DHQ: Digital Humanities Quarterly
2020
Volume 14 Number 4

The Stylometry of Film Dialogue: Pros and Pitfalls

Agata Hołobut <agata_dot_holobut_at_uj_dot_edu_dot_p>, Jagiellonian University in Kraków
Jan Rybicki <jan_dot_rybicki_at_uj_dot_edu_dot_pl>, Jagiellonian University in Kraków

Abstract

We examine film dialogue with quantitative textual analysis (stylometry, sentiment analysis, distant reading). Working with transcribed dialogue in almost 300 productions, we explore the complex way in which most-frequent-words-based stylometry and lexicon-based sentiment analysis produce patterns of similarity and difference between screenwriters and/or a priori IMDB-defined genres. In fact, some of our results show that counting and comparing very frequent word lists reveals further similarities: of theme, implied audience, stylistic patternings. The results are encouraging enough to suggest that such quantitative approach to film dialogue may become a welcome addition to the arsenal of film studies methodology.

Film dialogue

Although dialogue has been integral to film structure since the introduction of sound in the 1920s and used in silent film intertitles at least since 1904 [Kozloff 2000, 20] [Pitera 1979], it has garnered relatively little attention from film scholars. Stylistic immaturity and technical imperfections of the first talkies made early theorists, such as Rudolf Arnheim, Sergei Eisenstein or Siegfried Krakauer bemoan the loss of artistry inherent to silent films [Kozloff 2000, 7] and encouraged others to describe the compositional complexity of the new invention in contrast to its silent, though hardly wordless, predecessor [Hendrykowski 1999]. Still, as Sarah Kozloff argues in her ground-breaking monograph *Overhearing Film Dialogue* (2000), Western film scholarship has long been marked by " verbo-immunity" or even "verbophobia" in its approach to cinematic art, privileging the visual over the auditory and the non-verbal over the verbal: 1

> Although what the characters say, how they actually say it, and how the dialogue is integrated with the rest of the cinematic techniques are crucial to our experience and understanding of every film since the coming of sound, for the most part analysts incorporate the information provided by a film's dialogue and overlook the dialogue as signifier. Canonical textbooks on film aesthetics devote pages and pages to editing and cinematography but rarely mention dialogue. Visual analysis requires mastery of a recondite vocabulary and trained attentiveness; dialogue has been perceived as too transparent, too simple to need study [Kozloff 2000, 6]

This apparent disregard may have stemmed from a number of reasons, one of them being a deeply-grounded belief in the autonomy of various art forms. Scholars intent on presenting film as primarily a visual medium deemed dialogue negligible, because they considered it too closely related to other forms of expression, such as stage drama or prose-writing and hence unspecific to cinema [Piazza 2011, 8]. Film talk, however, differs considerably from verbal exchanges intended for the page or for the stage, as it is always affected by performers' interpretive choices and the "simultaneous signification of camerawork/mise-en-scene/editing" [Kozloff 2000, 16]. It is an integral element of fiction film and, as Sarah Kozloff argues, it performs unique narrative functions: (1) it anchors the diegesis and the characters, (2) it explains causal links between events presented onscreen, (3) it enacts narrative events, such as professing feelings, setting up deadlines or passing sentences, (4) it reveals the characters' personalities, (5) it controls the viewers' emotional and axiological reactions and (6) it helps create the impression of realism. It also plays an important aesthetic role, exploiting the artistic potential of language, conveying ideological messages and giving stars an opportunity to 2

show off their skills [Kozloff 2000, 33–4]. Since the publication of Kozloff's book a number of film scholars have pursued her line of study, tracing the evolution of screenwriting styles in America, emphasizing the requirements of particular genres (e.g. screwball comedy, science fiction, gangster film, western), investigating the idiom of selected auteurs (Cassavetes, Hawks, Welles, Sturges) and typical verbal representations of different social groups. The most important contribution to the field has been arguably the collected volume *Film Dialogue* edited by Jeff Jaeckle (2013).

3 Telecinematic dialogue has also received in-depth attention from linguists working within the fields of stylistics, pragmatics, sociolinguistics and discourse analysis (e.g. Bobrowski 2015; Piazza 2011; Piazza 2011a; Richardson 2010), as well as audiovisual translation experts interested in different techniques, strategies and modes of language transfer. Most research has been driven by qualitative approaches, yet quantitative measures have also been adopted by scholars using monolingual, as well as multilingual and multimedia corpora to investigate the relationship of scripted speech to real-life conversation, to examine the stylistics of television dialogue and the consistency of character identity revealed through language (e.g. Bednarek 2010; 2011; 2018; Quaglio 2009; Zago 2016; Zago 2018), to explore verbal discourse in different film genres [Veirano Pinto 2014], to analyze the sociolinguistic and pragmatic idiosyncrasies of interlingual dubbing (e.g. Freddi 2009; Freddi 2013) and its relationship to spontaneous speech (e.g. Heiss 2005; Romero Fresco 2009) or to study interlingual translation in the visual context (e.g. Valentini 2006; 2008; 2009).[1]

4 What many of these approaches demonstrate is that film and television genres are characterized by distinctive language patterns; dialogues in screwball comedy, melodrama, western, sitcom or police procedural are endowed with genre-specific functions and genre-specific stylistic, pragmatic and rhetorical features (such as the literary stylization of classic novel adaptations, the use of slang and jargon in gangster films or the omnipresence of techno-bubble in science fiction). As Kozloff explains:

> Partially, they [these verbal genre conventions] are motivated by the subject matter. Screenwriters are always concerned that dialogue be appropriate to characters' social backgrounds, and thus "realistic" (...). Partially, films are clearly copying preexisting expectations created by other forms of representation (...). And partially, I believe, dialogue patterns are related to the underlying gender dynamics of each genre: whether the genre is primarily addressed to male or female viewers and how each genre treats its male and female characters are crucial factors in its use of language. [Kozloff 2000, 137]

What some of the qualitative analyses (e.g. Jaeckle 2013; Miławska-Ratajczak 2019) also show is famous writer-directors' signature verbal style. All these insights have led us to believe that these gene- and *auteur* - specific verbal patterns, described by scholars and noticed by average viewers, may be further explored by stylometric measures.

Research objectives

5 In our research, we decided to adopt a variety of procedures, using quantitative textual analysis tools derived from stylometry (a.k.a. computational stylistics) and from automated sentiment analysis, a significant subdiscipline of text mining. As these approaches rely on computational work, they allow to analyze large sets of texts in ways similar to distant reading [Moretti 2007]Moretti 2007). Used on diverse corpora of literary and non-literary texts, they have already proven helpful in plagiarism detection, authorship attribution, as well as genre- and diachrony research, as will be explained below. Their application to film dialogue, however, has so far been limited [McKie 2014] [Hołobut et al. 2016] [Hołobut et al. 2017] [Van Zyl and Botha 2016] #byszuk2017 [Hołobut and Woźniak 2017] [Hołobut and Rybicki 2018] and it certainly merits further attention for a nuber of reasons. Firstly, stylometric analysis can complement other advancements in the realm of cinemetrics [Baxter 2014a] [Baxter 2014b] by bringing neglected cinematic speech to film scholars' attention. Secondly, it provides yet another level of potential comparison between film dialogue and other dialogic genres, stage drama in particular. Both theatre and screen plays combine literary and performative dimensions. Stage drama has already been analyzed in its literary dimension by qualitative and quantitative means, with obvious precedence to, often, heated arguments on Shakespearean authorship attribution (to name but a few of the latest: Vickers 2002; Craig and Kinney 2009; Vickers 2011, Rudman 2016). Equally obviously, this leaves the space open for the stylometry of filmic speech. Thirdly, stylometry and automated sentiment analysis can open new avenues for

qualitative research, revealing “patterns overlooked in close reading” [Hayles 2010, 75]. That is why in our research we intended to test the usefulness of the mentioned tools in identifying the potential stylometric similarities or differences between film fictional dialogues, revealing regularities typical of a given genre, theme, historical context, writer, director- and/ or film cycle. In our previous studies, we focused exclusively on historical films and literary adaptations, combining quantitative stylometric with qualitative stylistic and pragmalinguistic analysis of filmic speech (cf. Hołobut 2017b). This time, we collected a multigeneric corpus of transcribed dialogue lists to 278 Anglophone productions, spanning over eighty-four years of cinematic art (for a complete list of productions included, see Appendix 1), running a pilot quantitative analysis, to be potentially elaborated by qualitative means.[2] We labelled each dialogue list with the name of the screenwriter and the date of release and divided them into six rough generic/thematic categories following IMDb listings: chick flick, vampire, superhero (as distinguished by the dominant theme); romance, thriller, action (as distinguished by the dominant genre). This division is admittedly provisory and tentative for several reasons:

- film genres are in themselves fuzzy categories, as suggested by film scholars and experts in media semiotics (cf. Chandler 1997; Altman 1999; Grant 2003; Bondebjerg 2015);
- most productions included in the corpus evade easy categorization, being best ascribed to a number of major film genres simultaneously and fitting better more precise sub-generic distinctions (for example, IMDb classifies many superhero films as action/adventure/fantasy or action/adventure/science-fiction hybrids, yet some productions, such as *Deadpool* or *Thor: Ragnarok* have an additional comedic twist, while others, like Christopher Nolan’s visions of *Batman*, qualify as thriller/crime);
- thematic groupings obviously intersect with major generic ones; for example, some vampire films reveal horror and some – dominant romance features, as do a number of releases stereotypically categorized by IMDb viewers as “chick flicks”, although many of the latter belong to the sub-genre of romantic comedy; that is why we have decided to keep this dubious appellation;
- some productions included in the corpus conform better to *auteur* than genre classifications imposed on them, hence their status may be considered dubious.

Still, we did not consider these complexities and inconsistencies in categorization as detrimental to our stylometric analysis. On the contrary, we hoped that by inspecting in detail the correlations between particular film scripts, we may detect certain genre-, theme-, diachrony- or author-related regularities which would allow us to revise or reshuffle the adopted categories. We therefore subjected the corpus to a number of quantitative procedures, the origin and methodology of which will be explained below. 6

Introducing stylometry and sentiment analysis.

Quantitative research on literary language has been developing at least since John Burrows studied character idiolects in Jane Austen, showing that her major characters with similar functions in different novels tend to use the most common words of their lexicon in similar ways (Burrows 1989). This was an early application of multivariate analysis of word frequencies that had already been around since Mosteller and Wallace showed that authors differ in this respect strongly enough to allow attribution (1964). We now know that minute differences in individual usage of very frequent and seemingly very insignificant words such as “the,” “a,” “in,” “and,” “such ” and “as” – when treated with appropriate statistical tools – are enough to tell authors apart; or, more precisely, to point out which of several candidate writers is the real author of an anonymous text. These authorship attribution methods, sometimes still called non-standard authorship attribution methods, have in fact become a standard approach whenever the hand that held the pen (or that tapped on the keyboard) is obscured for any reason: plagiarism, publishers’ promotional policy or mere authorial whims. They have been recently used to check that Harper Lee did in fact write both *To Kill a Mockingbird* and *Go Set a Watchman* [Choiński et al. 2019]; that Robert Galbraith and J.K. Rowling are one and the same person [Juola 2015]; and that Elena Ferrante, the bestselling yet non-existent Italian woman, writes suspiciously like the real Italian man, Domenico Starnone [Tuzzi and Cortelazzo 2018]. 7

But many other stylometric experiments in method have also shown that what stylometrists call the authorial signal (meaning nothing more, in fact, than the authors’ individual idiosyncrasies in word usage that can be made evident through statistics) is just one of the classifications that can be made on any sets of texts. Not only texts by the same 8

authors are similar in this respect. On another level, texts written by authors writing over several literary periods tend to exhibit a chronological "evolution" [Burrows 1994] [Rybicki 2017a], and while this can be easily explained by simple evolution of language over centuries, it is also a fact that many writers (including Shakespeare, Dickens, James and Conrad) exhibit a chronological progression within their own *oeuvre* [Hoover 2007] [Rybicki 2017b]. Works of the same writer in two different genres exhibit different stylometric traits; this has already been shown for Shakespeare and Molière [Rybicki and Eder 2011] [Rybicki 2017b]; what is more, genre can be a factor in large-scale comparisons of authorial word usage [Mealand 1999]. Interestingly, there is much less of this "stylometric universal" in translation: some translators seem to exhibit their own and stable "stylometry" while others have a different word usage for each author they translate [Burrows 2002], and translations in a large set tend to group more often by their original authors than by translators [Rybicki 2012]; nevertheless, traces of translators can be traced within a collaborative work [Rybicki and Heydel 2013].

