most of the films in the corpus are consequently the highest grossing movies of their year of release. Where it was not possible to obtain a useable electronic version of the required script, I replaced this with an alternative that had also been a clear financial success.

The earliest screenplay in the corpus is *Jaws* (1975), widely regarded by critics as the film which ushered in the era of the modern blockbuster and which ultimately gave rise to the concept of the 'New Hollywood', wherein the profit potential of hit films was maximised (Schatz). The full list of screenplays is as follows:

1. *Jaws* (1975)

2. *Rocky* (1976)

3. *Star Wars* (1977)

4. *Alien* (1979)

5. *Raiders of the Lost Ark* (1981)

6. *Ghostbusters* (1984)

7. *Indiana Jones and the Last Crusade* (1989)

8. *Basic Instinct* (1992)

9. *Jurassic Park* (1993)

10. *Titanic* (1997)

11. *Armageddon* (1998)

12. *Mission Impossible II* (2000)

13. *Fantastic Four* (2005)

With the exception of *Fantastic Four*, all of these films (or their sequels or prequels) are referred to in Stringer's edited collection *Movie Blockbusters*, which suggests that a selection policy based purely on box office receipts results in a strong degree of consensus for classifying these films as blockbusters.

Once the screenplays had been selected, the next stage was to prepare them for corpus analysis. This entailed substantial pre-processing in order to facilitate a number of different analytical methods. Each screenplay was tagged in order to distinguish dialogue from screen directions, and in the case of dialogue, tags were used to identify the speaking character and their gender. I used a mark-up system developed in a small-scale pilot project (see McIntyre and Walker), which is reproduced in table 1:

Feature	Tags (elements and attributes)	Example
Screen	<sdir> screen	<sdir> INT. MISSION

directions	direction </sdir>	CONTROL – INSIDE THE GLASS–ENCASED ROOM Flip enters the room. Skip writes notes.... </sdir>
Character identification and gender	<char id="Name of character" gen="M or F"> dialogue </char>	<char id="Temple" gen="M"> I'm going to brief the President. What's going on here, Dan? Why didn't we have warning? </char>
Screen directions embedded in character dialogue	<char id="Name of character" gen="M or F"> <sdir> screen direction </sdir> dialogue </char>	<char id="Temple"> <sdir> (V.O.) </sdir> Can you go secure? </char>

Table 1 Mark-up system used for pre-processing

Using mark-up within angle brackets meant that it was then possible to extract all the female speech from the corpus, all the male speech and all the screen directions, in order to study them separately. The inclusion of the speaking character's name within the 'id' attribute also makes it possible to extract the

speech of single characters in order to compare their dialogue against that of other characters. I used Scott Piao's *Multilingual Corpus Toolkit*[1] software to do this, extracting the male speech, female speech and screen directions into three separate files. I also created a file composed of all the male and female dialogue together. While I concentrate in this article on dialogue, the extraction of the screen directions also allows for the separate study of these at a later date.

Once the corpus was prepared, I used the software packages WordSmith Tools (Scott "WordSmith") to calculate word frequencies, type/token ratios, n-grams and dispersal plots, and Wmatrix (Rayson "Wmatrix") to calculate key words and key semantic domains. All of these terms will be explained as and when necessary in the course of the analysis in section 4.

4. Analysis

Numerous claims have been made about the representation of gender in Hollywood blockbusters. Tasker, for instance, comments on the dominance of the white male hero in action blockbusters, while Langford (247) notes that a particular strain of blockbuster develops a 'parodically masculine action vernacular'. Langford (235) further notes the '[o]pposition to

---

[1] Freely available from Scott Piao's website: https://sites.google.com/site/scottpiaosite/software/mlct

authority' that is often displayed by the male hero and Neale
claims that '[d]isplays of the male body and of the hero's
physical prowess are traditional in all kinds of adventure
films' (75). Despite the difficulty of defining the blockbuster
movie, there is, it seems, a surprising degree of consensus as
to what it to be found in such films. In the case of the above
claims, we might reasonably assume a degree of difference in the
way that male and female characters are represented in
blockbuster movies. This, of course, is an area that could be
researched from a number of different angles; here I focus
primarily on what the linguistic aspects of characterisation are
for male and female characters. A useful place to start is with
some simple frequency information.

4.1 Frequency information

In an earlier publication exploring dialogue in film drama
(McIntyre and Walker), I and a colleague explored the issue of
gender specifically in action blockbusters. However, the corpus
we used was much smaller than the one constructed for this study
and focused particularly on action films. Furthermore, since the
paper was for a handbook as opposed to a journal, its purpose
was primarily didactic and pedagogical, leaving little space for
detailed discussion of our findings. In this article, then, my
aim is to develop the hypotheses proposed in that earlier

publication. The primary hypothesis there was that the difference in the way that male and female characters are represented in blockbusters will be reflected linguistically. One straightforward way in to exploring this issue in greater depth is through a simple count of the amount that male and female characters say. Using WordSmith Tools, the frequency information detailed in table 2 was extracted for both the male and female speech in the corpus:

	Male speech	Female speech
Tokens	85,081	17,770
Types	9220	3188
Standardised type-token ratio	43.78	42.74

