Digital Scholarship in the Humanities, 2023, **38**, 1678–1691
https://doi.org/10.1093/llc/fqad062
Full Paper

OXFORD

Film dialogue and R-stylo

Barry Salt *

*Correspondence address: barrysalt@postmaster.co.uk

Abstract

The dialogue of a large number of American feature films of the last 30 years is analysed with the stylometric tools contained in the R-stylo package. Various interesting results showing the capabilities and restrictions of this statistical package emerge.

Keywords: literary stylometry; dialogue; American; feature films

1. Introduction

The package R-stylo, devised by Maciej Eder and his associates, (Eder et al. 2016) brings together a number of established statistical multivariate algorithms under a menu control interface. The algorithms covered are bootstrap consensus networks, delta classifier, k-NN classifier, principal components analysis, and support vector machines (SVMs). It has been used for analysing literary texts, and has also been extended to analysing the dialogue of television shows in an impressive article by Byszuk (2020) and to film dialogue by Hołobut and Rybicki (2020). There has apparently been no earlier work on using these methods of statistical analysis on film dialogue, with the exception of a paper by Buckland (2019). This article is a demonstration of R-stylo's application to analysing the dialogue in American feature films of the last couple of decades, and takes a different approach to the article by Hołobut and Rybicki just mentioned, which is only concerned with generic classification of films using their dialogue. In their particular application, Hołobut and Rybicki are quite successful at classifying films by genre using the R-Stylo Consensus Tree algorithm, combined with the Gephi programme.

The classification of films by genre by the film industry has existed for almost the entire history of cinema, and has always been a bit fuzzy, as Hołobut and Rybicki remark. In recent times, it has got even fuzzier under the influence of the film fans contributing to the listings in the International Movie Database.

The texts I use in this article are gathered from the recordings of the subtitles from DVD copies of films and TV shows collected by *opensubtitles.org*. These subtitle records are not a perfect reproduction of the dialogue spoken in films, as briefly mentioned in the article by Joanna Byszuk. Their main shortcoming is that not all the words spoken in a film are subtitled, to allow for slow readers, and also to prevent the text overflowing the picture when a lot is said quickly in a scene. Also, a very small proportion of words are changed in pursuit of the same objective.

I give three examples from films with a large amount of dialogue to indicate the order of this effect. For *Mr Deeds Goes to Town* (1936) and *His Girl Friday* (1940) all the dialogue is subtitled, with just a few word changes. These subtitles are taken from new restoration disks which treat the films as classics. The idea is presumably that all connoisseurs want to see the complete dialogue, and also have a high reading speed. For *The American President* (1995), on the other hand, 7 per cent of the dialogue is not subtitled, and I believe that is more typical.

The second problem is that the subtitles can include extra text that is not spoken in the film. This is either the words of popular songs heard on the sound track or the description of significant sounds also heard on the sound track. The use of sound description in the subtitles for DVDs has only really appeared in the last couple of decades, and is nowadays most prominent in the subtitling of television drama. The films I am considering contain little of this, and in any case, I have removed the transcriptions of song lyrics from the films in my corpora. Even further, a quick test with the subtitles file of one recent film shows that removing the audio description text makes no difference to the classification of that film by R-stylo. To regain the complete and pure text of only the words spoken in a film requires a large expenditure of time and/or money, and the intent of this article is to show that it is possible to get good authorship discrimination without it. In any case, working with incomplete texts is an established part of literary stylometry.

There are two corpora I will work with. The first is all the American commercial feature films released in the USA in 1999 for which I could obtain subtitles, to the total of 80 films. The important point about this selection is that none of the writers who worked on these films contributed to more than one of them. The second corpus is a selection of films made by prominent writer/directors of the last few decades. The reason for concentrating on the scripts of writers who direct their own films is that there is much less likelihood of the production company interfering with the script by having other writers work on it uncredited.

As a preliminary to my investigation, I apply the other R-Stylo algorithms not used by Agata Hołobut and Jan Rybicki for generic classification. The result of using principal components analysis on eighty films from 1999 (titles available on *starword.com*) is shown in the following graph. The R-stylo algorithms ignore the prefixes to the dialogue file titles during their working, but afterwards add the colours to the finished graph by using the file prefixes.

(Although the Stylo programmes were set at 1,200 MFW, for some reason the programmes give 1102 MFW on all the graph captions.)

The dialogue files for the films have been generically classified by myself, as indicated by the prefixes to their names, along obvious lines. That is, drama, comedy (com), romantic comedy (romcom), thriller (thrill), horror (horr), and so on, with different coloured lettering for each. It can be seen that the genres mostly overlap on the graph, and their groups are mostly centred a bit above the zero point. However, there is the beginning of separation for the action and science fiction (sf) categories, placing them towards the negative region of the second principal component. The noticeable outliers, which are *Anna and the King* and *Wing Commander*, deserve a little discussion.

A quick visual scan down the dialogue text for *Anna and the King* shows that the English spoken by the Siamese characters entirely lacks articles, and this shows up in the recorded word frequencies using the AntConc concordance programme (Anthony 2017). The frequencies for 'the' and 'a' are down by 21 per cent and 9 per cent, respectively, from the norm for the whole corpus. So nothing new is being learned on this point from R-stylo, but in the case of *Wing Commander*, the peculiarity of its dialogue is not quite so obvious. A comparison of the *Wing Commander* dialogue keyword frequencies with the norm for the whole 1999 corpus shows a marked deficiency in the personal pronouns 'you' and 'I', which in *Wing Commander* are 34 per cent and 35 per cent below the norm for the whole corpus. The reason for this does not leap out at me from the dialogue itself, but more concordance inspection of it shows that the use of 'you' is mostly to give orders from a superior to an inferior, whereas in normal drama it is more often used to elicit information from the person being addressed. In other words, the film is mostly lacking ordinary emotional interaction between the characters. It is little more than a live action version of the 'shoot 'em up' video game on which it is based, which has hardly any plot of the ordinary dramatic kind. The film was unsurprisingly a commercial failure. Other action films in the sample are also displaced down into the bottom half of the graph along the PC2 axis for the same reason, but in a less extreme way. We have now gone beyond R-stylo, into the region where the meanings of words matter.

Returning to R-stylo, I need to mention the settings of the package that I am using for all of this investigation. These are—0 per cent culling with no pronoun deletion, list cut-off of 5,000, and the MFW settings are 0–1,200 with increments of 100, starting at frequency rank 1. The upper limit of the MFW setting was determined by trying values from 100 to 1,900, with the best discrimination achieved at 1,200. For general literary texts, the preferred setting for this kind of statistical classificatory programmes appears to be 2,000 words. My results show that film dialogue has a much reduced vocabulary compared with literary writing, and indeed it can be readily observed that in literary writing the extended vocabulary occurs in the descriptive parts of the text, and not in the dialogue.

Word digrams and trigrams proved much less useful in script classification. I tried working with longer and longer groups of words up to pentagrams (five-grams), but classification gets worse and worse the longer the group. In the statistics I use classic Delta, and for PCA I only use correlation, not covariance, and a consensus strength of 0.5 for the consensus tree. After many trials, I find that other settings make hardly any difference to the results. This includes the other possible Delta settings, and so I use classic Delta throughout. The reason for Delta settings other than classic Burrows not making any difference in my application of the algorithm is given by the investigation 'Understanding and explaining Delta measures for authorship attribution'. by Stefan Evert et al. reported in *Digital Scholarship in the Humanities*, Vol. 32, Supplement 2, (2017). Most of their tests show little difference between the different forms of delta, up to numbers of most frequent words (nMFW) of 2000. The marked divergence between results for the different deltas only occurs for nMFW values greater than 2,000.