9 That such results are stable, reproducible and robust in terms of statistics is obviously due to the fact that the features of text used for counting are very simple and, at the same time, very numerous. The first hundred most frequent word-types in any collection of texts (which usually account for as much as a half of every text) contains little more than various function words; "meaningful" words – the usual stuff of literary study – are still a minority in the first thousand. This could suggest that stylometry based on most frequent word counts misses the crucial element of literary texts, meaning, or *what*, and focuses on style, or *how*; but, on the one hand, word counts alone are hardly style, and, on the other, Hough insisted that style is just "an aspect of meaning" [Hough 1969, 8]. Attempts to bridge the gap between stylistics and stylometry continue (c.f. Hermann 2015), but there is still much work to do; one explanation of the power of the most frequent words is that they "do not function as discrete entities. Since they gain their full meaning only through the different sorts of relationship they form with each other, they can be seen as markers of those relationships and, accordingly, of everything that those relationships entail" [McKenna et al. 1999].

10 The reason why most-frequent-word stylometry, and not some other types of textual analysis, has been such a powerful newcomer into (digital) literary studies is perhaps just one of the many examples that computer-based methods – especially unsupervised ones – are notoriously fallible when dealing with the semantics of the literary text. This is very visible in the tribulations of sentiment analysis, which has been a focus of much quantitative textual research at least since Stone et al. made it one of the objects of their study [Stone et al. 1966]. It must be said that while sentiment analysis can be of use, and is often used, in large-scale browsing of non-literary texts and/or to gauge the overall mood of large sections of the Internet, news etc., its application to artistic writing is somewhat more problematic due to the features of the literary text that define it as literary. Ambiguity, irony, metaphor or even simple negation all cause problems in disambiguating the sentiment/emotional "value" of words and phrases.

11 Despite this problem, the presence of some interest in sentiment in literature can be attested to by the numerous software packages that try to make it possible. One of the most recent, Jockers's *syuzhet* package [Jockers 2016] for R [R Core Team 2014], employs a number of lexicons of sentiment-related terms to trace curves of negative and positive emotions in single novels, associating lexical items found in individual texts to emotive categories. Among these, the NRC Word-Emotion Association Lexicon [Mohammad and Turney 2010] has already been used to trace the evolution of sentiment across larger literary corpora and follows Plutchik's concept of eight "basic emotions" (anger, disgust, fear, sadness and anticipation, joy, surprise, trust; #plutchik1991; alternatively, it also identifies two "sentiments," respectively: negative and positive [Mohammad 2011]. A similar approach has been adopted by Acerbi et al. for English novels of the 20th century, who observed a general decrease, with time, of terms of emotion by tracing the evolution of terms related to some general categories such as fear or joy [Acerbi et al. 2013]. An even more extensive chronological study performed on three hundred years of English and Polish fiction showed a steady growth of negative sentiments and "negative" emotions with time [Rybicki 2018].

12 The main drawback sentiment analysis has is that most of the readily-available methods use a lexicon approach, where individual words are assigned a sentiment/emotion value. The NRC Word-Emotion Association Lexicon, which we used here, was a major crowdsourcing venture using Amazon's Mechanical Turk, where numerous users contributed to produce a statistics-based system of evaluating and quantifying emotion; yet even this tool may cause a variety of problems. Consider the following examples from the NRC Lexicon (Table 1):

emotion/sentiment	*baboon*	*dentistry*	*polish*
anger	0	0	0
anticipation	0	0	0
disgust	1	0	0
fear	0	1	0
joy	0	0	0
sadness	0	0	0
surprise	0	0	0
trust	0	0	0
negative	1	0	0
positive	0	0	1

Table 1. Example of emotional valence/sentiment values in the NRC Lexicon. Grey background applied for items discussed in the text.

The “baboon” example shows how a lexical item is assigned a value of “disgust” and “negative” despite the fact that, in a text about theriology or even Africa, these associations should be made perfectly neutral; the Lexicon also seems to ignore the fact that presence of baboons may just as plausibly trigger “fear,” or that one may thus address a human being in “anger” as well as in “disgust.” The example of “dentistry” seems to take into consideration the most stereotypical reaction to that noble profession; more importantly, the choice of this very rare word and the absence, in the Lexicon, of the much more frequent word “dentist” is an illustration of the numerous issues of representativeness in this approach. The final example, that of “polish,” shows an even more “mechanical” problem and the extent to which mere linguistic ambiguity may be a distorting factor: this word, when capitalized, may well require a very different sentiment/emotion value. More generally, distribution of emotions and sentiments within the literary text is uneven; smoothing any results may be risky, as evidenced by the difficulties with this problem in the *syuzhet* package (cf. Swafford 2015, Jockers 2017). 13

Despite these pitfalls, our attempt to measure sentiment and emotion in film dialogue seems worthwhile. After all, while the lexicon-based method may be wrong at times, it is important that it is wrong in a consistent way. Much as the tool is imprecise, there is no reason to suppose different texts will be treated in any different way. 14

Methods

For most-frequent-word-based stylometric analysis, the texts were treated with the well-established Delta procedure [Burrows 2002] implemented in the *stylo* package for R [Eder et al. 2016], the statistical programming environment [R Core Team 2014]. The entire ensemble of texts in electronic form serves as input for *stylo*, which then tokenizes the text (separates it into words). The word tokens are then counted in the whole set to produce a descending frequency list of word-types, so that the 100 or more of these may be identified. A given number of those very frequent words (usually, from 100 to 1000) is used as the features of the analysis: their frequency is now counted in each individual text to create a frequency table; its columns are treated as sequences of numbers that can be compared using an appropriate distance (dissimilarity) measure. Our study uses the so-called Cosine Delta distance, which has been shown to work best in authorship attribution [Evert et al. 2017]. It is based on cosine similarity of the angle between two vectors, x = z(T) and y = z(T1), and is calculated according to the formula: $\cos \propto = \frac{\sum_{i=1}^{n_s} x_i y_i}{\sqrt{(\sum_{i=1}^{n_s} x_i^2)}\sqrt{(\sum_{i=1}^{n_s} y_i^2)}}$ where ns is the number of most-frequent word-types used in the analysis, while z(T) and z(T1) are z-score values for the frequency of a word in, respectively, texts T and T1. Z-scores are calculated with the usual formula: $z(T) = \frac{f_s(T) - \mu_s}{\sigma_s}$ where fs(T) denotes relative frequency of word s in text T, μs is mean frequency of word s in the entire set of texts, and σs is standard deviation of frequency of word s in the same set [Smith and Aldridge 2011]. 15

This produces a distance matrix: a square table of “distances,” or degrees of dissimilarity, between each pair of texts – 16

the smaller the value, the more similar the two texts are. Usually, at least in literary texts of some length, the smallest values are those between texts written by the same author; our studies have shown that this is a much less marked phenomenon in film dialogue, and even less so in TV series [Hołobut et al. 2016] [Hołobut et al. 2017]. While this is already the kind of output that may be examined, analyses of larger numbers of texts require visualization of the patterns of similarity and difference between the texts, which can be achieved with further statistical procedures. We use Ward's hierarchical cluster analysis to draw tree diagrams that bring together the nearest neighbors (texts most similar to each other); usually, for greater reliability of the results, this is repeated for different numbers of most-frequent word types (in our study, for 100, 200, 300, … 1000) most frequent words, and the results are further processed in consensus network analysis [Eder 2017] in *Gephi* [Bastian et al. 2009].

While the stylometric part of the analysis was performed on all grammatical word forms, sentiment analysis requires that the texts be lemmatized, or converted to their basic grammatical forms. This was done with the standard automatic tool, Treetagger [Schmid 1994]. Sentiments and emotions (as defined by Mohammad 2011) were counted using the "NRC Lexicon" within the *syuzhet* package [Jockers 2016]. Counts of emotions were summed for each text and proportions between positive and negative sentiments established for each text. Graphs were produced to visualize the results; the Lexicon's eight emotions were treated similarly but relative to individual text lengths.

Results

Figure 1 presents a network analysis performed with the above-mentioned method on the full set of film dialogues.

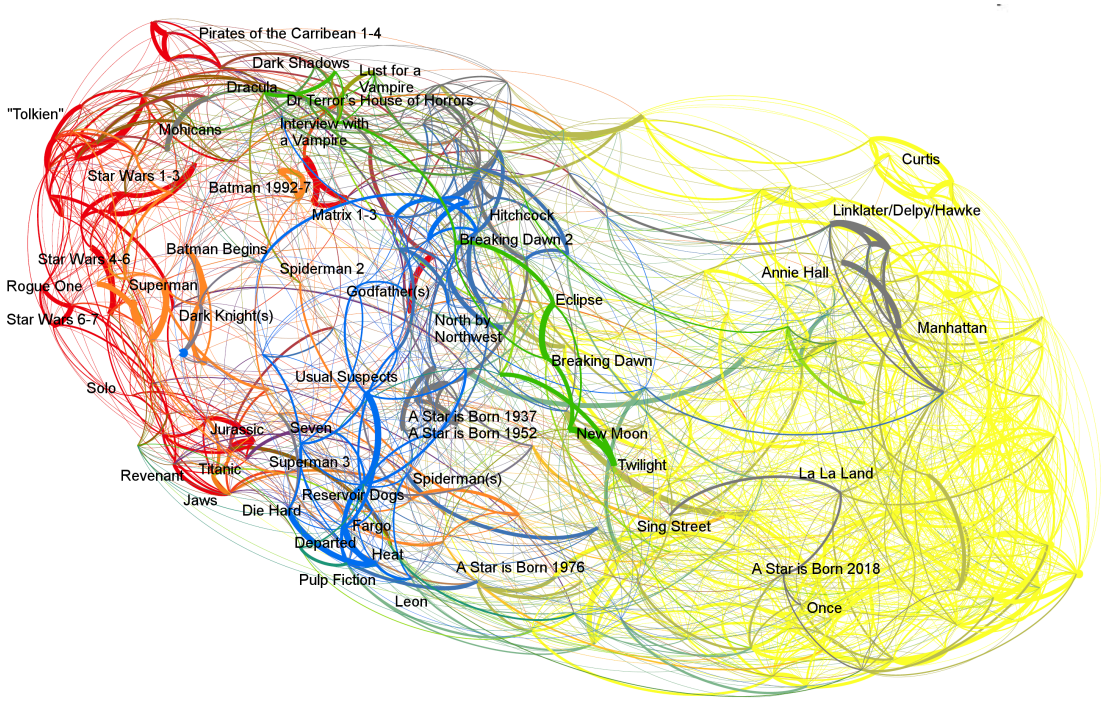

Figure 1. Network analysis of the entire set of film dialogues, color-coded by (principal) genre: action/adventure, red; chick flick, yellow; romance, grey; superhero, orange; thriller, blue and vampire, green).