Table 2 Frequency information for male and female speech

The corpus as a whole consists of 320,499 words. 85,081 words occur in the dialogue by male characters and 17,770 in the female dialogue. The remaining 217,648 words occur in screen directions. In table 2, 'tokens' refers to the total number of words found in the two sub-corpora, while 'types' refers to the number of different words (for example, in the sentence 'The cat

sat on the mat', there are six tokens and five types). These initial figures show such a clear and obvious difference between the amount of speech by male and female characters that a test for statistical significance is unnecessary. The type-token ratio (TTR) provides, to some degree, a measure of lexical richness, and is calculated by dividing the number of tokens by the number of types. The standardised type-token ratio referred to in table 2 is generated by calculating the TTR every 1000 words across the course of the two sub-corpora. What is apparent from this result is that while the amount spoken by male and female characters differs greatly, there is no significant difference in the vocabulary richness of each of these two types of character. Nonetheless, there is perhaps some indicative support here for at least one element of Tasker's claim that blockbusters are dominated by the white male hero; certainly it seems that in simple quantitative terms, male characters dominate in terms of amount of speech.

This frequency count hints at an imbalance in the way that male and female characters are represented in blockbusters but it is a relatively crude indicator of this. In addition to counting how much characters say, we also need to know what aspects of their dialogue are particularly significant and what they talk about. These are issues that I will consider in the next section.

4.2 Keyness

A common analytical tool within corpus linguistics is the keyness measure. This is achieved by comparing the target text or corpus against a larger reference corpus. By using a statistical test it is then possible to determine which items (words, parts-of-speech, semantic fields, etc) are over-represented in the target corpus in comparison with the reference corpus. For example, if a one-million-word corpus of political discourse is found to contain the word 'choice' ten times, then in a smaller 100,000-word sample of the same kind of data, statistically we would expect to see the word ten times less – i.e. once in 100,000 words. If, in fact, the 100,000-word corpus contains ten instances of 'choice' then it would seem that 'choice' is over-represented in the data when compared to the norm, and the analyst would need to find some explanation for this. Keyness, then, is a useful tool for determining what is idiosyncratic about a particular corpus.

An issue for any analysis of keyness is the choice of reference corpus. Xiao and McEnery have shown how it is possible to obtain almost identical keyword lists from very differently sized reference corpora. While this might appear to suggest that the choice of reference corpus is of minimal importance, this is not strictly true. Culpeper makes the salient point that 'The

closer the relationship between the target corpus and the reference corpus, the more likely the resultant keywords will reflect something specific to the target corpus'. In this respect, an ideal reference corpus for the target blockbuster movie corpus would be a larger corpus of screenplays representative of all film genres, marked-up in the same way as the target corpus. In practical terms, however, this is not possible. Without assistance it is extremely time-consuming to create such corpora, and it is usually prohibitively expensive to acquire such help.

Scott (2009), however, proposes at least a way around the issue. Scott ("In Search") reports the results of testing a number of different reference corpora on two target texts from the BNC (British National Corpus) to determine the extent to which different reference corpora affect keyword lists for the target corpus. The main findings from his experiment are: (i) when using a reference corpus made of non-domain-specific language, the larger the reference corpus the better; (ii) even an apparently inappropriately constructed reference corpus will allow the identification of keywords that indicate the target text's so-called 'aboutness'; and (iii) genre-specific reference corpora lead to the generation of different kinds of keywords. Point three is in accordance with Culpeper's point about choosing appropriate reference corpora, while point two relates

particularly to the aim of identifying the thematic elements of a target text. Point 1 is of most interest here. Since it was not possible to create the ideal reference corpus for my purposes, I chose the most appropriate alternative bearing in mind Scott's advice concerning size. Wmatrix, the software package that I used for calculating keyness, offers a number of different reference corpora. In some respects, the BNC Written Imaginative Sampler (222,541 words of fiction) is the most appropriate in terms of its comprising of fiction (albeit narrative as opposed to dramatic). On the other hand, the BNC Spoken Demographic sampler of 501,953 words of spoken English is larger. Additionally, it is appropriate in the sense that dramatic dialogue is, in essence, a fictionalised version of real-life dialogue (Mandala). This, then, is the reference corpus that I used.

    To begin, let us consider the positive keywords generated by Wmatrix for both the male and female dialogue (that is; those words which are statistically over-represented when compared against the norm constituted by the reference corpus). Wmatrix calculates keyness using the log-likelihood statistical test. This gives each word a log-likelihood score. Above the critical value of 15.13, we can be 99.99% confident that the words in the keyword list are indeed statistically significant and not over-

represented as a result of chance alone. Tables 3 and 4 detail

the first 20 keywords for the male and female dialogue:

Keyword	LL score (p < 0.0001; critical value = 15.13)
ya	446.72
ship	307.71
the	307.11
an'	270.69
sir	262.34
Luke	255.85
this	245.95
God	228.46
Mr.	220.72
cockpit	220.72
all_right	178.03
here	173.28
us	172.88
Reed	154.89
guys	153.73
me	144.44

Dr.	143.27
world	136.81
of	134.77
Rose	134.06

Table 3 First 20 keywords in the male dialogue

Keyword	LL score (p < 0.0001; critical value = 15.13)
Nick	182.18
me	177.54
Reed	168.69
Mr.	161.94
you	131.22
Jack	126.07
ship	117.35
Victor	107.96
God	101.21
Cal	94.46
Dr.	80.97
indy	74.22

all_right	68.26
Ash	67.47
Johnny	61.02
Rose	59.88
kill	55.79
help	54.54
kane	53.98
Paulie	53.98