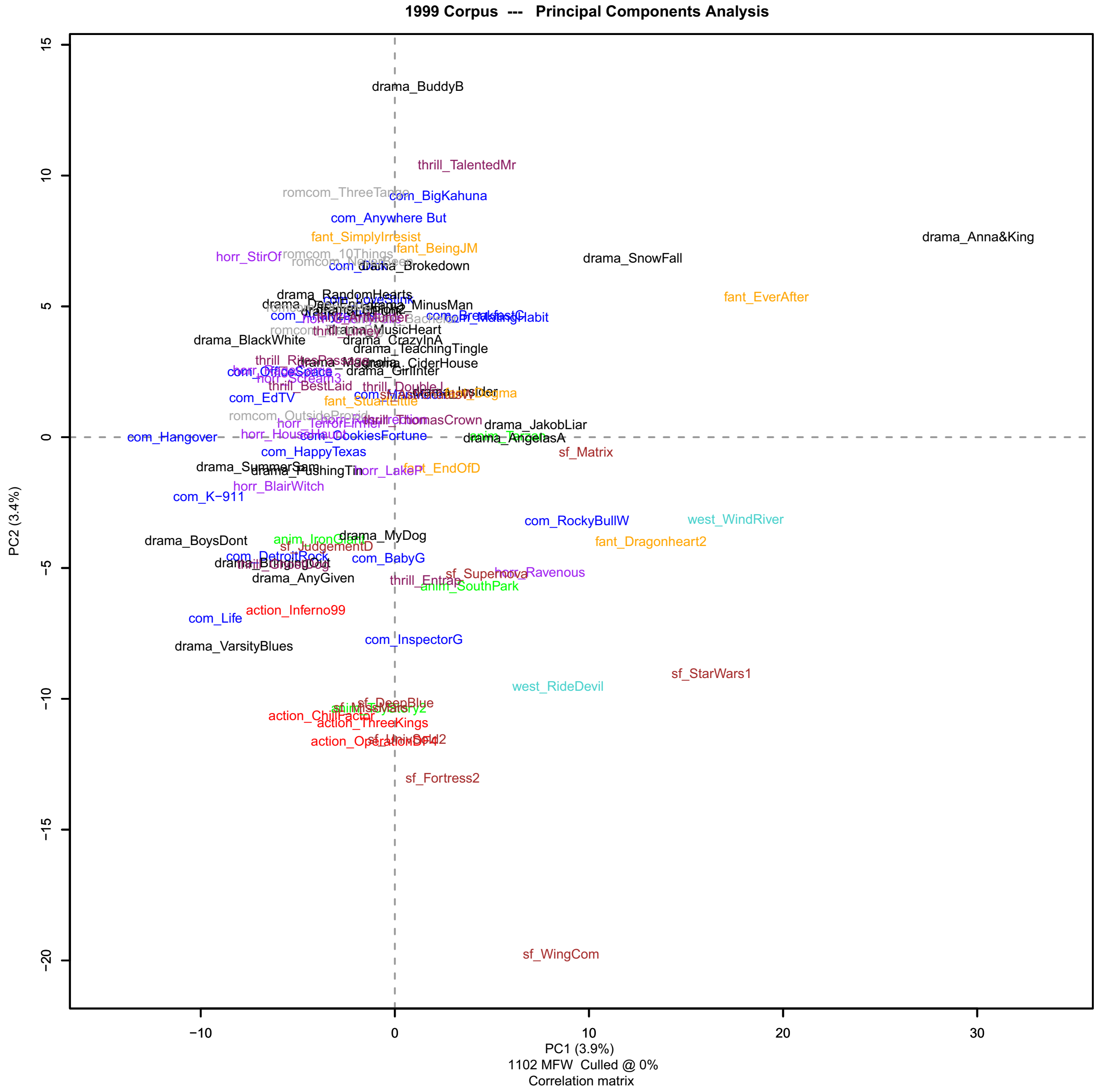


Figure 1.

Applying my settings to the cluster analysis of the sample from 1999 is given in Fig. 2.

As you can see, there is little clear separation by genre in general, though the science fiction films are mostly grouped fairly closely together. One obvious thing that the dialogue texts of the science fiction films have in common, which is missing from other films, is the basic vocabulary of space technology. Making a comparison using AntConc, and looking at the 100 most common keywords from the science fiction sample, while ignoring proper nouns, most of the words come from present-day space travel, like ‘ship’, ‘shuttle’, ‘fuel’, ‘coordinates’, ‘crew’, ‘commander’, ‘power’, ‘earth’, ‘orbit’, ‘surface’, ‘planet’, ‘rover’. Of the other keywords inside the 100 most common, only ‘jump’ (meaning space-time jump) and ‘dimension’ come from the common vocabulary of the imaginary world of science fiction.

An obvious feature of Hołobut and Rybicki’s work is that they are only using six genres in their classification, and excluding the most basic and common genres of drama and comedy.

So for comparison, I now rerun my use of the Stylo cluster algorithm on a reduced corpus of thirty-three films that only includes the science fiction (sf), fantasy

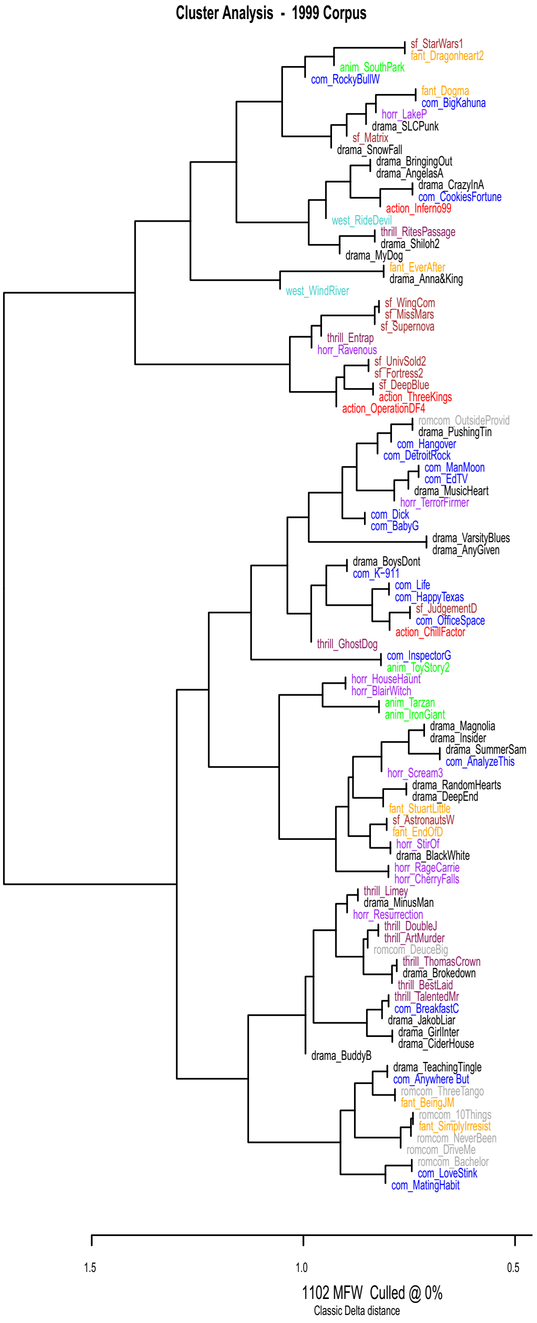


Figure 2.