As can be seen, clustering by generic-thematic grouping is especially successful for thrillers (blue), with particularly strong connections visible among films sharing sub-generic properties. For instance, Brian Singer's crime/drama/thriller *The Usual Suspects* (1995) written by Christopher McQuarrie shows on the one hand strong connections to David Fincher's crime/mystery drama *Seven* (1995) and on the other – to Quentin Tarantino's *Reservoir Dogs* (1993), the Coen brothers' *Fargo* (2006), and Martin Scorsese's *The Departed* (2006), thus demonstrating close stylometric affinities between different productions combining elements of crime, drama and thriller and raising interesting questions about the sensitivity of stylometric tools to the exponents of low register (possibly reflected in the grammatical simplicity of utterances affecting the frequency of function words) and the pragmatics of deduction and threat (cf. Kozloff 2000:

220) typical of crime films.

John McTiernan's *Die Hard* (1988), classified by IMDb as action/thriller, shows strong ties to both genres, being much closer to other action productions (such as Steven Spielberg's *Jaws*, 1975), but at the same time showing surprisingly tight stylometric bonds to Tarantino's *Pulp Fiction* (1994), Michael Mann's *Heat* (1995) and Luc Besson's *Leon* (1994), classified as crime/drama/thrillers. 20

A separate sub-area on our map is occupied by Alfred Hitchcock's masterpieces intertwined with classic noir productions. *Dial M for Murder* (crime, thriller 1954), written by a British playwright Frederick Knott, links strongly to Hitchcock's *Psycho* (an adaptation of Robert Bloch's novel; horror, mystery, 1960) and *Rebecca* (an adaptation of Daphne du Maurier's novel; drama, mystery, romance,1940), as well as to Carol Reed's *The Third Man*, written by Graham Greene (film-noir, mystery, thriller, 1949). *Rebecca* is closely related to his *Psycho* and *Vertigo* (an adaptation of Boileau-Narcejac's novel; mystery, romance, thriller, 1958), as well as to Billy Wilder's film noir *Double Indemnity* co-authored by the director and Raymond Chandler (crime, drama, film noir, 1944). At the same time, it connects strongly to David Lean's romance *Brief Encounter* and other representatives of the romance genre, such as Sydney Pollack's *Out of Africa* (1985) or Edward Zwick's *Legends of the Fall* (1994). Hitchcock's *Rear Window* (mystery, thriller 1954) is bonded with *Psycho*, as well as *The Third Man* and *Double Indemnity*. They constitute quite a tight-knit area in the diagram, allowing the diachronic, generic and *auteur* director signals to collude and encouraging further qualitative research into how Hitchcock's belief in the subservience of dialogue to "pure cinema" may have affected his collaborator's screenwriting style and reflected earlier film-making conventions. The only Hitchcock production that does not blend in with the other scripts analyzed is *North by Northwest* (adventure, mystery, thriller, 1959), written by Ernest Lehman and apparently intended by the writer as "the Hitchcock picture to end all Hitchcock pictures" (qtd. in Freedman 2015: 36). The dialogue lines demonstrate closer links to Roman Polański's neo-noir *Chinatown* (1974), and Jacques Tourneur's noir *Out of the Past* (1947) than to Hitchcock's other films. These neighbor with two representatives of the gangster genre in our corpus, Francis Ford Coppola's *Godfather* (1972, 1974). An in-depth inspection of this grouping suggests that further sub-division into mystery, gangster and film noir might reveal additional patterning in dialogue techniques. This supports Sarah Kozloff's "close-reading" observations on the functional and pragma-stylistic distinction to be made between noirs and gangster films, the former featuring characters who are "middle-class," solitary, "cynical and deliberately 'hardboiled'" and the latter featuring self-educated "groups of confederates" [Kozloff 2000, 203]. It is interesting, however, that films directed by Alfred Hitchcock (and written by different acclaimed authors) display stronger stylometric bonds than those both written and directed by other acclaimed *auteurs*, such as Quentin Tarantino or the Coen brothers. 21

Another grouping that forms a successful cluster is the thematic category "chick flick". As listed by IMDb users, it is dominated by romcoms, but it also contains romance drama and family drama, films that demonstrably tap into the realm of human emotions and range from humor to pathos. They consistently occupy the right-hand side of the map and intermingle with productions categorized as "romances," a grouping that clearly overlaps with the one discussed. Thus, the grey "intrusions" into the yellow-dominated area on the diagram are all *auteur* romances: John Carney's romantic musical dramas *Once* (2007) and *Sing Street* (2016), Damien Chazelle's *La La Land* (2016) and Bradley Cooper's most recent remake of *A Star is Born* (2018), as well as critically acclaimed verbal masterpieces: Richard Linklater's romantic drama trilogy *Before Sunrise* (1995), *Before Sunset* (2004) and *Before Midnight* (2013) co-created with performers Julie Delpy and Ethan Hawke, Woody Allen's *Manhattan* (1979) and *Annie Hall* (1977) and Spike Jonze's *Her* (2013), the last two presented with the Academy Award for Best Original Screenplay and clearly devoid of melodramatic traces typical of many representatives of the "romance" category. 22

Another interesting regularity to be observed in this area of the diagram is the clustering of films written or co-authored by Richard Curtis: *Four Weddings and a Funeral* (1994), the two Bridget Jones films (2001, 2004), *Notting Hill* (1999), *Love, Actually* (2003) and *About Time* (2013). The only aberration on the romcom/romance map is James Cameron's *Titanic* (1997), hidden in the depths of action/adventure/thriller, in the neighborhood of Spielberg's *Jaws* (1975) and Inarritu's *Revenant* (2015). It seems that James Cameron's unsentimental provenance and interest in non-romantic genres may have left stylometric traces in his dialogue technique. 23

Vampire films (green), by contrast, reveal their individual and complex generic affinities. The *Twilight* saga, spread around the central part of the diagram, shows marked stylometric affinities across the episodes, yet it clearly demonstrates the gradual loss of romantic undertone, with the first two episodes: *Twilight* (2008) and *New Moon* (2009) classified as fantasy/drama/romance showing bonds with romantic comedies and romances and the more recent parts of the series uniting with classic thriller productions. Another area is occupied by horror rather than fantasy representations of the vampire theme: Coppola's *Dracula* (1992), Burton's *Dark Shadows* (2012), Jordan's *Interview with the Vampire* (1994) and dated horrors: Sangster's *Lust for a Vampire* (1971) and Francis's *Dr Terror's House of Horrors* (1965). These regularities may, indeed, suggest that stylometric measures based on most frequent word frequencies are sensitive to the changing functionalities of filmic speech, romance depending more on the explicit expression of emotion [Kozloff 2000, 249] and horror – on the implicit build-up of suspense. 24

As for other groupings, action/adventure films, including superhero productions, all cluster to the left-hand side of the diagram. What merits attention is the stylometric consistency of big franchises, with particular episodes clustering regardless of their authorial parentage. Some blockbuster series can be attributed to the same screenwriter throughout, allowing the authorial and thematic signals to coincide: for instance, the *Pirates of the Caribbean* swashbucklers (written by Ted Elliott and Terry Rossio), the *Matrix* trilogy (by Lana and Lilly Wachowski), the *Lord of the Rings* and the *Hobbit* series, which intertwine in our analysis (written by Fran Walsh, Philippa Boyens, Peter Jackson). 25

The Batman series is clearly divided into two main areas: a cluster of Christopher Nolan's most recent productions: *Batman Begins* (2005), *The Dark Knight* (2008) and *The Dark Knight Rises* (2012) and another cluster of three older episodes: *Batman Returns* (1992) directed by Tim Burton as well as *Batman Forever* (1995) and *Batman and Robin* (1997) directed by Joel Schumacher. Tim Burton's original *Batman* (1989), however, occupies a distant location on the map, tied closely to thriller films, and showing little stylometric relation to the following pictures, possibly inspired more by the gangster/crime genre to which it visually alludes. 26

The *Star Wars* franchise is, unsurprisingly, separated into three areas. The two oldest episodes: *Star Wars* (1977) and *The Empire Strikes Back* (1980) are united with a strong stylometric bond. Lawrence Kasdan and George Lucas's *Return of the Jedi* (1983) connects them to the more recent trilogy: *The Phantom Menace* (1999), *Attack of the Clones* (2002) and *Revenge of the Sith* (2005), written by George Lucas. They neighbour closely with Lucas's other productions, namely Steven Spielberg's *Indiana Jones and the Raiders of the Lost Ark* (written by Kasdan, 1982) and *Indiana Jones and the Last Crusade* (written by Boam, 1989). The most recent *Star Wars* incarnations produced by Disney: *The Force Awakens* (2015), *The Last Jedi* (2017) and *Rogue One* (2016) form a separate cluster, showing a different dialogue technique. Interestingly, Lawrence Kasdan's most recent addition to the series, *Solo: A Star Wars Story* (2018) does not reveal any stylometric affinities to the remaining pictures; indeed *Solo* does seem to fly solo. 27

As for other series, *Jurassic Park* and *Jurassic World* episodes show stylometric links. Richard Donner's *Superman* (1978) bonds with its recent remake, *Superman Returns* (2006) are Richard Lester's *Superman II* (1980). Significantly, the comedic third episode, *Superman III* (1983) is separated from the remaining ones. Other remakes, such as *The Last of the Mohicans* (1937, 1992) and two out of the four versions of *A Star is Born* (1937; 1954) also demonstrate marked stylometric similarity. 28

When isolated from other categories, the sub-generic nuances of action/adventure productions become somewhat more marked (Figure 2). This configuration also exhibits a strong authorial signal and, very characteristically for film dialogues, a tendency to cluster by thematic universe [Hołobut et al. 2016] #holobut2017H. As a result, the mob-ridden environment of Gotham and Spiderman's New York becomes very similar to that of the more direct depictions of the mafia by Coppola. 29

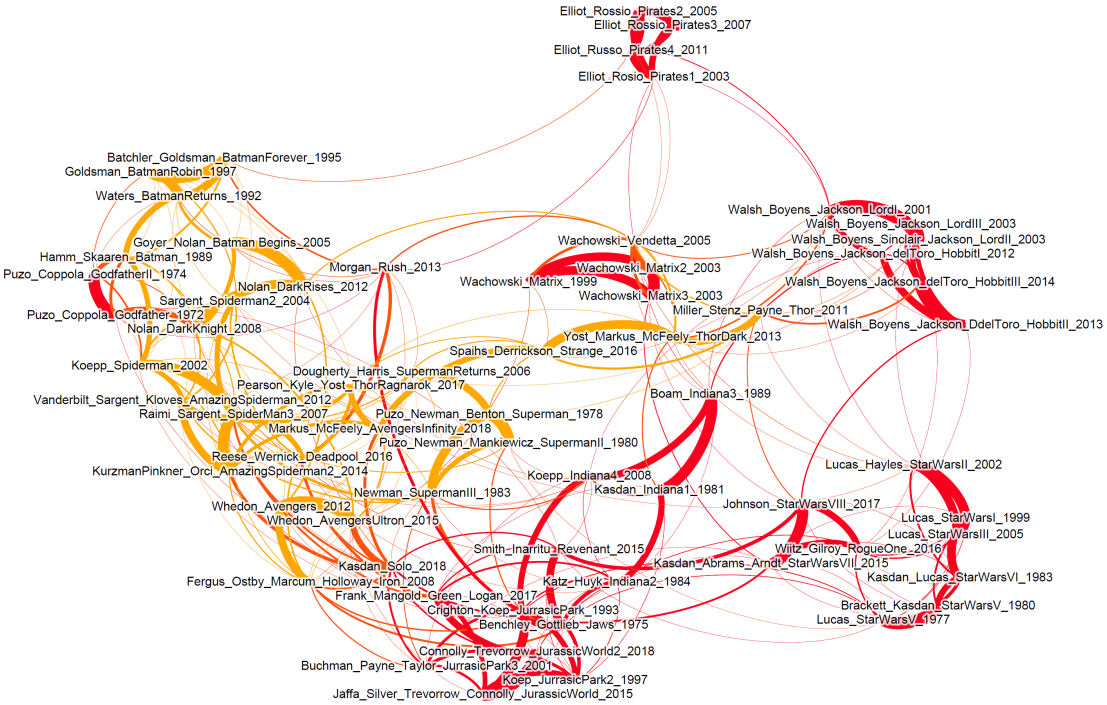

Figure 2. Network analysis of the action/adventure (red) and superhero (orange) genre.