Table 4 First 20 keywords in the female dialogue

One issue with interpreting lists of keywords is that, due to the large amount of data from which they are drawn, they tend to be too long for it to be possible to analyse each word individually. What is needed, then, is some way of reducing the keyword list to a manageable amount. One crude way to do this is to choose an arbitrary cut-off point; in this case, I have taken the first twenty from each sub-corpus. However, even this leaves a large amount of data to deal with. A second means of reducing the list, then, is to eliminate those items which are specific to one particular text, since these are indicative of features to do with that text in particular as opposed to the corpus in general. For example, in the list of keywords from the male dialogue (henceforth, male keywords), *ya* and *'an* are abbreviated

forms of, respectively, *you* and *and*. However, by examining

concordances for these words (that is, lists of the keyword in

context) it becomes clear that they are to be found only in the

screenplay for *Rocky*, and while they suggest a particular

dialectal form of speech for the speaking characters, they are

not common to the blockbuster corpus as a whole. A similar issue

concerns the proper noun *Luke*, which is found only in the *Star*

Wars screenplay. Proper nouns are often featured in keyword

lists simply as a result of their not being present in the

reference corpus. What is interesting about *Luke* is appearance

so high up in the list. Its presence is a result of the other

characters in the film referring to him by name, perhaps in

order to focus the audience's attention on his character and to

confirm him as the main protagonist and focus of the action. By

the same token, *Rose* is specific to *Titantic*, and, it would

appear, for much the same reason. Similarly, *Reed* is found only

in the *Fantastic Four* screenplay. *God* is found in most of the

screenplays where it tends to be used as a minor expletive,

though its appearance as a keyword is perhaps explained by the

numerous references to religion in two films particularly:

Indiana Jones and the Last Crusade and *Titanic*.

   Ship is found primarily in *Star Wars* (as an abbreviation of

starship) and *Titanic*, which is unsurprising and unrelated to

the blockbuster corpus as a whole. These are keywords relating

to what Philips has called 'aboutness' as a result of their intrinsic connection to what might be termed the propositional content of the story. The same is true for *cockpit*, which is again restricted to *Star Wars*. *Dr* is found in a number of screenplays and it certainly the case that blockbusters often feature characters who are academics (albeit unconventional ones). Nonetheless, it is difficult to assert this feature as a central characteristic of blockbusters (not least because of a certain degree of skew arising from the presence of two Indiana Jones films in the corpus, the main protagonist of which is Dr Jones), though we might reasonably argue for its being a peripheral characteristic.

Having eliminated a number of the male keywords, we are left with the following: *the*, *sir*, *this*, *Mr*, *all-right*, *here*, *us*, *guys*, *me*, *world* and *of*. Culpeper (38) notes that 'Generally, it appears to be the case that aboutness keywords relate to "open class" words, whilst stylistic keywords relate to "closed class" words'. That is, grammatical words are more likely to be indicators of aspects of character, text or authorial style than open class words. Nonetheless, this is not to dismiss open class words out of hand. For example, the open class words in the above list may well be indicative of certain central characteristics of blockbusters. References to the *world*, for instance, may be indicative of scale being particularly

important in blockbuster narratives, while *guys* is noteworthy as an indicator of the prevalence of male to male conversation (*guys* is used primarily to refer to male characters in the corpus). *All-right* tends to be used as an assessment of the situation and seems related to the ongoing resolution of complicating actions (see Labov and Waletsky) that is common to blockbuster narratives.

With regard to the closed class words there are clear connections between some of them. *Sir* and *Mr* are both vocative terms addressed to male characters. Taken together with *guys* they provide some evidence of the 'masculine vernacular' noted by Langford, thereby validating this more subjective analysis. *Me* is interesting, particularly when contrasted with *you* in the female keyword list, suggesting perhaps a greater degree of self-interest on the part of male characters when compared against females. While *me* is also a keyword for female characters, *you* suggests a focus on other characters too. In the Brown and Levinson politeness model, this might be described as other-directed facework (as opposed to self-aggrandisement), and is perhaps characteristic of female characters in blockbuster movies.

In the male keyword list this leaves the grammatical words *the*, *this*, *here* and *of* (although *here* is strictly speaking open-class, it is not the kind of adverb that is open to

morphological adaptation). Of course, statistical significance does not necessarily equate to interpretative significance and it is difficult to see an overwhelmingly clear reason for the presence of *of* as a keyword. The same goes for *the*, though here it is possible to note that the preponderance of definite articles in the male speech also indicates a preponderance of existential presuppositions, and this may well constitute a central characteristic of blockbusters. Finally, we can note that *this* and *here* and both proximal deictic terms, and this seems important when we consider that blockbusters tend to be fast-moving thrillers with a focus on the here-and-now (cf. sedate, reflective art-house films).