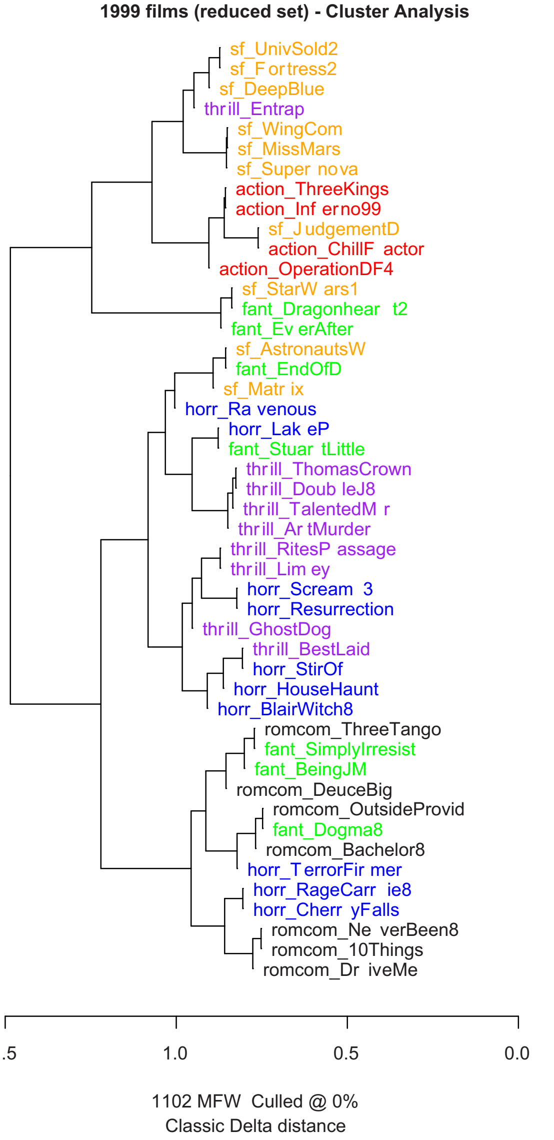


Figure 3.

(fant), horror (horr), action (action), thriller (thrill), and romantic comedy (romcom) (Fig. 3).

This shows better separation of these six genres, though it is still not quite as good as the result of Hołobut and Rybicki's method using Consensus Tree and Gephi.

Having established that R-stylo does not perform well in my tests for separating out scripts by genre, the next step is to look at how well it does in differentiating authorship. This application has not been previously tested.

2. Authorship

To investigate authorship, I work with a corpus of the dialogue from films from the last three decades, most of which were both written and directed by a group of successful American film scriptwriters, though some scripts are included that were directed by others, although written by the named author. In nearly all these examples, the writer–directors concerned had sufficient standing in the films industry to rank as producers of these films, and hence prevent the usual studio interference on their scripts. I have also included a group of the earliest ‘mumblecore’ films (prefix ‘mum_’), whose significance will become clear in due course.

Performing a principal components analysis on this author corpus produces the correlation matrix Fig. 4.

This shows more discrimination and grouping for the different authors than Fig. 1 did for genres. Nevertheless, there is a pile-up near the centre involving Marc Lawrence, Woody Allen, M. Night Shyamalan, and Charlie Kaufman. This is a surprise, as I think that most people, like myself, subjectively think of Shyamalan, Kaufman, and Woody Allen as having distinctive ‘voices’, so seeing their films mixed up with each other, and also those of Marc Lawrence is a shock. However, a little further thought suggests that a lot of the distinctiveness of Shyamalan, Kaufman, and Allen’s films come from their special plot conceptions. Aaron Sorkin’s films are tightly grouped on the edge of this central cluster, and almost separated from it, which is subjectively more satisfactory. Mamet’s scripts are edging away from the main central cluster, and overlapping the tighter group of Gilroy-scripted films. The obvious outliers are Shyamalan’s *The Village*, and Mamet’s *Heist*. The first of these is unusual in having

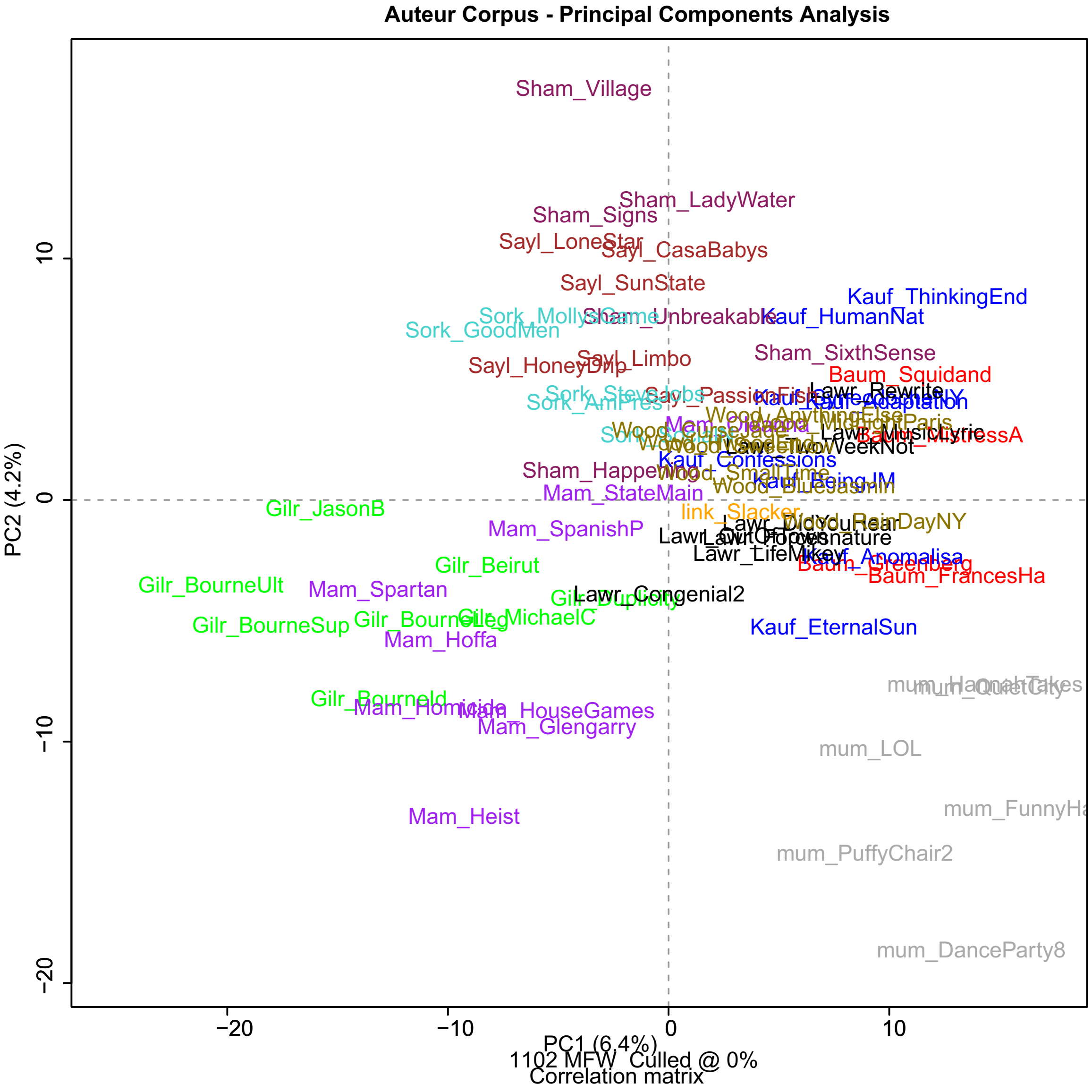


Figure 4.

the plot take place entirely inside a cult, rather than in ordinary society, as in his other films. This is reflected in the dialogue, with 'not' and 'will' in the imperative form appearing at about three times the norm for Shyamalan films, and also with respect to their norm in the whole corpus. David Mamet's *Heist* is way down south past the group of Gilroy films, and gets there by a preponderance of words from the rough action genre, such as 'yeah', 'gonna', and 'go', if one ignores the words that are specific to this story. The fact that *Jason Bourne* is part of a tight cluster with the other Bourne films, despite not being written by Tony Gilroy, shows the importance of genre, and indeed subject matter, over authorship in this case.

Again applying the cluster analysis tool with the basic settings already mentioned, we get (Fig. 5).