Sentiment analysis can be mostly presented with much less statistical manipulation, simply by plotting the various sentiment/emotion scores against time of film release. Figure 3 shows the proportion of positive to negative sentiments in our entire set of film dialogues (data points) and the linear trend (lines) of this ratio. In the figure, balanced positive-to-negative contents would place a given film at 0.5 on the vertical scale; anything above that value contains more positive than negative terms; anything below 0.5 descends into negativity. As can be seen, values for all groupings are highly scattered; nevertheless, some trends in the appearance of positive and negative sentiments may be observed. The oldest genres in this set, romances and thrillers, exhibit a fairly stable proportion, with, unsurprisingly, a much higher participation of the positive despite a slight downwards trend. By contrast, the initially much more negative thrillers rise slightly from darker to lighter moods. A much more marked fall in positive sentiments can be observed in the dialogues of action/adventure films and even more so in the superhero subgenre; in the second decade of the 21st century, they are more negative in sentiment than even the thrillers. Still, negativity flourishes the most in an entirely different class: in the vampire movies; their verbally expressed positive sentiments dwindle almost to 0.5 at the end of the timeline, perhaps not unrelatedly to a very marked and mirror-image ascent of positive values in chick flicks.

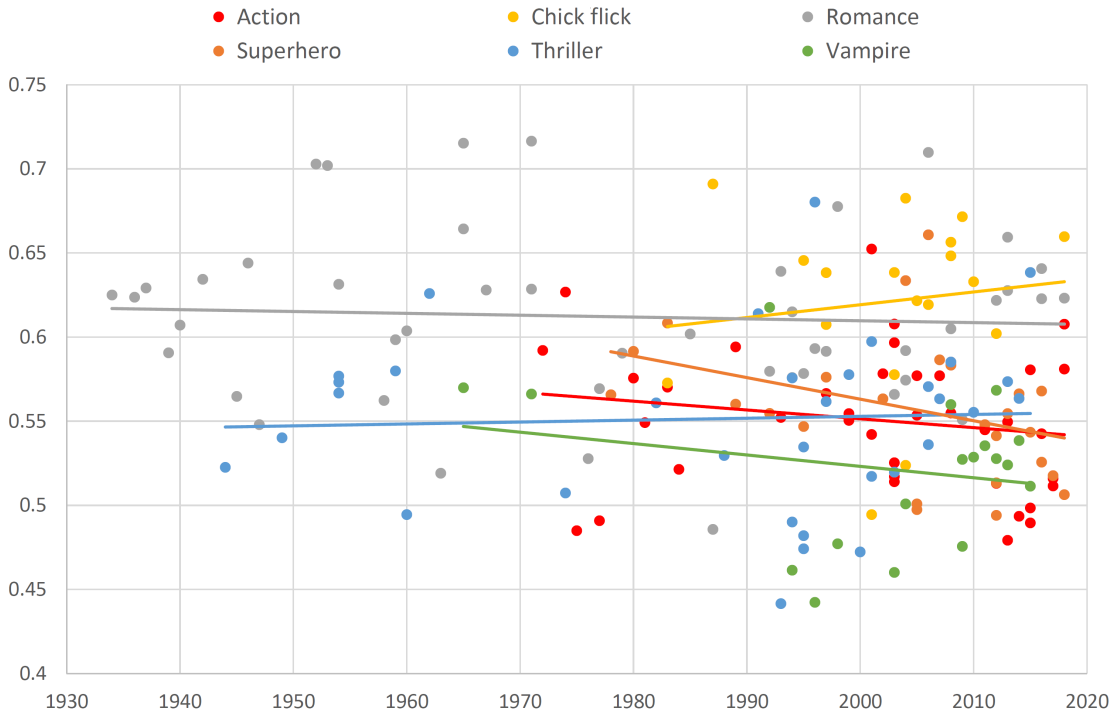

Figure 3. Positive-to-negative sentiment ratio in the entire set of film dialogues, color-coded by (principal) genre, with linear trendlines.

Numeric results for the eight basic Plutchik emotions are consistent with both those for positivity/negativity and with each other. Based on the results of the Kruskal-Wallis tests obtained for each pair of emotions (well below any values that could suggest statistical insignificance), their representation in the boxplots in Figure 4 shows a very strong correlation for high levels of anticipation and joy between chick flicks and romances. These two categories are also very much alike in their equally low values of anger, disgust, and fear. At the same time, in terms of these five emotions, both romances and chick flicks differ with very high statistical significance from the other genres. Sadness and fear are also statistically higher in films about vampires; this grouping is also significantly deficient in trust. It may be said more generally that, in terms of sentiment and emotion analysis, three rather than six categories are discernible here: 1) chick flicks and romances, 2) vampire films, 3) action, superhero and thriller movies.

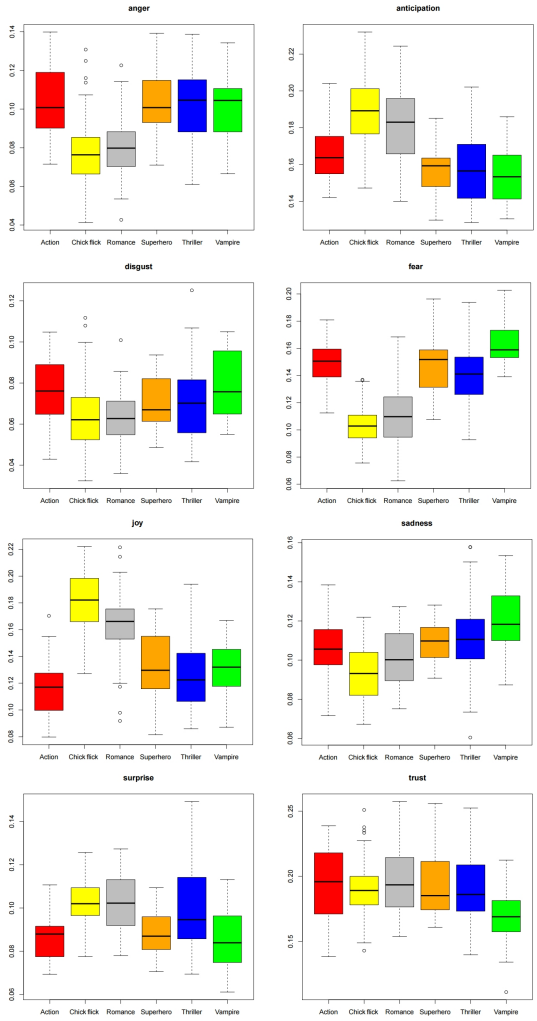

Figure 4. Relative frequencies of words associated with the eight basic emotions in the entire set of film dialogues. Boxes denote the range of the 25th-75th percentiles; the short horizontal line inside each box is the median; whiskers and circles denote outlying (extreme) values.

Discussion

Several interesting observations can be made on the basis of this study. Not the least of these is that the tentative theme/genre classifications that we have started with have been verified by our quantitative approach in both stylometry and sentiment analysis. Our most-frequent-word analysis shows that while some of our initial groupings do seem to create discernible communities, individual films do not seem to care. These stylometric “misclassifications” (such as the affinity of Woody Allen’s sophisticated comedies to other comedic productions rather than for instance period romance dramas) are understandable and attest to the inadequacy of our *a priori* film classification rather than that of the stylometric approach; indeed, these misclassifications often correct the rough pigeon-holing that we used, revealing the influence of authorial and sub-generic nuance on the measurable features of film dialogue style and mutual relationships between dialogue techniques in different groups of films considered.

32

Film dialogue seems to successfully reflect the fact that individual films rarely conform to the requirements of a single, unadulterated, major-level genre such as action films or thrillers. This becomes quite evident when one views in detail the network analysis graph in full complexity of its signals. At times, films cluster by screenwriter as if in accordance with the *Schreiber* theory of film authorship; at times, the director’s choices seem to bring together different screenwriters’ efforts, as if to vindicate the claims of *auteur* theorists who believe in the director’s influence over the collective work of different specialists, screenwriters included. Then there are mainstream films that do cluster by genre. Elsewhere, the collaborative nature of making movies takes over to create a strong franchise signal, which in turn may be overridden by that of the studio: after all, in our graphs, Disney *Star Wars* have left the orbit of their predecessors. Diachrony is

33

another significant element which deserves more in-depth qualitative analysis, possibly looking at the evolution of non-verbal cinematic techniques; the language of film dialogue in the 1940s and the 1950s stands apart from later productions. What is more, the diachronic element is not limited to the date of the films' release; our stylometric results can also reflect their temporal setting, quite plausibly due to archaic stylization of the spoken parts. This has already been noted in an earlier study on historical films and series #holobut2017.

34 It is quite remarkable that, despite all the above caveats against lexicon-based sentiment analysis, much of our results of this type of distant reading agrees with what we know about the genres from "close viewing." The fact that computational sentiment analysis agrees so well, indeed so boringly, with human presuppositions might be a result of the somewhat crude approach – through decontextualized lexical items. At the same time, however, the crude nature of the lexicon approach may be reflected in the fact that, contrarily to most-frequent-word stylometry, sentiment analysis shows fewer, rather than more, valid divisions: chick flicks merge with romances, action/superhero films with thrillers. Meanwhile, vampire movies, an internally conflicted category according to Delta, acquires a separate status due to higher levels of "fear" and lower levels of "trust." This quantitative procedure can act as a starting point for more in-depth qualitative evaluation of emotions (both those expressed by characters and those evoked in viewers) and the dominance of particular speech acts in the verbal exchanges on screen (such as teasing in comedy, threatening in gangster films, commanding in drama or explaining in science-fiction, cf. Kozloff 2000).