Turning to the female keyword list, the presence of many of these words have been accounted for in the above analysis of the male keywords. What is striking here, however, is the large number of proper nouns; eleven out of twenty keywords are names. Although these are accounted for by the fact that they are specific to particular screenplays, what is interesting is that all but one of these are the names of male characters. The only female name that turns up as key in both the male and female dialogue is *Rose*, despite the fact that there are at least two other films in the corpus featuring strong female leads (*Alien* and *Basic Instinct*). This is perhaps further evidence for the notion of a male vernacular being common to blockbusters.

The other two interesting keywords in the female list are
kill and *help*. The concordance of *kill* (figure 1) reveals that
none of these are instances of female characters taking
responsibility for killing. In line 2, where the character says
'we have to kill it', it is noteworthy that the subject of this
sentence is the plural pronoun *we* as opposed to the singular *I*
(see also line 3) In other cases, *kill* occurs in structures that
indicate or speculate that a female character cannot be
responsible for the act of killing.

```
1.  mber . We can't grab it. We can't    kill  it ... Yeah . Where 's its mouth . L
2.  ducks in the freezers. We have to     kill  it first . We can't kill it . If we
3.  e have to kill it first. We can't     kill  it. If we do , it will spill its bod
4.  dto us . Those bastards. How do we     kill  it. How . No way . We've had enoug
5.  asting your time. Catherine didn't     kill  him . I know who you are . How did h
6.  Do you thinkI 'd be dumb enough to     kill  anyone in the exact way I 've describ
7.  rite a book about a killing and then  kill  him the way I described in my book .
8.  . For my book . How does it feel to    kill  someone ? I do n't know . But you do
9.  watch me all the time . She tried to   kill  you , didn't she ? Do you think I t
10. 't she ? Do you think I told her to    kill  You ? Everybody that I care about di
11. ou really think I... that I could      kill  someone ... I never even met Johnny
12. What possible motive would I have to   kill  him ? She's evil . She 's brilliant
13. eport about her . You still think I    kill  people , don't you ? Liar . I decid
14. ll be killed . The fall alone would    kill  you. How cold ? No . I know what ic
15. re , can you ? Then you'll have to      kill  me before it 's too late . Before I
16. ey could make us worse, maybe even     kill  us. Please don't make this persona
```

Figure 1 Concordance of *kill* in the female dialogue

What is perhaps significant about *help* is that 13 of the 25 instances are direct requests for help, confirming the stereotypical notion of female characters as passive and helpless.

Keywords, then, are useful as means of revealing characteristic features of a text, and in this particular case, of the stylistic aspects of male and female dialogue. Wmatrix also offers the facility to calculate key semantic domains; that is, those semantic fields that are over-represented in the target corpus. These are calculated by assigning every word in the corpus a particular semantic tag, based on what is essentially an in-built thesaurus. Key domains can be indicative of thematic elements of a text (see, for example, Afida, and Archer, Culpeper, and Rayson) and in this respect are of value in attempting to uncover some of the prototypical textual elements of blockbusters. Tables 5 and 6 detail the first 20 key semantic domains in the male and female dialogue (key domains are conventionally rendered in small capitals):

Key domain	LL score (p < 0.0001; critical value = 15.13)

UNMATCHED	5799.90
SAILING, SWIMMING, ETC.	533.83
IN POWER	465.81
LOCATION AND DIRECTION	439.43
DEAD	433.32
FLYING AND AIRCRAFT	345.43
THE UNIVERSE	328.14
PEOPLE: MALE	307.86
WARFARE, DEFENCE AND THE ARMY; WEAPONS	294.19
LIGHT	290.42
ALIVE	257.37
OBJECTS GENERALLY	241.79
SPEECH ACTS	235.19
CAUSE & EFFECT/CONNECTION	230.98
GEOGRAPHICAL TERMS	228.69
TIME: GENERAL	180.12
HELPING	179.78
SCIENCE AND TECHNOLOGY IN GENERAL	177.40
RELIGION AND THE SUPERNATURAL	166.30
NO CONSTRAINT	137.29

Table 5 First 20 key domains in the male dialogue

Key domain	LL score (p < 0.0001; critical value = 15.13)
UNMATCHED	1428.99
SAILING, SWIMMING, ETC.	166.48
DEAD	90.73
SPEECH ACTS	90.04
ALIVE	86.28
CAUSE & EFFECT/CONNECTION	68.59
LIGHT	60.73
TIME: GENERAL	58.16
PERSONAL NAMES	57.18
PEOPLE: MALE	47.62
LOCATION AND DIRECTION	46.59
POLITE	44.98
HELPING	41.90
WARFARE, DEFENCE AND THE ARMY; WEAPONS	37.45
OPEN; FINDING; SHOWING	33.84
MENTAL OBJECT: MEANS, METHOD	32.68
MENTAL ACTIONS AND PROCESSES	31.28
EXPECTED	30.67
LAW AND ORDER	25.87

IMPORTANT	23.94

Table 6 First 20 key domains in the female dialogue

As a means of making the analysis of key domains manageable, table 7 details those domains that are found in one sub-corpus but not the other, and those that are common to both:

Domains in male but not female dialogue	Domains in female but not male dialogue	Domains common to both male and female dialogue
IN POWER	PERSONAL NAMES	UNMATCHED
FLYING AND AIRCRAFT	POLITE	SAILING, SWIMMING, ETC.
GEOGRAPHICAL TERMS	OPEN; FINDING; SHOWING	DEAD
SCIENCE AND TECHNOLOGY IN GENERAL	MENTAL OBJECT: MEANS, METHOD	SPEECH ACTS
RELIGION AND THE SUPERNATURAL	MENTAL ACTIONS AND PROCESSES	ALIVE
NO CONSTRAINT	EXPECTED	CAUSE & EFFECT/CONNECTION
	LAW AND ORDER	LIGHT
	IMPORTANT	TIME: GENERAL
		PEOPLE: MALE
		LOCATION AND DIRECTION
		HELPING
		WARFARE, DEFENCE AND THE ARMY; WEAPONS

Table 7 Reduced key domain list

If we now concentrate on table 7, at the top of the list of male

domains is IN POWER. Below is every tenth concordance line for

this domain, to give an indication of its contents:

1.	out on you . Oh , no ,	**sir**	. I 've never had a woma
2.	n thing sneaking in ---	**Chief**	! Show a little respect
3.	e me -- fast ! Grab the	**leader**	. He ai n't normal , thi
4.	cond chance . Ya need a	**manager**	. An advisor . I been in
5.	ere is the Ambassador ?	**Commander**	, tear this ship apart u
6.	ush on us ? Excuse me ,	**sir**	, but that R2 unit is in
7.	ng as .... The Imperial	**Senate**	will no longer be of any
8.	e Owen ! And , now Your	**Highness**	, we will discuss the lo
9.	rough him . You mean it	**controls**	your actions ? Partially
10.	es . There 's no one on	**board**	, sir . According to the
11.	e . Everything is under	**control**	. Situation normal . Wha
12.	learner ; now I am the	**master**	. Only a master of evil
13.	he Incom T-sixty-five ?	**Sir**	, Luke is the best bushp
14.	Leader ... This is Gold	**Leader**	We 're starting out at
15.	e wingmen . I 'm on the	**leader**	. 'S SHIP . Luke 's ship
16.	re n't mentioned in the	**order**	. See it from their poin
17.	x- pects progress . You	**led**	me to be- lieve- Nothing
18.	m Roger Delacorte – the	**Head**	Librarian . Are you the
19.	easures , Strategic Air	**Command**	... Black belt in Karate
20.	l got it ? Well , yes ,	**sir**	. It 's right here ! I '
21.	ut if you 're aScottish	**lord**	, then I am MickeyMouse
22.	ase . Fahrscheine meine	**Dame**	. Bitte . Guten Tag , He
23.	er you , Junior . Yes ,	**sir**	! Haaa ! Who was this fu
24.	ks . What does he say ?	**Insists**	nothing 's wrong on the
25.	ake DNA information and	**organize**	it . In this room , we t
26.	Rica . When 's the damn	**power**	coming on ? Strong legs
27.	, it was with auxiliary	**power**	. Jesus . The auxiliary
28.	m in the sitting room .	**Heading**	for bedroom B-54 . Stay
29.	e practically goddamned	**royalty**	, ragazzo mio ! ! You se
30.	at he stays there . Yes	**sir**	! Reminds me of my Harva

```
31. ! She is made of iron ,  sir            . I assure you , she can
32. d children only ! Sorry  sir            , no men yet . You 're a
33. hell was it ? They 're   sayin'         it 's space rocks . Rock
34. ave enough left over to  duke           it out with Burt Reynold
35. he camera ? Oh , I also  di-rec-tor     . Russian cinema . MIR m
36. Colonel Sharp , this is  Chief          of Staff Collinswood . H
37. l evidence . All done ,  chief          . - right - - do n't go
38. 's McCloy , the Biocyte  CEO            . Nekhorvich 's boss ? Y
39. eous . I wo n't let you  take control   of my company . Sit down
40. screws ? He 's smart ,   powerful       , successful -- Well may
41. stock from livestock .   Sir            . Reed 's comments at th
42. into the Baxter 's main  power          to generate enough volta
```

Figure 2 Concordance of IN POWER in the male dialogue

To validate Langford's assertion that an opposition to authority is common to male characters would require further qualitative analysis, though the prevalence of IN POWER as a semantic domain does accord his notion of a masculine vernacular in blockbusters. Indeed, here is where we find the keyword *sir*. As King ("New Hollywood") points out, 'The dominant genres of the contemporary blockbuster tend to be strongly male-oriented' and in the IN POWER domain we see the predominance of power as a theme of male speech. This may be seen as tangentially related to the notion of strength and prowess that Neale (75) claims as characteristic of male figures in blockbusters.

As with the keyword lists, not all statistically significant domains will have interpretative significance.

Staying with the male domains, FLYING AND AIRCRAFT is specific only to a small number of films and does not therefore offer any generalisable finding for blockbusters as a whole. The same is true of RELIGION AND THE SUPERNATURAL. SCIENCE AND TECHNOLOGY IN GENERAL, on the other hand is dispersed fairly evenly across the corpus, and we might therefore point to this as a characteristic feature of blockbuster movies, central particularly to male dialogue. GEOGRAPHICAL TERMS is also fairly evenly dispersed and consists of such items as *hemisphere*, *earth*, *island*, *bay*, *desert*, *mountain*, *well*, etc. Here we might note a focus on location that I will pick up on below in my discussion of the LOCATION AND DIRECTION domain. NO CONSTRAINT incorporates such lexemes as *release*, *unlocking*, *let it out*, *confined* and *freedom*, and its status as a key domain for male speech perhaps suggests another thematic aspect of blockbusters that is tied particularly to male characters.