The separation of scripts by their authorship is nearly complete, with a few exceptions that are informative in themselves. David Mamet's scripts (Mam_) are broken into two groups, and the Steven Zaillian scripts (Zaill_) are broken into three groups, with *All the King's Men* and *A Civil Action* place next to the group of Aaron Sorkin films, and the other two Zaillian films placed separately into each of the two groups of David Mamet films (Table 1). The division of the Mamet films into two separate groups is itself an outright failure of the cluster analysis to identify an author. The least unsatisfactory reason that I can give for this is that the larger group of Mamet films, including *Heist* and *Hoffa*, have more violent criminality in them. The fact that the *Jason Bourne* script is firmly classified with the other Bourne films scripts, despite not being written by Tony Gilroy, is a significant demonstration of the weakness of these methods in attributing authorship. It is not quite so surprising that Zaillian's script for *Hannibal* is paired at the final level of clustering with Mamet's unproduced script for that same film. I will return to this important occurrence later.

An obvious contributing factor to the success of classification by author here is that many of the writers concerned only work in one genre (or sub-genre) of film.

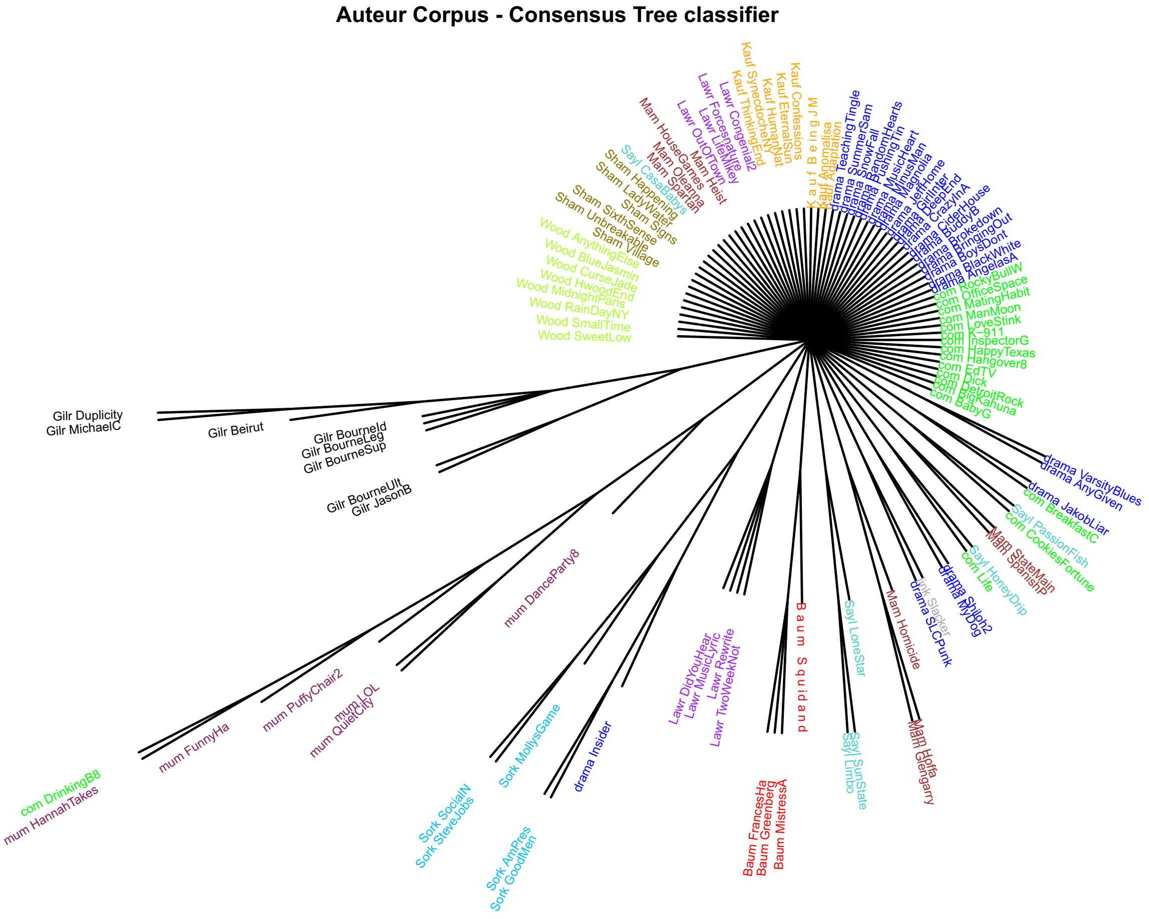


Figure 5.

Table 1. Author Dialogue Corpus.

Dialogue file name	Film title	Year	Director	Script writers
Baum_FrancesHa.txt	Frances Ha	2012	Baumbach, Noah	Baumbach, N. and Gerwig, G.
Baum_Greenberg.txt	Greenberg	2010	Baumbach, Noah	Baumbach, N. and Leigh, J.J.
Baum_MistressA.txt	Mistress America	2015	Baumbach, Noah	Baumbach, N. and Gerwig, G.
Baum_Squidand.txt	Squid and the Whale, The	2005	Baumbach, Noah	Baumbach, N.
Gilr_Beirut.txt	Beirut	2018	Anderson, Brad	Gilroy, Tony
Gilr_BourneId.txt	Bourne Identity, The	2002	Liman, Doug	Gilroy, T. and Herron, W.B.
Gilr_BourneLeg.txt	Bourne Legacy, The	2012	Gilroy, Tony	Gilroy, T. and D.
Gilr_BourneSup.txt	Bourne Supremacy, The	2004	Greengrass, Paul	Gilroy, T. and Ludlum, R.
Gilr_BourneUlt.txt	Bourne Ultimatum, The	2007	Greengrass, Paul	Gilroy, T. and Burns, S. and Nolfi, G.
Gilr_Duplicity.txt	Duplicity	2009	Gilroy, Tony	Gilroy, T.
Gilr_JasonB.txt	Jason Bourne	2016	Greengrass, Paul	Greengrass, P. and Rouse, C.
Gilr_MichaelC.txt	Michael Clayton	2007	Gilroy, Tony	Gilroy, Tony
Kauf_Adaptation.txt	Adaptation	2002	Jonze, Spike	Kaufman, Charlie and Orlean, Susan
Kauf_Anomalisa.txt	Anomalisa	2015	Johnson, D. and Kaufman, C.	Kaufman, Charlie
Kauf_BeingJM.txt	Being John Malkovich	1999	Jonze, Spike	Kaufman, Charlie
Kauf_Confessions.txt	Confessions of a Dangerous Mind	2002	Clooney, George	Kaufman, Charlie and Barris, Chuck
Kauf_EternalSun.txt	Eternal Sunshine of the Spotless Mind	2004	Gondry, Michael	Kaufman, C. and Gondry, M. and Bismuth, P.
Kauf_HumanNat.txt	Human Nature	2001	Gondry, Michael	Kaufman, Charlie
Kauf_SynecdocheNY.txt	Synechdoche, New York	2008	Kaufman, Charlie	Kaufman, Charlie
	I'm Thinking of Ending Things	2020	Kaufman, Charlie	Kaufman, Charlie and Reid, Ian
Lawr_Congenial2.txt	Miss Congeniality 2	2005	Pasquin, John	Lawrence, Marc
Lawr_DidYouHear.txt	Did You Hear About the Morgans	2009	Lawrence, Marc	Lawrence, Marc
Lawr_Forcesnature.txt	Forces of Nature	1999	Hughes, Bronwen	Lawrence, Marc
Lawr_LifeMikey.txt	Life with Mikey	1993	Lapine, James	Lawrence, Marc
Lawr_MusicLyric.txt	Music and Lyrics	2007	Lawrence, Marc	Lawrence, Marc
Lawr_OutOfTown.txt	Out-of-Towners, The	1999	Weisman, Sam	Lawrence, Marc and Simon, Neil
Lawr_Rewrite.txt	Rewrite, The	2014	Lawrence, Marc	Lawrence, Marc
Lawr_TwoWeekNot.txt	Two Weeks Notice	2002	Lawrence, Marc	Lawrence, Marc
Mam_Glengarry.txt	Glengarry Glen Ross	1992	Foley, James	Mamet, David
Mam_Heist.txt	Heist	2001	Mamet, David	Mamet, David
Mam_Hoffa.txt	Hoffa	1992	DeVito, Danny	Mamet, David
Mam_Homicide.txt	Homicide	1991	Mamet, David	Mamet, David
Mam_HouseGames.txt	House of Games	1987	Mamet, David	Mamet, David and Jonathan Katz
Mam_Oleanna.txt	Oleanna	1994	Mamet, David	Mamet, David
Mam_SpanishP.txt	Spanish Prisoner, The	1997	Mamet, David	Mamet, David
Mam_Spartan.txt	Spartan	2004	Mamet, David	Mamet, David
Mam_StateMain.txt	State and Main	2000	Mamet, David	Mamet, David
Mam_HANNIBAL.txt	Hannibal	2001	Scott, Ridley	Mamet, David and Zaillian, Steven
mum_DanceParty8.txt	Dance Party, USA	2006	Katz, Aaron	Katz, Aaron
mum_FunnyHa.txt	Funny Ha Ha	2002	Bujalski, Andrew	Bujalski, Andrew
mum_HannahTakes.txt	Hannah Takes the Stairs	2007	Swanberg, Joe	Swanberg, Gerwig, Osborne, etc.
mum_LOL.txt	LOL	2006	Swanberg, Joe	Brewersdorf and Swanberg and Wells
mum_PuffyChair2.txt	Puffy Chair, The	2005	Duplass, Jay, and Mark	Duplass, Jay, and Mark
mum_QuietCity.txt	Quiet City	2007	Katz, Aaron	Katz, A. and Fisher, E. and Lankenau, C.
Sayl_CasaBabys.txt	Casa de los babys	2003	Sayles, John	Sayles, John
Sayl_HoneyDrip.txt	Honeydripper	2007	Sayles, John	Sayles, John
Sayl_Limbo.txt	Limbo	1999	Sayles, John	Sayles, John
Sayl_LoneStar.txt	Lone Star	1996	Sayles, John	Sayles, John
Sayl_PassionFish.txt	Passion Fish	1992	Sayles, John	Sayles, John
Sayl_SunState.txt	Sunshine State	2002	Sayles, John	Sayles, John
Sham_Happening.txt	Happening, The	2008	Shyamalan, M. Night	Shyamalan, M. Night
Sham_LadyWater.txt	Lady in the Water	2006	Shyamalan, M. Night	Shyamalan, M. Night