35 It may seem paradoxical that of the two approaches used in this study, the one that seems to have less to do with semantics – the comparison of frequencies of very frequent words, grammatical words, function words, "synsemantic" words – also seems to reflect, albeit indirectly, the semantics of the film dialogues – their differing functionalities, their stylistic patternings, their themes, their thematic subgenres, their implied audience. At the same time, an approach more directly aimed at meaning: the search for terms of sentiment, *syuzhet*, emotion, suffers from the computers' still-underdeveloped ability to "understand" the meaning of literary texts and can only confirm – at best – what we already know about film genres. In fact, the little information we obtain from this latter part of the study makes our flawed (oversimplified) classification seem almost adequate – which it certainly is *not*. Still, it is enough to teach us never to trust a vampire...

36 Dialogue is only a small part of film. At first glance – and for quite a long time in the history of film studies – it seemed much less important to the overall form and meaning of the art than image or non-verbal sound. Film scholars, cognitive scientists and digital experts have already undertaken extensive research into film narrative structure (e.g. Murtagh 2009; McKie 2014), film digital archiving (e.g. Bateman 2016) and other "measurable" parameters of cinematic art, such as shot lengths, types of shots, motion types and/or visual characteristics of the image onscreen such as light intensity (e.g. Salt 2007; Cutting 2011; Suchan 2016; Heftberger 2018). Our study shows that results of these more standard digital film studies could be interestingly confronted with quantitative analysis of films' textual layer for better insight into the specificities and the mutual influences of film genres, authorial *oeuvre* or chronology; if only to see to what extent the sentiments and the emotions and the style in film dialogue go hand in hand – or not – with the visual and the non-verbal auditory components in various film genres and periods of film history.

37 Such a multifaceted analysis of film would be a dream interdisciplinary project, and, in Shakespearean phrase, devoutly to be wished. The fact that it has not been yet done, at least to our knowledge, is a result of obvious difficulties. It is possible to produce a large corpus of film dialogue texts, for instance editing the screenplays available online (cf. Murtagh 2009: 302); distant reading was made for big data (in this case, textual data too extensive to be processed by the naked eye); once the texts are acquired, the machines do the rest. However, the same can hardly be said of the more "technical" sides of filmmaking listed in the preceding paragraph: there, each item would require much annotation and coding, resulting in a much higher complexity of the project. The good news is that now that (we believe) interesting textual features of film dialogue have been found, the incentive to combine it with other features of film may result in a much larger project than the one that could be described within the limited space of this paper.

Appendix 1.

List of films included in the corpus divided into subcategories (following IMDBs classifications) and ordered chronologically:

- **Action**
    - The Godfather, dir. F. F. Coppola, USA (1971) [crime, drama]
    - The Godfather, dir. FF. Copola, USA (1974) F. F. Coppola [crime, drama]
    - Jaws, dir. S. Spielberg, USA (1975) [adventure, drama, thriller]
    - Star Wars, dir. G. Lucas, USA (1977) [action, adventure, fantasy]
    - Star Wars Episode V: The Empire Strikes Back, dir. I. Kershner, USA (1980) [action, adventure, fantasy]
    - Raiders of the Lost Ark, dir. S. Spielberg, USA (1981) [action, adventure]
    - Star Wars Episode VI: Return of the Jedi, dir. R. Marquand, USA (1984) [action, adventure, fantasy]
    - Indiana Jones and the Temple of Doom, dir. S. Spielberg, USA (1984) [action, adventure]
    - Indiana Jones and the Last Crusade, dir. S. Spielberg, USA (1989) [action, adventure, fantasy]
    - Jurassic Park, dir. S. Spielberg, USA (1993) [adventure, sci-fi, thriller]
    - The Lost World: Jurassic Park, dir. S. Spielberg (1997) [action, adventure, sci-fi]
    - The Matrix, dir. L. and L. Wachowski, USA (1999) [action, sci-fi]
    - Lord of the Rings: The Fellowship of the Ring, dir. P. Jackson, News, Zealand, USA (2001) [adventure, drama, fantasy]
    - Star Wars Episode I: The Phantom Menace, dir. G. Lucas, USA (2001) [action, adventure, fantasy]
    - Jurassic Park III, dir. J. Johnston, USA (2001) [action, adventure, sci-fi]
    - Lord of the Rings: The Two Towers, dir. P. Jackson, New Zealand, USA (2002) [adventure, drama, fantasy]
    - Star Wars Episode II: Attack of the Clones, dir. G. Lucas, USA (2002) [action, adventure, fantasy]
    - The Lord of the Rings: The Return of the King, dir. P. Jackson, New Zealand, USA (2003) [action, adventure, drama]
    - The Matrix Reloaded, dir. L. and L. Wachowski, USA (2003) [action, sci-fi]
    - The Matrix Revolutions, dir. L. and L. Wachowski, USA (2003) [action, sci-fi]
    - The Pirates of the Caribbean: he Curse of the Black Pearl, dir. G. Verbinski, USA (2003) [action, adventure, fantasy]
    - Star Wars Episode III: Revenge of the Sith, dir. G. Lucas, USA (2005) [action, adventure, fantasy]
    - The Pirates of the Caribbean: The Dead Man’s Chest, dir. G. Verbinsku, USA (2006) [action, adventure, fantasy]
    - Pirates of the Caribbean: At World’s End, dir. G. Verbinski, USA (2007) [action, adventure, fantasy]
    - Indiana Jones and the Kingdom of the Crystal Skull, dir. S. Spielberg, USA (2008) [action, adventure, fantasy]
    - Pirates of the Caribbean: On Stranger Tides, dir. R. Marshall, USA (2012) [action, narrative, fantasy]
    - The Hobbit: An Unexpected Journey, dir. P. Jackson, USA, New Zealand (2012) [adventure, family, fantasy]
    - The Hobbit: The Desolation of Smaug, dir. P. Jackson, USA, New Zealand (2013) [adventure, fantasy]
    - Rush, dir. R. Howard, UK, Germany, USA (2013) [action, biography, drama]
    - Hobbit: The Battle of the Five Armies, dir. P. Jackson, USA, New Zealand (2014) [adventure, fantasy]
    - Jurassic World, dir. C. Trevorrow, USA (2015) [action, adventure, sci-fi]
    - Star Wars Episode VII: The Force Awakens, dir. J. J. Abrams, USA (2015) [action, adventure, fantasy]

- The Revenant, dir. A. G. Iñárritu, USA, Hong Kong, Taiwan (2015) [action, adventure, biography]
- Rogue One, dir. G. Edwards, USA (2016) [action, adventure, sci-fi]
- Star Wars: The Last Jedi, dir. R. Johnson, USA (2017) [action, adventure, fantasy]
- Jurassic World: The Fallen Kingdom, dir. J. A. Bayona, USA (2018) [action, adventure, sci-fi]
- Solo: A Star Wars Story, dir. R. Howard, USA (2018) [action, adventure, fantasy]

- **Chick flick**

  - Ice Castles, dir. D. Wrye, USA (1978) [drama, romance, sport]
  - Terms of Endearment, dir. J.L. Brooks, USA (1983) [comedy, drama]
  - Splash (1984), dir. R. Howard, USA [comedy, fantasy, romance]
  - Pretty in Pink, dir. H. Deutch, USA (1986) [comedy, drama, romance]
  - Dirty Dancing, dir. E. Ardolino, USA (1987) [drama, music, romance]
  - Moonstruck, dir N. Jewison, USA (1987) [comedy drama, romance]
  - Beaches, dir. G. Marshall, USA (1988) [comedy, drama, music]
  - Steal Magnolias, dir. H. Ross, USA (1989) [comedy, drama, romance]
  - When Harry Met Sally, dir. R. Reiner, USA (1989) [comedy, drama, romance]
  - Ghost, dir. Jerry Zucker, USA (1990) [drama, fantasy, romance]
  - Pretty Woman, dir. G. Marshall, USA (1990) [comedy, romance]
  - Father of the Bride, dir. Ch. Shyer, USA (1991) [comedy, family, romance]
  - Fried Green Tomatoes, dir. J. Avnet, USA (1991) [drama]
  - Thelma and Louise, dir. R. Scott, USA (1991) [adventure, crime, drama]
  - Indecent Proposal, dir. A. Lyne, USA (1993) [drama, romance]
  - Sleepless in Seattle, dir. N. Ephron, USA (1993) [comedy, drama, romance]
  - Untamed Heart, dir. T. Bill (1993), USA [comedy, drama, romance]
  - Four Weddings and a Funeral, dir. M. Newell, UK [comedy, drama, romance]
  - Boys on the Side, dir. H. Ross, USA, France (1995) [comedy, drama]
  - Clueless, dir. A. Heckerling, USA (1995) [comedy, romance]
  - A Month by the Lake, dir. J. Irvin, UK, USA (1995) [comedy, drama, romance]
  - While You Were Sleeping, dir. J. Turteltaub, USA (1995) [comedy, drama, romance]
  - The Truth About Cats and Dogs, dir. M. Lehmann, USA (1996) [comedy, romance]
  - Romy and Michele's High School Reunion, dir. D. Mirkin, USA (1997) [comedy]
  - My Best Friend's Wedding, dir. P.J. Hogan, USA (1997) [comedy, drama, romance]
  - Titanic, dir. J. Cameron, USA (1997) [drama, romance]
  - Home Fires, dir. D. Parisot, USA (1998) [comedy, romance, drama]
  - Hope Floats, dir. F. Whitaker, USA (1998) [drama, romance]
  - Ever After, dir. A. Tennant, USA (1998) [comedy, drama, romance]
  - Stepmom, dir. Ch. Columbus, USA (1998) [comedy, drama]
  - How Stella Got Her Groove Back, dir. K. R. Sullivan, USA (1998) [comedy, drama, romance]
  - Practical Magic, dir. G. Dunne, USA (1998) [comedy, drama, fantasy]
  - The Wedding Singer, dir. F. Coraci, USA (1998) [comedy, music, romance]
  - You've Got Mail, dir. N. Ephron, USA (1998) [comedy, drama, romance]
  - Notting Hill, dir. R. Curtis, UK, USA (1999) [comedy, drama, romance]
  - Forces of Nature, dir. B. Hughes, USA (1999) [comedy, romance]
  - 10 Things I Hate About You, dir. G. Junger, USA (1999) [comedy, drama, romance]
  - Never Been Kissed, dir. R. Gosnell, USA (1999) [comedy, drama, romance]
  - She's All That, dir. R. Iscove, USA (1999) [comedy, romance]
  - Runaway Bride, dir. G. Marshall, USA (1999) [comedy, romance]
  - Miss Congeniality, dir. D. Petrie, USA (2000) [action, comedy, crime]
  - Save the Last Dance, dir. T. Carter, USA (2000) [drama, music, romance]