If we now turn to the key semantic fields in the female dialogue we can note some very different thematic domains. It seems particularly noteworthy that PERSONAL NAMES and POLITE are key, since this suggests a degree of interpersonal attention that is not present in the male dialogue, and is potentially gender-related. We saw in the analysis of female keywords that proper nouns were over-represented in the female dialogue; the keyness of the related semantic domain shows that this focus on names is

not restricted to a specific few. 15 of the 16 items in the
domain POLITE are thanking expressions, and it again seems
interpretatively significant that these are over-represented to
such a degree in the female dialogue.

Of the 52 items in OPEN; FINDING; SHOWING, 28 refer to the
action of finding, as figure 3 shows:

1. n , I should have expected to **find**   you holding Vader 's leash . I
2. , we've got to get across . **Find**   the control that extends the b
3. analyzed , a weakness can be **found**   . It 's not over yet ! It is f
4. re ... Working on it. Eureka. **Found**   it . What the hell are we doin
5. m the ship . First we have to **find**   it . I 've checked on the supp
6. 's another problem. How do we **find**   it . There 's novisual communi
7. t . And what do we do when we **find**   it . Trap it somehow . Why do
8. his meathead . I thought I 'd **find**   you here . We 've got an hour
9. anyway . Now if we could only **find**   it . What 's it key on . We se
10. ere accusing him . If I could **find**   the commandcomputer key , I co
11. xcept for that's crap which I **found**   near his chair . Here is the l
12. r us to stay a little longer. **Find**   something ? My God , I must be
13. a cigarette ? Yes you do . I **found**   some in my purse ; would youli
14. ased . That 's whyyou did n't **find**   it in your computer . She said
15. er ? It 's ... gold ! Grandpa **found**   gold . That 's a million year
16. o ? Put in a piece of amber , **find**   a mosquito , drill it out . Ri
17. is before . Right . I need to **find**   Wu . I have to run a few tests
18. d , what 's with the phones ? **Find**   what ? Look , there is a sick
19. lose to the animals ! Where I **found**   Freda 's baby . Told you ! Act
20. was just wondering if you had **found**   the " Heart of the Ocean " yet
21. o you live , Mr. Dawson ? You **find**   that sort of rootless existenc
22. . And you will not attempt to **find**   me . In return I will keep my
23. an easy choice . Uh-huh . Go **find**   the wealthy lady you came with
24. nd if I 'm on top ? Damn it ! **Find**   what ? – right.where is it ? W
25. how in the world did you ever **find**   me ? How do you that . Ahh . Y

```
26. feeling alright ? We need to   find         him . Victor , I 'm sorry I --
27. Ten . Waiting for you to come   find         me . That would have kinda def
28. ... I 'm saying it now . Come   find         me . That 's my nose , genius
```

Figure 3 Lemmas of 'find' in concordance of OPEN; FINDING; SHOWING in the female dialogue

Here we might note a central characteristic of blockbuster movies – that of detection in all its variants. That this characteristic is located primarily in the female dialogue potentially indicates one of the functions of female characters in such films.

It is difficult to accord any interpretative significance to MENTAL OBJECT: MEANS, METHOD or MENTAL ACTIONS AND PROCESSES. The former is dominate by instances of 'way', while the latter contains very few items which are not widely dispersed across the corpus as a whole. LAW AND ORDER is also not widely dispersed enough to suggest any general characteristics. EXPECTED, on the hand, suggests that expectation (e.g. lemmas of *hope*, *expect* and *anticipate*) is a significant characteristic of female dialogue in the blockbuster. IMPORTANT, too, seems to have a similar function.

At the top of the list of domains common to both male and female speech is UNMATCHED. This is Wmatrix's category for words it is unable to assign to any other category. A look at the

contents of this category reveals that is composed primarily of neologisms and elements of punctuation (e.g. --). I propose to disregard the category here, for reasons of space, though we might briefly note that neologisms would be worthy of further study in relation to the blockbuster.

SAILING, SWIMMING, ETC. and WARFARE, DEFENCE AND THE ARMY; WEAPONS are best classified as 'aboutness' domains, pertaining particular to, respectively, *Jaws* and *Titanic*, and *Star Wars* and *Armageddon*. DEAD and ALIVE, in contrast, are thematic oppositions that appear to be central characteristics of the blockbuster. The high degree of spectacle and thrills associated with blockbusters are perhaps intensified as a result of being motivated by life-or-death scenarios.

The prevalence of SPEECH ACTS is perhaps related to the need for plot-advancing dialogue in the blockbuster. This hypothesis appears to be confirmed when we examine the n-grams generated by Wmatrix. N-grams are sequences of words that are repeated in the data ('n' stands for any number, hence a 4-gram is a sequence of 4 words). If we calculate 5-grams for the male speech and 4-grams for the female speech (the are no 5-grams in the female dialogue), the results are as follows:

No.	Male 5-grams	Frequency

1	what are you doing here	8
2	the well of the souls	6
3	to get out of here	6
4	what are you talking about	6
No.	**Female 4-grams**	**Frequency**
1	what are you doing	10
2	what do you want	9
3	get me out of	6
4	help me obi wan	5
5	i don't want to	5
6	me obi wan Kenobi	5

Table 8 N-grams in the male and female dialogue

The first 5-gram in the male dialogue is an interrogative whose function is perhaps to reflect the element of surprise typical of blockbuster plots. The fourth 5-gram is another interrogative that facilitates the delivery of plot-advancing dialogue from the addressee. The first and second of the female 4-grams have a similar function. I will consider the other n-grams momentarily.