(continued)

Table 1. (continued)

Dialogue file name	Film title	Year	Director	Script writers
Sham_Signs.txt	Signs	2002	Shyamalan, M. Night	Shyamalan, M. Night
Sham_SixthSense.txt	Sixth Sense, The	1999	Shyamalan, M. Night	Shyamalan, M. Night
Sham_Unbreakable.txt	Unbreakable	2000	Shyamalan, M. Night	Shyamalan, M. Night
Sham_Village.txt	Village, The	2004	Shyamalan, M. Night	Shyamalan, M. Night
sork_AmPres.txt	American President, The	1995	Reiner, Rob	Sorkin, Aaron
sork_GoodMen.txt	Few Good Men, A	1992	Reiner, Rob	Sorkin, Aaron
sork_MollysGame.txt	Molly's Game	2017	Sorkin, Aaron	Sorkin, Aaron and Bloom, Molly
sork_SocialN.txt	Social Network, The	2010	Fincher, David	Sorkin, Aaron and Mezrich, Ben
sork_SteveJobs.txt	Steve Jobs	2015	Boyle, Danny	Sorkin, Aaron and Isaacson, Walter
Wood_AnythingElse.txt	Anything Else	2003	Allen, Woody	Allen, Woody
Wood_BlueJasmin.txt	Blue Jasmine	2013	Allen, Woody	Allen, Woody
Wood_CurseJade.txt	Curse of the Jade Scorpion, The	2001	Allen, Woody	Allen, Woody
Wood_HwoodEnd.txt	Hollywood Ending	2002	Allen, Woody	Allen, Woody
Wood_MidnightParis.txt	Midnight in Paris	2011	Allen, Woody	Allen, Woody
Wood_RainDayNY.txt	Rainy Day in New York, A	2019	Allen, Woody	Allen, Woody
Wood_SmallTime.txt	Small Time Crooks	2000	Allen, Woody	Allen, Woody
Wood_SweetLow.txt	Sweet and Lowdown	1999	Allen, Woody	Allen, Woody
Zaill_Hannibal.txt	Hannibal	2001	Scott, Ridley	Zaillian, Steven
Zaill_CivilAction	Civil Action, A	1998	Zaillian, Steven	Zaillian, Steven
Zaill_AmerGangster	American Gangster	2007	Scott, Ridley	Zaillian, Steven
Zaill_AllKings	All the King's Men	2006	Zaillian, Steven	Zaillian, Steven

The third operation to perform on this corpus is using R-stylo's Bootstrap Consensus Trees classifier on it (Fig. 6).

The differentiation of the works of the different authors is not perfect, and not superior to that in the preceding Cluster Analysis output.

3. Mumblecore

One of the most notable things about these graphs is that the mumblecore group of films, together with three of the Noah Baumbach scripts, are almost completely separated from all the other films. The language in them that causes this is hesitational interjections like 'um' and 'uh', which are part of what gives mumblecore its name, and these words are seven times more frequent than in the rest of the films of the author sample. But the most important difference is that 'I' is used nearly three times more frequently in mumblecore than in the other films. This corresponds to the self-obsessed nature of the characters in these films, which is their main distinguishing trait, as opposed to the characters in all other films.

4. Joint authorship analysis

The R-stylo package provides an SVM rolling classify algorithm intended to reveal any joint authorship of a text, and I have applied this to the two dialogue scripts of *Hannibal* in the approved way. The setting was mfw = 100, slice size = 1000, and slice overlap = 750. For the reference texts in this analysis, I used *A Civil Action*, *American Gangster*, and *All the King's Men* by Steven Zaillian, and *Heist, Hoffa*, and *Homicide* by David Mamet. The test specimen was Zaillian's dialogue for *Hannibal*. The result is provided in Fig. 7.

The suggestion from this analysis that one or more sections of the Zaillian script includes material written by Mamet is completely wrong. The first script for *Hannibal* was written by David Mamet, and then discarded for a completely rewritten script by David Zaillian, according to the testimony of those writers. As well as visual inspection of the dialogue texts of the two screenplays, I ran them through Plagiarism Checker X. When set to detect four word groups, this programme highlights many banal phrases like 'Special Agent Clarice Starling', 'Do you know what', 'What do you say', and 'What do you think', so I used the six-word setting. This gave only thirteen six-word sections of dialogue in common, out of approximately 10,000 words. In both scripts, the dialogue in the *Hannibal* novel by Thomas Harris is almost completely re-written, except for these very short sections, which are taken almost verbatim in both. But these short sections do not occur in any of the places indicated in the above two graphs, so the SVM algorithm is in error in this case.

A further trial of the SVM algorithm can be made on the dialogue of *Jason Bourne*. Although this is classified with the other Bourne films, all written by Tony Gilroy, Gilroy actually had no hand in it, and the script was written by the director, Paul Greengrass, with the help of his editor, Christopher Rouse. The SVM programme using a reference set of the Gilroy written films *Beirut* and

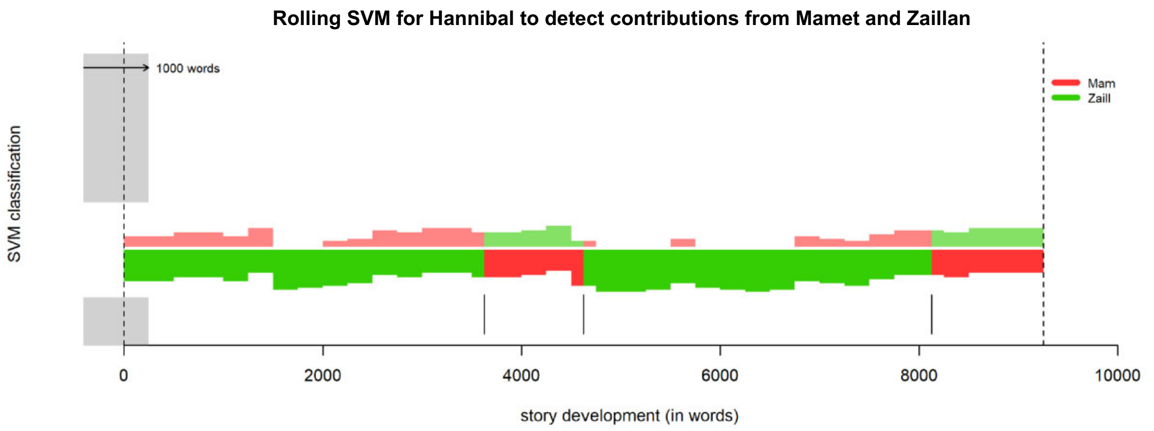


Figure 6.