- The Wedding Planner, dir. A. Shankman, Germany, USA (2001) [comedy, romance]
- Serendipity, dir. P. Chelsom, USA (2001) [comedy, romance]
- Bridget Jones' s Diary, dir. S. Maguire, UK, France, USA (2001) [comedy, drama, romance]
- The Princess Diaries, dir. G. Marshall, USA (2001) [comedy, family, romance]
- Legally Blonde, dir. R. Luketic, USA (2001) [comedy, romance]
- Maid in Manhattan, dir. W. Wang, USA (2002) [comedy, drama, romance]
- Sweet Home Alabama, dir. A. Tennant, USA (2002) [comedy, romance]
- My Big Fat Greek Wedding, dir. J. Zwick, Canada, USA (2002) [comedy, drama, romance]
- Crossroads, dir. T. Davies, USA (2002) [comedy, drama, romance]
- Divine Secrets of the Ya-Ya Sisterhood, dir. C. Khouri, USA (2002) [drama]
- A Walk t Remember, dir A. Shankman, USA (2002) [drama, romance]
- How to Lose a Guy in 10 Days, dir. D. Petrie, USA (2003) [comedy, romance]
- Uptown Girls, dir. B. Yakin, USA (2003) [comedy, drama, romance]
- Le Divorce, dir. J. Ivory, USA (2003) [drama, romance, comedy]
- Love, Actually, dir. R. Curtis, UK, USA, France (2003) [comedy, drama, romance]
- Alex and Emma, dir. R. Reiner, USA (2003) [comedy, romance]
- Under the Tuscan Sun, dir. A. Wells, USA, Italy (2003) [comedy, drama, romance]
- Legally Blonde 2, dir. Ch. Herman-Wormfeld, USA (2003) [comedy]
- Love Don't Cost a Thing, dir. T. Byer, USA (2003) [comedy, romance, drama]
- Bridget Jones: The Edge of Reason, dir. B. Kidron, UK, USA (2004) [comedy, drama, romance]
- The Notebook, dir. N. Cassavetes, USA (2004) [drama, romance]
- 13 Going on 30, dir. G. Winick, USA (2004) [comedy, fantasy, romance]
- 50 First Dates, dir. P. Segal, USA (2004) [comedy, drama, romance]
- Raisin Helen, dir. G. Marshall, USA (2004) [comedy, drama, romance]
- The Little Black Book, dir. N. Hurran, USA (2004) [comedy, romance, drama]
- Mean Girls, dir. M. Waters, USA (2004) [comedy]
- Fever Pitch, dir. B and P. Farelly, USA (2005) [comedy, drama, romance]
- Monster-in-Law, dir. R. Luketic, USA, Germany (2005) [comedy, romance]
- The Wedding Date, dir. C. Kilner, SUA (2005) [comedy, romance]
- The Perfect Man, dir. M. Rosman, USA (2005) [comedy, family, romance]
- Hitch, dir. A. Tennant, USA (2005) [comedy, romance]
- Prime, dir. B. Younger, USA (2005) [comedy, drama, romance]
- The Sisterhood of the Travelling Pants, dir. K. Kwapis (2005) [comedy, drama, romance]
- Must Love Dogs, dir. G. D. Goldberg, USA (2005) [comedy, romance]
- Last Holiday, dir. W. Wang, USA (2006) [comedy]
- Failure to Launch, dir. T. Dey, USA (2006) [comedy, romance]
- The Break-Up, dir. P. Reed, USA (2006) [comedy, drama, romance]
- The Devil Wears Prada, dir. D. Frankel, USA, France (2006) [comedy, drama]
- P.S. I Love You, dir. R. LaGravenese, USA (2007) [drama, romance]
- Enchanted, dir. K. Lima, USA (2007) [animation, comedy, family]
- Music and Lyrics, dir. M. Lawrence, USA (2007) [comedy, music, romance]
- Mamma Mia!, dir. Ph. Lloyd, USA, UK, Germany (2007) [comedy, musical, romance]
- Definitely, Maybe, dir. A. Brooks, USA (2008) [comedy, drama, romance]
- Made of Honor, dir. P. Weiland, USA, UK (2008) [comedy, romance]
- 27 Dresses, dir. A. Fletcher, USA (2008) [comedy, romance]
- Sex and the City, dir. M. P. King, USA (2008) [comedy, drama, romance]
- The Sisterhood of the Traveling ants 2, dir. S. Hamri, USA (2008) [comedy, drama, romance]
- The Time Traveler's Wife, dir. R. Schwentke, USA (2008) [drama, fantasy, romance]
- Bride Wars, dir. G. Winick, USA (2009) [comedy, romance]
- He's Just Not That Into You, dir. K. Kwapis, Germany, USA (2009) [comedy, drama, romance]

- The Proposal, dir. A. Fletcher, USA (2009) [comedy, drama, romance]
- Couples Retreat, dir. P. Billingsley, USA (2009) [comedy]
- Julie and Julia, dir. N. Ephron, USA (2009) [biography, drama, romance]
- The Ugly Truth, dir. R. Luketic, USA (2009) [comedy, romance]
- Labor Pains, dir. L. Shapiro, USA (2009) [comedy, romance]
- Valentine's Day, dir. G. Marshall, USA (2009) [comedy romance]
- The Switch, dir. J. Gordon, W. Speck, USA (2010) [comedy, drama, romance]
- Letters to Juliet, dir. G. Winnick, USA (2010) [adventure, comedy, drama]
- Dear John, dir. L. Hallström, USA (2010) [drama, romance, war]
- The Back-Up Plan, dir. A. Poul, USA (2010) [comedy, romance]
- Sex in the City, dir. M. P. King, USA (2010) [comedy, drama, romance]
- Easy A, dir. W. Gluck, USA (2010) [comedy, drama, romance]
- Something Borrowed, dir. L. Greenfield, USA [comedy, drama, romance]
- Love, Wedding, Marriage, dir. D. Mulroney, USA (2011) [comedy]
- Bridesmaids, dir. P. Feig, USA (2011) [comedy, romance]
- Crazy, Stupid Love, dir. G. Ficarra, J. Requa, USA (2011) [comedy, drama, romance]
- What to Expect When You're Expecting, dir. K. Jones, USA (2012) [comedy, drama, romance]
- Pitch Perfect, dir. J. Moore, USA (2012) [comedy, music, romance]
- Safe Have, dir. L. Hallström, USA (2012) [drama, romance, thriller]
- About Time, dir. R. Curtis, UK (2013) [comedy, drama, fantasy]
- The Other Woman, dir. N. Cassavetes, USA (2014) [comedy, romance]
- The Single Moms Club, dir. T. Perry, USA (2014) [comedy, drama]
- Lila and Eve, dir. Ch. Stone III, USA (2015) [crime, drama, mystery]
- How to Be Single, dir. Ch. Ditter, USA (2016) [comedy, drama, romance]
- Home Again, dir. H. Meyers-Shyer, USa (2017) [comedy, drama, romance]
- Crazy Rich Indians, dir. J. M. Chu, USA (2018) [comedy, romance]
- Book Club, dir. B. Holderman, USA (2018) [comedy, drama, romance]

- **Romance**
  - It happened One Night, dir. F. Capra, USA (1934) [comedy, romance]
  - The Last of the Mohicans, dir. G. B. Seitz, USA (1936) [adventure, drama, history]
  - A Star is Born, dir. W. W. Wellman, J. Conway, USA (1937) [drama]
  - Gone with the Wind, dir. V. Flemming, G. Cukor, USA (1939) [drama, history, romance]
  - Rebecca, dir. A. Hitchcock, USA (1940) [drama, mystery, romance]
  - Casablanca, dir. M. Curtiz, USA (1942) [drama, romance, war]
  - Brief Encounter, dir. D. Lean, USA (1945) [drama, romance]
  - The Best Years of Our Lives, dir. W. Wyler, USA (1946) [drama, romance, war]
  - Out of the Past, dir. J. Tourneur, USA (1947) [crime, drama, film-noir]
  - Singin' in the Rain, dir. S. Donen, G. Kelly, USA (1952) [comedy, musical, romance]
  - Roman Holiday, dir. W. Wyler, USA (1953) [comedy, romance]
  - A Star is Born, dir. G. Cukor, USA (1954) [drama, musical, romance]
  - Vertigo, dir. A. Hitchcock, USA (1958) [mystery, romance, thriller]
  - Some Like It Hot, dir. B. Wilder, USA (1959) [comedy, romance]
  - The Apartment, dir. B. Wilder, USA (1960) [comedy, drama, romance]
  - Charade, dir. S. Donen, USA (1963) [comedy, mystery, romance]
  - Doctor Zhivago, dir. D. Lean, USA, Italy, UK (1965) [drama, romance, war]
  - Sounds of Music, dir. R. Wise, USA (1965) [biography, drama, family]
  - The Gradate, dir. M. Nichols, USA (1967) [comedy, drama, romance]
  - Fiddler on the Roof, dir. N. Jewison, USA (1971) [drama, family, musical]
  - Harold and Maude, dir. H. Ashby, USA (1971) [comedy, drama, romance]

- A Star is Born, dir. F. Pierson (1976) [drama, music, romance]
- Annie Hall, dir. W. Allen, USA (1977) [comedy, romance]
- Manhattan, dir. W. Allen, USA (1979) [comedy, drama, romance]
- Out of Africa, dir. S. Pollack, USA, UK (1985) [biography, drama, romance]
- The Princess Bride, dir. R. Reiner, USA (1987) [adventure, family, fantasy]
- The Last of the Mohicans, dir. M. Mann, USA (1992) [action, adventure, drama]
- Groundhog Day, dir. H. Ramis, USA (1993) [comedy, fantasy, romance]
- Forrest Gump, dir. R. Zemeckis, USA (1994) [drama, romance]
- Legends of the Fall, dir. E. Zwick, USA (1994) [drama, romance, war]
- Before Sunrise, dir. R. Linklater, USA (1995) [drama, romance]
- Good Will Hunting, dir. G. Van Sant, USA (1997) [drama, romance]
- The English Patient, dir. A. Minghella, USA, UK (1997) [drama, romance, war]
- The Horse Whisperer, dir. R. Redford, USA (1998) [drama, romance, western]
- Big Fish, dir. T. Burton, USA (2003) [adventure, drama, fantasy]
- Before Sunset, dir. R. Linklater (2004), USA [drama, romance]
- Eternal Sunshine of the Spotless Mind, dir. M. Gondry, USA (2004) [drama, romance, sci-fi]
- Once, dir. J. Carney, Ireland (2007) [drama, music, romance]
- Slumdog Millionaire, dir. D. Boyle, L. Tandan, UK, USA, France, Germany, India (2008) [drama, romance]
- Mr Nobody, dir. J. Van Dormael, USA (2009) [drama, fantasy, romance]
- The Perks of Being a Wallflower, dir. S. Chbosky, USA (2012) [drama, romance]
- Before Midnight, dir. R. Linklater, USA (2013) [drama, romance]
- Her, dir. S. Jonze, USA (2013) [drama, romance, sci-fi]
- Sing Street, dir. J. Carney, Ireland, UK, USA (2016) [comedy, drama, music]
- La La Land, dir. D. Chazelle, USA, Hing Kong (2016) [comedy, drama, music]
- A Star is Born, dir. B. Cooper, USA (2018) [drama, music, romance]

- **Superhero**

  - Superman, dir. R. Donner, USA, UK, Switzerland, Canada, Panama (1978) [action, adventure, drama]
  - Superman II, dir. R. Lester, R. Donner, USA, UK, Canada (1980) [action, adventure, sci-fi]
  - Superman III, dir. R. Lester, UK, USA (1983) [action, comedy, sci-fi]
  - Batman, dir. T. Burton, USA (1989) [action, adventure]
  - Batman Returns, dir. T. Burton, USA (1992) [action, crime, fantasy]
  - Batman Forever, dir. J. Schumacher, USA (1995) [action, adventure, fantasy]
  - Batman and Robin, dir. J. Schumacher, USA, UK (1997) [action, sci-fi]
  - Spider-Man, dir. S. Raimi, USA (2002) [action, adventure, sci-fi]
  - Spider-Man 2, dir. S. Raimi, USA (2004) [action, adventure, sci-fi]
  - Batman Begins, dir. Ch. Nolan, USA (2005) [action, adventure, thriller]
  - V for Vendetta, dir. J. McTeigue, USA (2005) [action, drama, sci-fi]
  - Superman Returns, dir. B. Singer, USA (2006) [action, adventure]
  - Spider-Man 3, dir. S. Raimi, USA (2007) [action, adventure, sci-fi]
  - The Dark Knight, dir. Ch. Nolan, USA (2008) action, crime, drama]
  - Iron Man, dir. J. Favreau, USA (2008) [action, adventure, sci-fi]
  - Thor, dir. K. Branagh, USA (2011) [action, adventure, fantasy]
  - The Amazing Spider-Man, dir. M. Webb, USA (2012) [action, adventure, fantasy]
  - The Avengers, dir. J. Whedon, USA (2012) [action, adventure, sci-fi]
  - The Dark Knight Rises, dir. Ch. Nolan, USA (2012) [action, thriller]
  - Thor: The Dark World, dir. A. Taylor, USA (2013) [action, adventure, fantasy]
  - The Amazing Spider-Man 2, dir. M. Webb, USA (2014) [action, adventure, sci-fi]