If SPEECH ACTS are important aspects of the blockbuster then so too is the concept of CAUSE AND EFFECT/CONNECTION. Discussing the concept of narrative, King ("New Hollywood" 183) notes that:

One tendency in debates about the relationship between
narrative and spectacle in the contemporary blockbuster has
been to exaggerate the importance of classical [i.e.
linear, cause and effect] narrative in the studio era [i.e.
before the New Hollywood], at the expense of other appeals.
Another has been to *under*estimate the importance of
narrative – 'classical' and otherwise – today.

If it is indeed the case that film scholars have underestimated
the integral nature of narrative to the blockbuster, then the
CAUSE AND EFFECT/CONNECTION domain highlights its importance. Many of
the items within this domain are related to narrative drive and
constitute elements of the n-grams in table 8, particularly the
word *why* as a facilitator of plot-advancing dialogue.

LOCATION AND DIRECTION is also key to the blockbuster. Here we
can make a connection with the proximally deictic keywords *this*
and *here*, which form part of this domain, and note that a sense
of immediacy and a strong sense of place is of particular
importance in generating a feeling of involvement and excitement
for the blockbuster audience.

Finally, the remaining domain common to both male and
female dialogue which has interpretative significance is PEOPLE:
MALE. The contents of this domain seem inevitably to relate to
King's ("New Hollywood" 138) assertion that the target audience

for blockbusters is primarily males in their teens and twenties.
Furthermore, the keyness of this domain for both male and female
character dialogue would seem to confirm a degree of
institutional sexism inherent in most blockbusters.

5. Conclusion

My aim in this article has been to demonstrate how techniques
from corpus linguistics might be employed to uncover some of the
prototypical stylistic characteristics of dialogue in
blockbuster movies. I have refrained from making the claim that
any of my findings constitute defining features of the
blockbuster. Rather I would argue that they are best seen as, to
varying degrees, central, secondary and peripheral features,
dependent on such extra-textual factors as the particular viewer
and the context of viewing. In some cases, the analysis above
confirms some of the more qualitative judgements of film
critics, the value of which is to highlight the accuracy of the
original critical method. Beyond this, a corpus linguistic
approach offers new insights into what might constitute valuable
areas for future research. As Rayson ("Keywords") has pointed
out, one particularly useful feature of Wmatrix is its capacity
for generating candidate research questions. For example, the
notion that EXPECTATION is a key semantic domain for female
characters might prompt a qualitative study of how this emotion

is revealed, or a non-linguistic study of its multimodal disclosure. What should be clear is that the above corpus linguistic analysis does not, nor is intended to, provide conclusive answers. It is a tool to be deployed alongside the other techniques used by film and literary critics that has the capacity to provide objective quantitative support for qualitative or subjective claims.

Works Cited

Afida, Mohamad Ali. "Semantic Fields of Problem in Business
    English: Malaysian and British Journalistic Business
    Texts." (2007) *Corpora* 2.2: 211–39. Print.

Archer, Dawn, Jonathan Culpeper, and Paul Rayson. "Love – A
    Familiar or a Devil? An Exploration of Key Domains in
    Shakespeare's Comedies and Tragedies." *What's in A Word-
    list? Investigating Word Frequency and Keyword Extraction*.
    Ed. Dawn Archer. Aldershot: Ashgate, 2009. 137–58. Print.

Bordwell, David. *Narration in the Fiction Film*. London:
    Routledge, 1986. Print.

Bordwell, David. *Making Meaning: Inference and Rhetoric in the
    Interpretation of Cinema*. Cambridge, MA: Harvard University
    Press, 1989. Print.

Brown, Penelope, and Stephen Levinson. *Politeness: Some
    Universals in Language Usage*. Cambridge: Cambridge
    University Press, 1987. Print.

Buckland, Warren. "A Close Encounter with *Raiders of the Lost
    Ark*: Notes on Narrative Aspects of the New Hollywood
    Blockbuster." *Contemporary Hollywood Cinema*. Eds. Steve
    Neale and Murray Smith. London: Routledge, 1998. 166–77.
    Print.

Cohan, Steven. "Censorship and Narrative Indeterminacy in *Basic
    Instinct*: 'You Won't Learn Anything From Me I Don't Want

You to Know." *Contemporary Hollywood Cinema*. Eds. Steve

Neale and Murray Smith. London: Routledge, (1998). 263–79.

Print.

Culpeper, Jonathan. "Keyness: Words, Parts-of-Speech and

Semantic Categories in the Character-Talk of Shakespeare's

Romeo and Juliet." *International Journal of Corpus

Linguistics* 14.1 (2009): 29–59. Print.