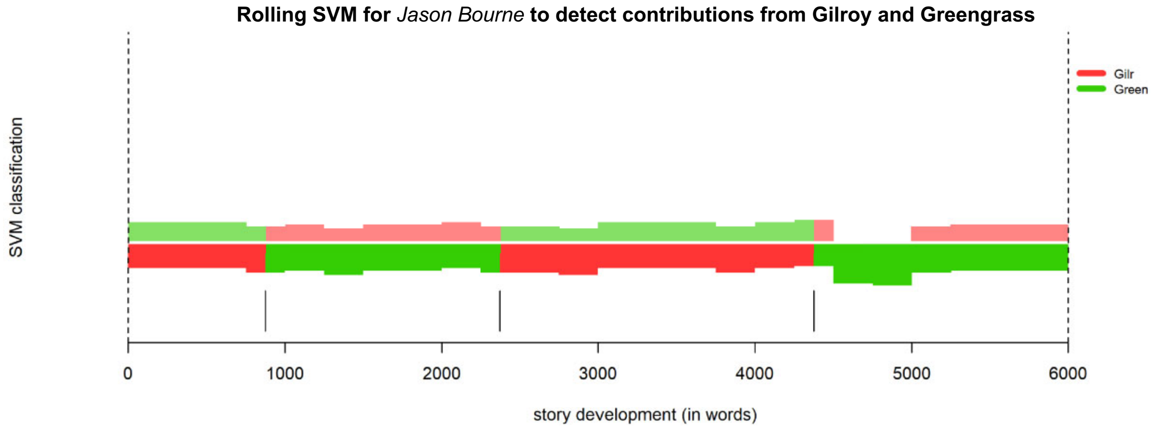


Figure 7.

The Bourne Legacy, together with the Greengrass scripted films *Bloody Sunday* and *United 93*, and working on the test dialogue of *Jason Bourne* (Fig. 8).

The implication that Gilroy wrote half the script is just plain wrong, so this is another failure for the SVM programme. It has been fooled by the intentional recreation by Paul Greengrass and Christopher Rouse of Gilroy's dialogue writing style used in the previous Jason Bourne films.

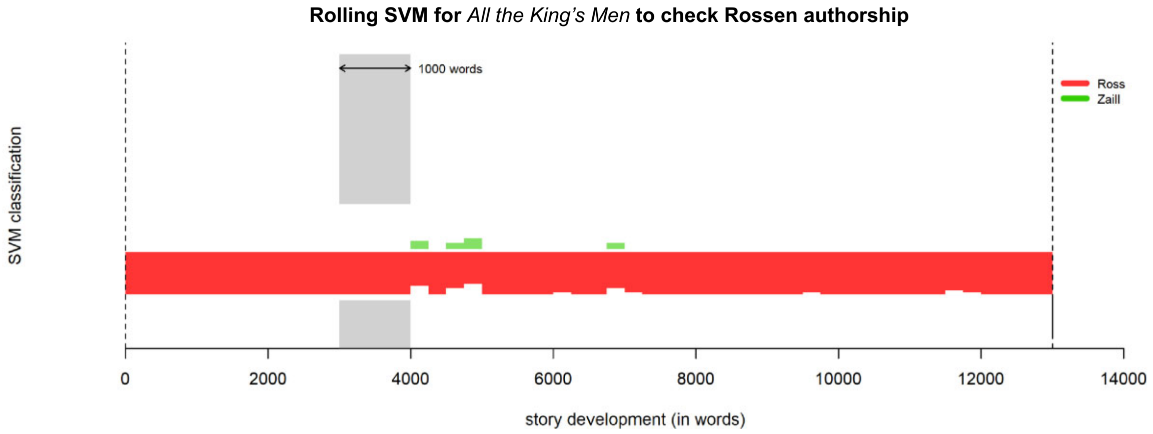
However, a further test of the SVM programme with the Robert Rossen screenplay of *All the King's Men*, based on Robert Penn Warren's novel, which was directed by Rossen and released in 1949, using reference scripts by Robert Rossen (*Alexander the Great*, *Johnny O'Clock*, *A Walk in the Sun*), and by Steve Zaillian (*American Gangster*, *A Civil Action, Hannibal*) who wrote and directed a new version of the Penn Warren novel in 2006 (Fig. 9).

It is just as well that this test shows no trace of Zaillian's writing, as he had not been born in 1949. And the SVM rolling tool is not wrong in this particular case. But it *is* being aided by the 60-year period difference between the Rossen and Zaillian texts, which swamps the fact that they are working from the same material.

5. The zeitgeist speaks

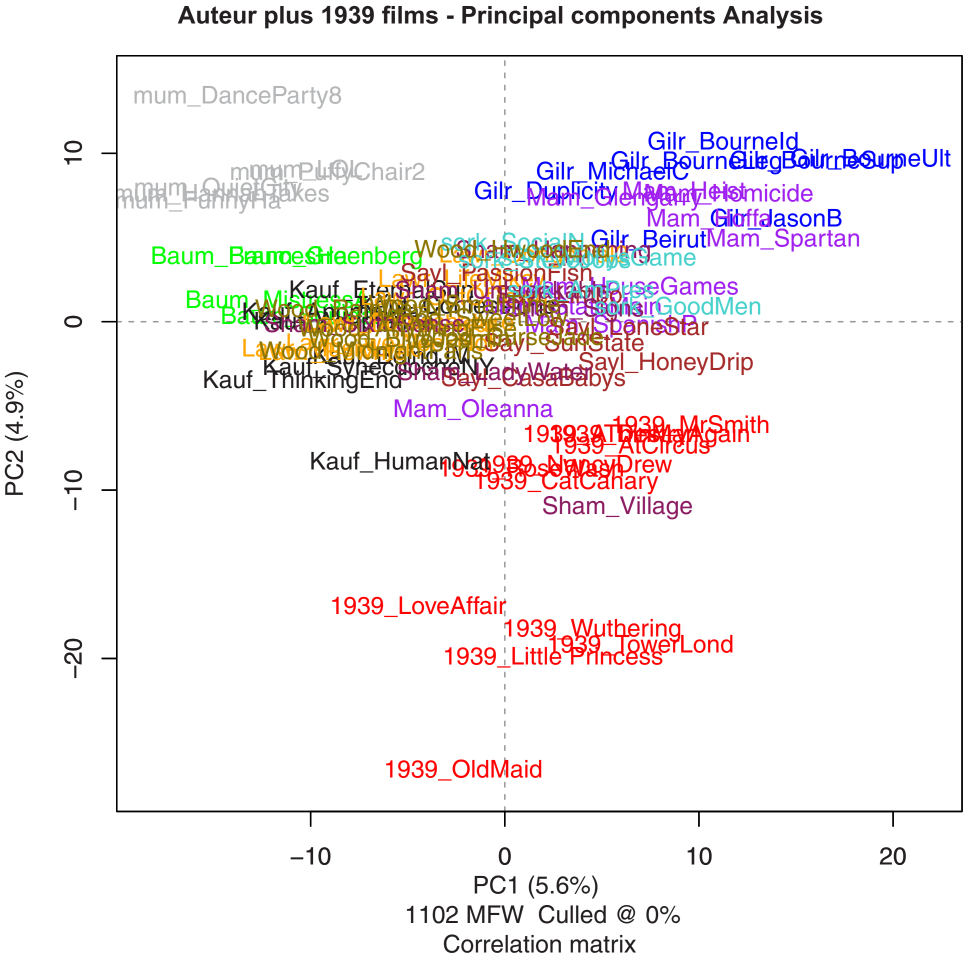
History exerts a pressure on film dialogue, as can be shown by comparing my author corpus with dialogue from thirteen American films released in 1939. Including these films gives the following principal components (Fig. 10).