- The Avengers: Age of Ultron, dir. J. Whedon, USA (2015) [action, adventure, sci-fi]
- Deadpool, dir. T. Miller, USA (2016) [action, adventure, comedy]
- Doctor Strange, dir. S. Derrickson, USA (2016) [action, adventure, fantasy]
- Logan: Wolverine, dir. J. Manglod, USA (2017) [action, drama, sci-fi]
- Thor: Ragnarok, dir. T. Waititi, USA (2017) [action, adventure, comedy]
- The Avengers: Infinity War, dir. A. Russo, J. Russo, USA (2018) [action, adventure, fantasy]

- **Thriller**

  - The Third Man, dir. C. Reed, UK (1949) [film-noir, mystery, thriller]
  - Dial M for Murder, dir. A. Hitchcock, USA (1954) [crime, thriller]
  - Double Indemnity, dir. B. Wilder, USA (1954) [crime, drama, film-noir]
  - On the Waterfront, dir. E. Kazan, USA (1954) [crime, drama, thriller]
  - Rear Window, dir. A. Hitchcock, USA (1954) [mystery, thriller]
  - North by Nothwest, dir. A. Hitchcock, USA (1959) [adventure, mystery, thriller]
  - Psycho, dir. A. Hitchcock, USA (1960) [horror, mystery, thriller]
  - What Ever Happened to Baby Jane?, dir. R. Aldrich, USA (1962) [drama, horror, thriller]
  - Chinatown, dir. R. Polanski, USA (1974) [drama, mystery, thriller]
  - Blade Runner, dir. R. Scott, USA, Hong Kong (1982) [sci-fi, thriller]
  - Die Hard, dir. J. McTiernan, USA (1988) [action, thriller]
  - Silence of the Lambs, dir. J. Demme, USA (1991) [crime, drama, thriller]
  - Reservoir Dogs, dir. Q. Tarantino, USA (1992) [crime, drama, thriller]
  - Léon, dir. L. Besson, France, USA (1994) [crime, drama, thriller]
  - The Heat, dir. M. Mann, USA (1995) [crime, drama, thriller]
  - Pulp Fiction, dir. Q. Tarantino, USA (1994) [crime, drama]
  - The Usual Suspects, dir. B. Singer USA, Germany (1995) [crime, mystery, thriller]
  - Fargo, dir. J. and E. Coen, USA (1996) [crime, drama, thriller]
  - L.A. Confidential, dir. C. Hanson, USA (1997) [crime, drama, mystery]
  - Seven, dir. D. Fincher, USA (1997) [crime, drama, mystery]
  - The Sixth Sense, dir. M. Night Shyamalan, USA (1999) [drama, mystery, thriller]
  - Memento, dir. Ch. Nolan, USA (2000) [mystery, thriller]
  - Donnie Darko, dir. R. Kelly, USA (2001) [drama, sci-fi, thriller]
  - Mulholland Drive, dir. D. Lynch, USA (2001) [drama, mystery, thriller]
  - Kill Bill dir. Q. Tarantino, USA (2003) [action, crime, thriller]
  - The Departed, dir. M. Scorsese, USA (2006) [crime, drama, thriller]
  - The Prestige, dir. Ch. Nolan, USA, Hong Kong (2006) [drama, mystery, sci-fi]
  - No Country for Old Men, dir. A. and J. Coen, USA (2007) [crime, drama, thriller]
  - Shutter Island, dir. M. Scorsese, USA (2010) [mystery, thriller]
  - Prisoners, dir. D. Villeneuve, USA (2013) [crime, drama, mystery]
  - Gone Girl, dir. D. Fincher, USA (2014) [crime, drama, mystery]
  - Room, dir. L. Abrahamson, USA (2015) [drama, thriller]
  - Vampire
  - Dr Terror’s House of Horrors, dir. F. Francis, UK (1965) [horror]
  - Lust for a Vampire, dir. J. Sangster, UK (1971) [horror]
  - Dracula, dir. F. Ford Coppola, USA (1992) [horror]
  - Interview with the Vampire: The Vampire Chronicles, dir. N. Jordan, USA (1994) [drama, horror]
  - From Dusk till Dawn, dir. R. Rodriguez, USA (1996) [action, crime, horror]
  - Blade, dir. S. Norrington, USA (1998) [action, horror, sci-fi]
  - Underworld, dir. L. Wiseman, USA, UK, Germany, Hungary (2003) [action, fantasy, thriller]
  - Van Helsing, dir. S. Sommers, USA, Czech Republic (2004) [action, adventure, fantasy]
  - Twilight, dir. C. Hardwicke, USA (2008) [drama, fantasy, romance]

- Jennifer's Body, dir. K. Kusama, USA (2009) [comedy, horror]
- The Twilight Saga: New Moon, sir. Ch. Weitz, USA (2009) [adventure, drama, fantasy]
- The Twilight Saga: Eclipse, dir. D. Slade, USA (2010) [adventure, drama, fantasy]
- The Twilight Saga: Breaking Dawn Part 1, dir. B. Condon, USA (2011) [adventure, drama, fantasy]
- The Twilight Saga: Breaking Dawn Part 2, dir. B. Condon, USA (2012) [adventure, drama, fantasy]
- Dark Shadows, dir. T. Burton, USA, Australia (2012) [comedy, fantasy, horror]
- The Mortal Instruments: City of Bones, dir. H. Zwart, USA, Germany, Canada, UK (2013) [action, fantasy, horror]
- Vampire Academy, dir. M. Waters, USA, UK (2014) [action, comedy, drama]
- Goosebumps, R. Letterman, USA, Australia (2016) [adventure, comedy, family]

Notes

[1] For an overview of corpus research into audiovisual translation, see Baños-Piñero 2013. For an overview of research into scripted speech, see Bednarek 2017.

[2] These were edited intralingual subtitles, randomly checked with the screen version for accuracy. It goes without saying that intralingual subtitles may occasionally involve the simplification of screen exchanges to match an average reading speed. As all forms of transcription of scripted speech back into writing, it also requires the sanitation of overlapping fragments and intrusive repetitions. However, the dialogue lists we randomly compared with complete audiovisual material revealed only minimal traces of condensation, quite insignificant for the procedures we applied and the methods used are insensitive to the edition techniques usually used to condense dialogues in captions (such as grammatical tense revision).

Works Cited

Acerbi et al. 2013 Acerbi A, Lampos V., Garnett P., and Bentley A. R. "The Expression of Emotions in 20th Century Books", *PLoS ONE* 8(3) (2013): e59030, doi: 10.1371/journal.pone.0059030.

Altman 1999 Altman, R. *Film/Genre*. British Film Institute, London (1999).

Bastian et al. 2009 Bastian, M., Heymann, S. and Jacomy, M. "Gephi: An Open Source Software for Exploring and Manipulating Networks", *Proceedings of the International AAAI Conference on Weblogs and Social Media*, San Jose, Ca (2009).

Bateman et al. 2016 Bateman, J. A., Tseng, Chiao-I., Seizov, O., Jacobs, A., Lüdtke, A., Müller, M. G. and Herzog, O. "Towards Next-generation Visual Archives: Image, Film and Discourse", *Visual Studies* 31 (2016); pp. 131-154.

Baxter 2014a Baxter, M. "Cinemetric Data Analysis" (2014), http://www.cinemetrics.lv/dev/Cinemetrics_Book_Baxter.pdf.

Baxter 2014b Baxter, M. "Quantitative Film Studies [a bibliography]" (2014), http://www.cinemetrics.lv/dev/bibliography_with_essay_Baxter.pdf.

Baños-Piñero et al. 2013 Baños-Piñero, R., Bruti, S. and Zanotti, S (eds) *Perspectives: Studies in Translatology. Special Issue: Corpus Linguistics and Audiovisual Translation: In Search of an Integrated Approach*, 21(4) (2013).

Bednarek 2010 Bednarek, M. *The Language of Fictional Television. Drama and Identity*, Continuum, London – New York (2010).

Bednarek 2011 Bednarek, M. "The Stability of the Televisual Character: A Corpus Stylistic Case Study". In R. Piazza, M. Bednarek and F. Rossi F. (eds), *Telecinematic Discourse*. Approaches to the Language of Films and Television Series, John Benjamins, Amsterdam – Philadelphia (2011), pp.185–204.

Bednarek 2018 Bednarek, M. *Language and Television Series: A Linguistic Approach to Television Dialogue*. Cambridge University Press, Cambridge (2018).

Bednarek and Zago 2017 Bednarek and M., Zago, R. "Bibliography of Linguistic Research on Fictional (Narrative, Scripted) Television Series and Films/Movie, Version 1" (2017), http://unipv.academia.edu/RaffaeleZago

Bobrowski 2015 Bobrowski, J. "Płaszczyzny stylizacji językowej w dialogu filmowym", *Polonica*, 35 (2015): 179–189.

http://dx.doi.org./10.17651/POLON.35.14.

Bondebjerg 2015 Bondebjerg, I. “Film: Genres and Genre Theory”. In J. D. Wright (ed.), *International Encyclopedia of the Social and Behavioral Sciences*, 2nd edition, Vol 9, Oxford (2015), pp. 160–164.

Burrows 1987 Burrows, J. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Clarendon, Oxford (1987).

Burrows 1994 Burrows, J. “Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative”, *Research in Humanities Computing* 2 (1994): pp 1-33.

Burrows 2002 Burrows, J. “The Englishing of Juvenal: Computational Stylistics and Translated Texts”, *Style* 36 (4) (2002): 677-698.

Chandler 1997 Chandler, D. *An Introduction to Genre Theory* (1997), http://www.aber.ac.uk/media/Documents/intgenre/chandler_genre_theory.pdf.

Choiński et al. 2019 Choiński, M., Eder, M. and Rybicki, J. “Harper Lee and Other People: A Stylometric Diagnosis”, *Mississippi Quarterly*, 70(3): 355-374.

Craig and Kinney 2009 Craig, H. and Kinney, A. “Shakespeare, Computers, and the Mystery of Authorship”, Cambridge University Press, Cambridge (2009).

Cutting et al. 2011 Cutting, J. E., Brunick, K. L., DeLong, J. E. and Iricinschi, C. “Quicker, Faster, Darker: Changes in Hollywood Film over 75 Years”, *i-Perception* 2 (2011): pp 569-576.