Culpeper, Jonathan, and Dan McIntyre. "Activity Types and

Characterisation in Dramatic Discourse." *Characters in

Fictional Worlds: Interdisciplinary Perspectives*. Eds. Ralf

Schneider, Fotis Jannidis and Jens Eder. Berlin: De

Gruyter, 2010. 1–33. Print.

Hills, Matt. "*Star Wars* in Fandom, Film Theory, and The Museum:

The Cultural Status of the Cult Blockbuster." *Movie

Blockbusters*. Ed. Julian Stringer. London: Routledge,

(2003): 178–89. Print.

King, Geoff. *Spectacular Narratives: Hollywood in the Age of the

Blockbuster*. London: I. B. Taurus, 2000. Print.

King, Geoff. *New Hollywood Cinema: An Introduction*. London: I.

B. Taurus, 2002. Print.

Kozloff, Sarah. *Overhearing Film Dialogue*. Berkeley, CA:

University of California Press, 2000. Print.

Labov, William, and Joshua Waletzky. "Narrative Analysis: Oral

Versions of Personal Experience." *Essays on the Verbal and*

Visual Arts. Ed. J. Holm. Seattle: University of Washington
Press, 1967. 12–44. Print.

Langford, Barry. *Film Genre: Hollywood and Beyond*. Edinburgh:
Edinburgh University Press, 2005. Print.

Lavik, Erlend. "The Battle for the Blockbuster: Discourses of
Spectacle and Excess." *New Review of Film and Television
Studies* 6.2 (2008): 169–87. Print.

Mandala, Susan. *Twentieth Century Drama Dialogue as Ordinary
Talk*. Farnham: Ashgate, 2007. Print.

McIntyre, Dan. "Integrating Multimodal Analysis and the
Stylistics of Drama: A Multimodal Perspective on Ian
McKellen's *Richard III*." *Language and Literature* 17.4
(2008): 309–34. Print.

McIntyre, Dan. "Dialogue and Characterisation in Quentin
Tarantino's *Reservoir Dogs*: A Corpus Stylistic Analysis."
Language and Style. Eds. Dan McIntyre and Beatrix Busse.
Basingstoke: Palgrave, 2010. Print.

McIntyre, Dan, and Brian Walker. "How Can Corpora Be Used to
Explore The Language of Poetry and Drama?" *The Routledge
Handbook of Corpus Linguistics*. Eds. Mike McCarthy and Anne
O'Keefe. Abingdon: Routledge, 2010. 516–30. Print.

Neale, Steve. 'Action–Adventure as Hollywood Genre." *Action and
Adventure Cinema*. Ed. Yvonne Tasker. London: Routledge,
2004. 71–83. Print.

Phillips, Martin. *Lexical Structure of Text*. [Discourse Analysis
     Monographs 12] Birmingham: University of Birmingham, 1989.
     Print.

Piazza, Roberta. "Voice-Over and Self-Narrative in Film: A
     Multimodal Analysis of Antonioni's *When Love Fails* (*Tentato
     Suicidio*)." *Language and Literature* 19.2 (2010): 1-22.
     Print.

Piazza, Roberta. *Let Cinema Speak*. London: Continuum,
     forthcoming. Print.

Piazza, Roberta., F. Rossi, F, and Monika Bednarek. (eds)
     *Telecinematic Discourse: An Introduction to the Fictional
     Language of Cinema and Television*. Amsterdam: John
     Benjamins, forthcoming. Print.

Rayson, Paul. "Keywords are Not Enough." JAECS (Japan
     Association for English Corpus Studies). Chuo University,
     Tokyo, Japan. 27 November 2004. Reading.

Rayson, Paul. *Wmatrix: A Web-based Corpus Processing
     Environment*. Computing Department, Lancaster University.
     <http://www.comp.lancs.ac.uk/ucrel/wmatrix/>, 2008.
     Software.

Rosch, Eleanor. (1975) "Cognitive Representations of Semantic
     Categories." *Journal of Experimental Psychology: General*
     104.3 (1975): 192-233. Print.

Richardson, Kay. *Television Dramatic Discourse: A*
    Sociolinguistic Study. Oxford: Oxford University Press,
    forthcoming. Print.

Schatz, Thomas. "The New Hollywood." *Film Theory Goes to the*
    Movies. Eds. Jim Collins, Hilary Radner, and Ava Preacher
    Collins. London: Routledge, 1993. 8–36. Print.

Scott, Mike. *WordSmith Tools* 4.0. Oxford: Oxford University
    Press, 2004. Software.

Scott, Mike. "In Search of a Bad Reference Corpus." *What's in A*
    Word-list? Investigating Word Frequency and Keyword
    Extraction. Ed. Dawn Archer. Aldershot: Ashgate, 2009. 79–
    92. Print.

Stringer, Julian. "Introduction." *Movie Blockbusters*. Ed. Julian
    Stringer. London: Routledge, 2003. 1–14. Print.

Stringer, Julian. (ed.) *Movie Blockbusters*. London: Routledge,
    2003. Print.

Tasker, Yvonne. *Spectacular Bodies: Gender, Genre and the Action*
    Cinema. London: Routledge, 1993. Print.

Xiao, Richard, and Tony McEnery. (2005) "Two Approaches to Genre
    Analysis: Three Genres in Modern American English', *Journal*
    of English Linguistics 33(1): 62–82. Print.