The 1939 scripts are all in the upper part of the graph, almost separated from the recent films. However, the gangster film *Each Dawn I Die* is almost



Rolling SVM for *All the King's Men* to check Rossen authorship

Figure 8.



Auteur plus 1939 films - Principal components Analysis

Figure 9.

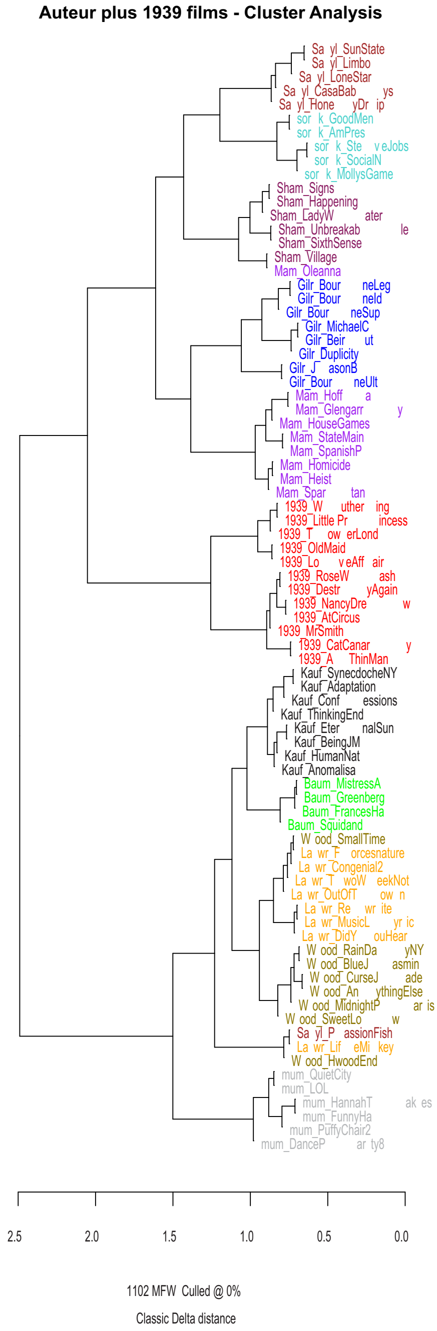


Figure 10.

down into the bottom half of the graph, where the more recent tough stuff lives.

There are obvious changes in conversational vocabulary over six decades in both the real world, and also the filmic world, but the one that stands out for me is the occurrence of the word ‘church’. In the thirteen films of the 1939 sample, ‘church’ is used seventy times, whereas in the seventy-three films of the author corpus, it is only used twenty-seven times. That is, sixteen times less. More importantly, there are vast numbers of obscene words, such as ‘fuck’, in most of the recent films, but none at all in the 1939 films. More importantly, there are many informal contractions of words and phrases that were hardly used in 1939 films, but which are now common. One instance is ‘gonna’ for ‘going to’. It existed in 1939, but parents and teachers beat it out of children in those days. Other 1939 lower class colloquialisms that differentiate recent films from pre-war films include ‘gotta’, ‘wanna’, ‘yeah’, ‘Hi’, and ‘hey’. For the record, the result of using the cluster algorithm on this expanded corpus including the 1939 films is provided in Fig. 11.

6. The doctor will see you now

Some interesting results follow from inserting the films about Dr. Hannibal Lecter’s life and works into my author group of films. The complete group of these Lecter films is provided in Table 2.

And when these are added to a reduced author corpus of drama scripts only, R-stylo cluster analysis is produced (Fig. 12).

Nearly all the Lecter films come together into a cluster between the Sorkin films and the Sayles films, except for *Hanibal Rising*, which is far away among the 1939 films. This film is, of course, unusual in that Thomas Harris wrote the screenplay, as well as the source novel, and Harris has had no previous experience in writing film scripts. Also, not surprisingly, 16 per cent of the dialogue in Harris’ script comes straight from his novel. This compares with the under 1 per cent of his dialogue used by Mamet and Zaillian when writing their scripts from his novel *Hannibal*. A glance at the dialogue text of the film *Hannibal Rising* shows that the low-life characters don’t use as much argot in a way that is usual in other contemporary films, and it is this that moves this film (and also *The Village*) back into the past. I think that the old-fashioned dialogue writing in *The Village* was used intentionally by M. Night Shyamalan, and is part of the plot, but I am not sure that Thomas Harris did it on purpose.

7. Conclusion

The main conclusions I draw from my use of the R-stylo package on the dialogue texts of American feature films are that the dialogues in present day feature films are less individual than might be supposed. The genre of a film does have a limited effect on the dialogue in it, but only in the case of the recent new category of mumblecore films does it have an influence that surpasses that of their individual writers. The world represented in that particular genre of films shows through clearly

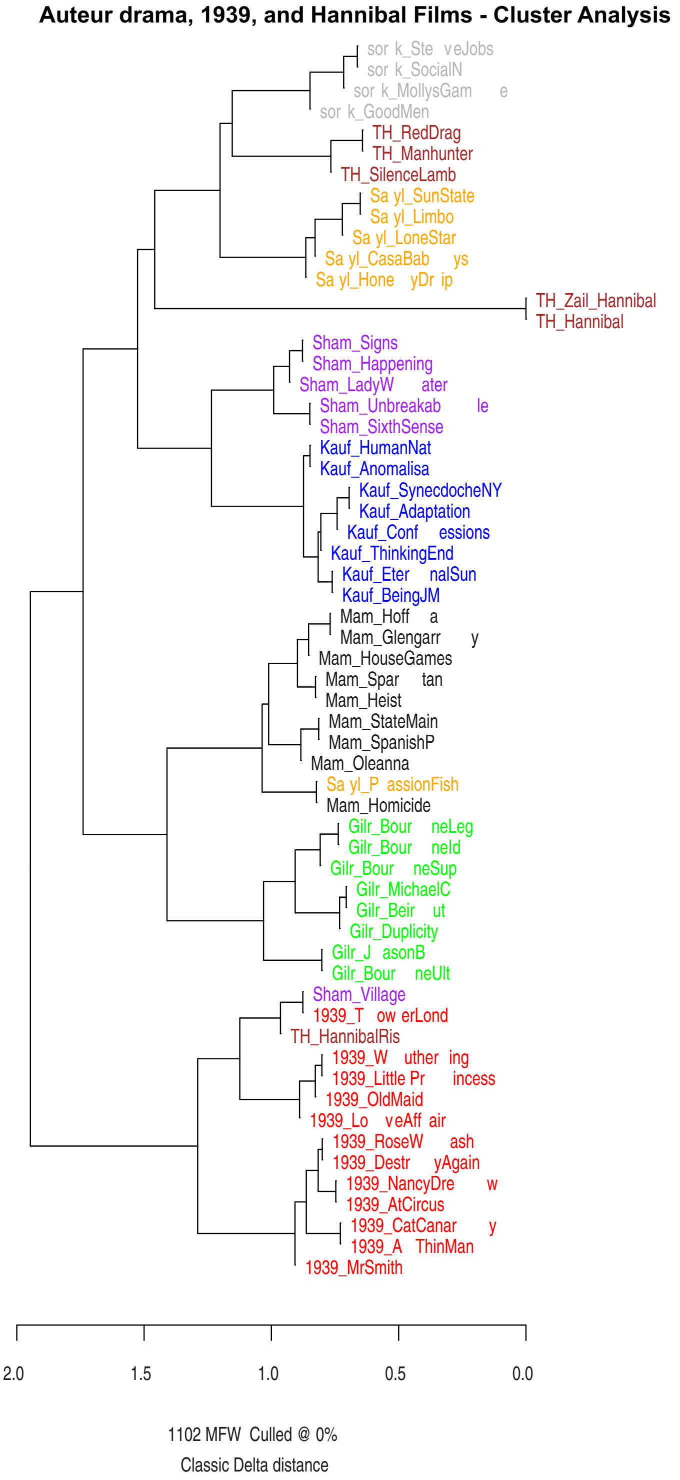

Auteur drama, 1939, and Hannibal Films - Cluster Analysis
sork_SteveJobs
sork_SocialN
sork_MollysGame
sork_GoodMen
TH_RedDrag
TH_Manhunter
TH_SilenceLamb
Sayl_SunState
Sayl_Limbo
Sayl_LoneStar
Sayl_CasaBabys
Sayl_HoneyDrip
TH_Zail_Hannibal
TH_Hannibal
Sham_Signs
Sham_Happening
Sham_LadyWater
Sham_Unbreakable
Sham_SixthSense
Kauf_HumanNat
Kauf_Anomalisa
Kauf_SynecdocheNY
Kauf_Adaptation
Kauf_Confessions
Kauf_ThinkingEnd
Kauf_EternalSun
Kauf_BeingJM
Mam_Hoffa
Mam_Glengarry
Mam_HouseGames
Mam_Spartan
Mam_Heist
Mam_StateMain
Mam_SpanishP
Mam_Oleanna
Sayl_PassionFish
Mam_Homicide
Gilr_BourneLeg
Gilr_BourneId
Gilr_BourneSup
Gilr_MichaelC
Gilr_Beirut
Gilr_Duplicity
Gilr_JasonB
Gilr_BourneUlt
Sham_Village
1939_TowerLond
TH_HannibalRis
1939_Wuthering
1939_LittlePrincess
1939_OldMaid
1939_LoveAffair
1939_RoseWash
1939_DestryAgain
1939_NancyDrew
1939_AtCircus
1939_CatCanary
1939_AThinMan
1939_MrSmith
2.0
1.5
1.0
0.5
0.0
1102 MFW Culled @ 0%
Classic Delta distance

Figure 11.

Table 2.

Dialogue file name	Film title	Year	Director	Script writers
TH_Manunter.txt	Manhunter	1986	Mann, Michael	Mann, Michael
TH_RedDrag.txt	Red Dragon	2002	Ratner, Brett	Tally, Ted
Mam_HANNIBAL.txt	Hannibal	2001	Scott, Ridley	Mamet, David and Zaillian, Steven
Zaill_Hannibal.txt	Hannibal	2001	Scott, Ridley	Mamet, David and Zaillian, Steven
TH_SilenceLamb.txt	Silence of the Lambs, The	1991	Demme, Jonathon	Tally, Ted
TH_HannibalRis.txt	Hannibal Rising	2007	Webber, Peter	Harris, Thomas

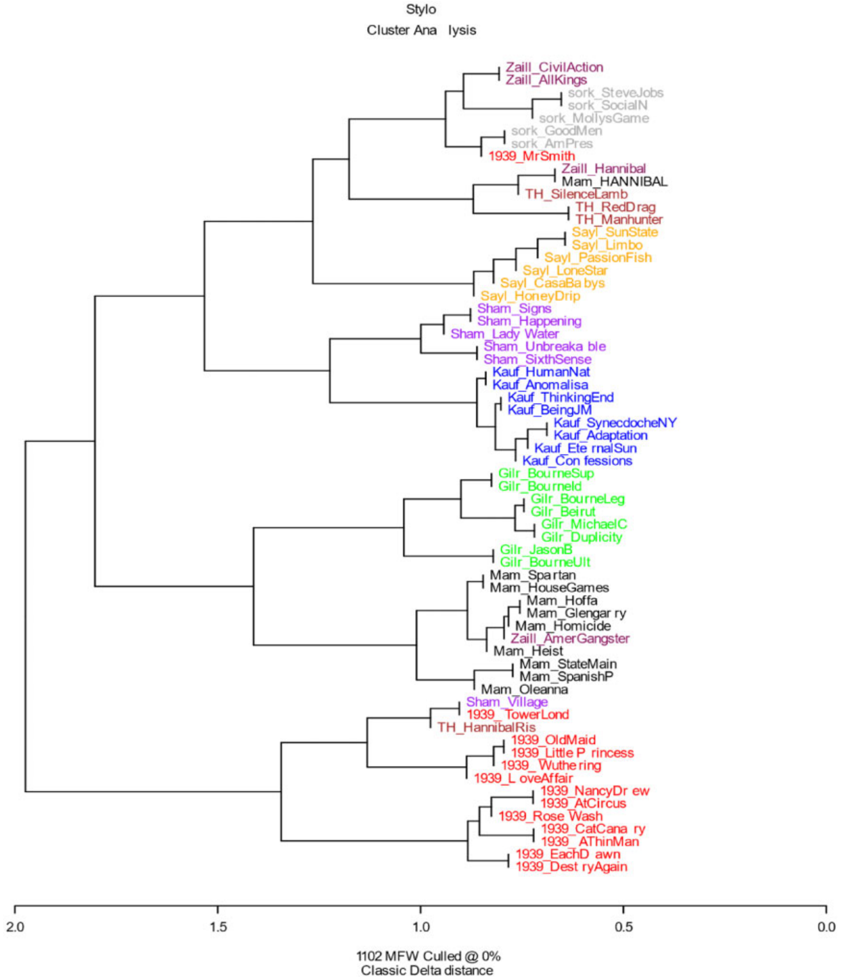


Figure 12.

into the statistics. Another very important general result is that the setting of the range of the most frequent word vectors has a big influence on the results one gets. Yet another surprising result is that using n-gram frequencies rather than single word frequencies makes classification worse rather than better. An obvious supposition about literary style is that part of it resides in the habitual phrases that any writer uses. This effect, if it exists, is not recognizable in film dialogue using the methods in the R-stylo package.

From my tests, it seems that the SVMs rolling classification algorithm can be uncertain about correctly identifying sections of a text written by different writers.

As can be clearly seen in the case of the Bourne films, it is quite possible for new writers to imitate the dialogue style of the preceding films in a series, when writing a new script. In most ordinary films, the dialogue is there to advance the plot, to represent the push and pull between the characters, and literary distinctiveness is not generally helpful in this. To go beyond classification to the reason it works, I have had to use the AntConc programme Keywords tool. Here, the use of negative keywords is particularly useful in finding the particular words that power the R-stylo classification process. The next level beyond word counting requires functional analysis of the relation of the dialogue words to the narrative. That is easy to say, but hard to do.

Authors' contributions

Barry Salt (Writing—Review & Editing)

Conflict of interest statement

None declared.

Funding

None declared.

References

Anthony, L. (2017) *AntConc (Version 3.5.0) [Computer Software]*. Tokyo, Japan: Waseda University. http://www.antlab.sci.waseda.ac.jp/

Buckland, W. (2019) '"Mind our Mouths and Beware Our Talk": Stylometric Analysis of Character Dialogue in The Darjeeling Limited', *Journal of Screenwriting*, **10**: 131–47.

Byszuk, J. (2020) 'The Voices of Doctor Who—How Stylometry Can be Useful in Revealing New Information About TV Series', *Digital Humanities Quarterly*, **14**.

Eder, M., Rybicki, J., and Kestemont, M. (2016) 'Stylometry with R: A Package for Computational Text Analysis', *R Journal*, **8**: 107–21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Evert, S., *et al.* (2017) 'Understanding and Explaining Delta Measures for Authorship Attribution', *Digital Scholarship in the Humanities*, **32**: ii4–16.

Hołobut, A., and Rybicki, J. (2020) 'The Stylometry of Film Dialogue: Pros and Pitfalls', *Digital Humanities Quarterly* **14**.