Eder 2017 Eder, M. “Visualization in Stylometry: Cluster Analysis Using Networks”, *Digital Scholarship in the Humanities* 32 (1) (2017): 50-64

Eder et al. 2016 Eder, M., Rybicki and J., Kestemont, M. “Stylometry with R: A Package for Computational Text Analysis”, *R Journal* 8 (1) (2016): 107-121.

Evert et al. 2017 Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C. and Vitt, T. “Understanding and Explaining Delta Measures for Authorship Attribution”, *Digital Scholarship in the Humanities* 32 (sup. 2) (2017): 4-16.

Freddi 2013 Freddi M. “Constructing a Corpus of Translated Films: A Corpus View of Dubbing”, *Perspectives: Studies in Translatology*, 21 (4) (2013): 491–503. DOI: 10.1080/0907676X.2013.831925.

Freddi and Pavesi 2009 Freddi M. and Pavesi M. “The Pavia Corpus of Film Dialogue: Methodology and Research Rationale”. In M. Freddi, M. Pavesi (eds), *Analysing Audiovisual Dialogue: Linguistic and Translational Insights*, Clueb, Bologna (2009), pp. 95–100.

Freedman 2015 Freedman, J. *The Cambridge Companion to Alfred Hitchcock*, Cambridge University Press, Cambridge, 2015.

Grant 2003 Grant, B.K. (ed.). *Film Genre Reader*. University of Texas Press, Austin ([1984] 2003).

Hayles 2010 Hayles, K. N. “How We Read: Close, Hyper, Machine”, *ADE Bulletin*, 150, (2010): 62-79.

Heftberger 2018 Heftberger, A. *Digital Humanities and Film Studies*, Springer, Berlin (2018).

Heiss and Soffritti 2005 Heiss, Ch. and Soffritti, M. “Parallelkorpora ‘gesprochener Sprache’ aus Filmdialogen? Ein multimedialer Ansatz für das Sprachenpaar Deutsch-Italienisch”. In J. Schwitalla, E. Wegstein (eds) *Korpus Linguistik Deutsch – synchron, diachron, kontrastiv*, Niemeyer, Tübingen (2005): 207-217.

Hendrykowski 1999 Hendrykowski, M. *Język ruchomych obrazów*, Ars Nova, Poznań (1999).

Hermann et al. 2015 Herrmann, J. B., van Dalen-Oskam, K. and Schoech C. “Revisiting Style, a Key Concept in Literary Studies”, *Journal of Literary Theory* 9 (1) (2015): 25-52.

Hoover 2007 Hoover, D. “Corpus Stylistics, Stylometry, and the Styles of Henry James”, *Style* 41(2) (2007): 174-203.

Hough 1969 Hough, G. *Style and Stylistics*, Routledge & Kegan Press, London (1969).

Hołobut and Rybicki 2018 Hołobut, A. and Rybicki, J. “Pride and Prejudice and Programming: A Stylometric Analysis”. In L. Raw (ed), *Adapted from the Original: Essays on the Value and Values of Works Remade for a New Medium*, McFarland, Jefferson (2018): pp. 134-147.

Hołobut and Woźniak 2017 Hołobut, A. and Woźniak, M. *Historia na ekranie: Gatunek filmowy a przekład audiowizualny*, Wydawnictwo UJ, Kraków.

Hołobut et al. 2016 Hołobut, A., Rybicki, J. and Woźniak, M. "Stylometry on the Silver Screen: Authorial and Translatorial Signals in Film Dialogue", Digital Humanities 2016: Conference Abstracts, Jagiellonian University and Pedagogical University, Krakow (2016): pp. 561–565.

Hołobut et al. 2017 Hołobut, A., Rybicki, J. and Woźniak, M. "Old Questions, New Answers: Computational Stylistics in Audiovisual Translation Research". In M. Deckert (ed), *Audiovisual Translation: Research and Use*, Peter Lang, Frankfurt (2017): pp. 203-216.

Jaeckle 2013 L. Jaeckle (ed), *Film Dialogue*, Wallflower Press, London – New York (2013).

Jockers 2016 Jockers, M. L. "Syuzhet: Extracts Sentiment and Sentiment-Derived Plot Arcs from Text" (2016), https://cran.r-project.org/web/packages/syuzhet/index.htm.

Jockers 2017 Jockers, M. L. "Resurrecting a Low Pass Filter (well, kind of)", Matthew L. Jockers Blog (2017), http://www.matthewjockers.net/2017/01/12/resurrecting/

Juola 2015 Juola, P. "The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions", *Digital Scholarship the Humanities* 30 (supp. 1): 100-113.

Kozloff 2000 Kozloff, S. *Overhearing Film Dialogue*, University of California Press, Berkeley (2000).

McKenna et al. 1999 McKenna, W., Burrows, J. and Antonia, A. "Beckett's Trilogy: Computational Stylistics and the Nature of Translation", *Revue Informatique et Statistique dans les Sciences humaine*s 35 (1999): 151-171.

McKie 2014 McKie, S. *Screenwriting 2.0: What Are the Possibilities of Screenplay "Datafication"? How the Screenplay as Sata Can Impact Creating and Managing, Presenting and Sharing, Analyzing and Visualizing Textual Screenplay Content*, PhD thesis, Royal Holloway College, London (2014).

Mealand 1999 Mealand, D. "Style, Genre, and Authorship in Acts, the Septuagint, and Hellenistic Historians", *Literary and Linguistic Computing* 14 (4) (1999): 479-506.

Miławska-Ratajczak 2019 Miławska-Ratajczak, M. *Dialog w roli głównej. Polszczyzna we współczesnym kinie na przykładzie wybranych utworów*, Universitas, Kraków (2019).

Mohammad 2011 Mohammad, S. "From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales", *Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH), Portland, OR, June 2011.

Mohammad and Turney 2010 Mohammad, S., and Turney P. "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon", *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, June 2010.

Moretti 2007 Moretti, F. *Graphs, Maps, Trees: Abstract Models for Literary History*, Verso, London, New York (2007).

Mosteller and Wallace 1964 Mosteller, F. and Wallace, D. "Inference and Disputed Authorship: The Federalist", Addison-Wesley, Reading (1964).

Murtagh et al. 2009 Murtagh, F, Ganz, A. and McKie, S. "The Structure of Narrative: The Case of Film Scripts", *Pattern Recognition*, 42 (2009): 302 –312.

Piazza 2011 Piazza, R. *The Discourse of Italian Cinema and Beyond: Let Cinema Speak*, Continuum, London–New York (2011).

Piazza et al. 2011 Piazza, R., Bednarek, M. and Rossi, F. (eds). *Telecinematic Discourse: Approaches to the Language of Films and Television Series*, John Benjamins, Amsterdam–Philadelphia (2011).

Pitera 1979 Pitera, Z. *Miłe kina początki*. Wydawnictwa Artystyczne i Filmowe, Warszawa (1979).

Plutchik 1980 Plutchik, R. "A General Psychoevolutionary Theory of Emotion". In R. Plutchik and H. Kellerman (eds), *Emotion: Theory, Research, and Experience: Vol. 1. Theories of Emotion*, Academic, New York (1980), pp. 3-33.

Quaglio 2009 Quaglio, P. *Television Dialogue: The Sitcom Friends vs Natural Conversation*, John Benjamins, Amsterdam–Philadelphia (2009).

R Core Team 2014 R Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical

Computing, Wien (2014), http://www.R-project.org/.

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

Romero Fresco 2009 Romero Fresco, P. “Naturalness in the Spanish Dubbing Language: A Case of Not-so-close Friends”, *Meta* 54(1) (2009): 49–72. doi:10.7202/029793ar

Rudman 2016 Rudman, J. “Non-Traditional Authorship Attribution Studies of William Shakespeare’s Canon: Some Caveats”, *Biblioteca di Studi di Filologia Moderna: Collana*, Riviste e Laboratorio (2016).

Rybicki 2012 Rybicki, J. “The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation”. In M. Oakes, M. Ji, *Quantitative Methods in Corpus-Based Translation Studies*, John Benjamins, Amsterdam (2012), 231-248.

Rybicki 2017a Rybicki, J. “A Second Glance at a Stylometric Map of Polish Literature”, *Forum of Poetics* 10 (2017): pp. 6-21.

Rybicki 2017b Rybicki, J. “Reading Novels with Statistics: What Numbers of Words Tell Us about Authorship, Genre, or Chronology”. In J.A. Dobelman (ed), *Models and Reality: Festschrift For James Robert Thompson*, T&NO Company, Chicago (2017), pp. 207-224.

Rybicki 2018 Rybicki, J. “Sentiment Analysis Across Three Centuries of the English Novel: Towards Negative or Positive Emotions?” EADH Conference Abstracts, Galway (2018).

Rybicki and Eder 2011 Rybicki, J.and Eder, M. “Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?”, *Literary and Linguistic Computing* 26 (3) (2011): 315-332.

Rybicki and Heydel 2013 Rybicki, J. and Heydel, M. “The Stylistics and Stylometry of Collaborative Translation: Woolf's ‘Night and Day’ in Polish”, *Literary and Linguistic Computing* 28 (4) (2013): 708-717.

Salt 2007 Salt, B. *Moving into Pictures: More on Film History, Style, and Analysis*, Starword, London (2007).

Schmid 1994 Schmid, H. “Probabilistic Part-of-Speech Tagging Using Decision Trees”. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Schmidt 2012 Schmidt, B. “Making Downton More Traditional” (2012), http://www.prochronism. com/2012/04/making-downton-more-traditional.html.

Smith and Aldridge 2011 Smith, P. and Aldridge, W. “Improving Authorship Attribution: Optimizing Burrows’ Delta Method”, *Journal of Quantitative Linguistics*, 18(1) (2011): 63–88.

Stone et al. 1966 Stone, P. J., Dunphy, D.C., and Smith, M. S. *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press, Cambridge, MA (1966).

Suchan and Bhatt 2016 Suchan, J. and Bhatt, M. “The Geometry of a Scene: On Deep Semantics for Visual Perception Driven Cognitive Film Studies”, *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, NY (2016), pp. 1-9.

Swafford 2015 Swafford, A. “Problems with the Syuzhet Package”, Anglophile in Academia: Annie Swafford’s Blog (2015), https://annieswafford.wordpress.com/2015/03/02/syuzhet/

Tuzzi and Cortelazzo 2018 Tuzzi, A. and Cortelazzo, M. (eds), *Drawing Elena Ferrante’s Profile*, Padova University Press, Padova (2018).

Valentini 2008 Valentini, C. “Forlixt1: The Forli Corpus of Screen Translation: Exploring Macrostructures”. In D. Chiaro, D. Heiss, Ch. Bucaria (eds), *Between Text and Image: Updating Research in Screen Translation*, John Benjamins, Amsterdam–Philadelphia (2008), pp. 37–51.

Valentini and Linardi 2009 Valetini, C. and Linardi S. “Forlixt 1: A Multimedia Database for AVT Research”, *inTRAlinea*, Special Issue: The Translation of Dialects in Multimedia (2009), http://www.intralinea.org/specials/article/Forlixt_1_A_multimedia_database_for_AVT_research.

Van Zyl and Botha 2016 Van Zyl, M. And Botha Y. “Stylometry and Characterisation in *The Big Bang Theory*”, *Literator*, 37 (2), a1282 (2016), http://dx.doi.org/10.4102/lit.v37i2.1282.

Veirano Pinto 2014 Veirano Pinto M. “Dimensions of Variation in North American Movies”. In T. Berber Sardinha, M. Veirano Pinto (eds), *Multi-Dimensional Analysis, 25 Years On: A Tribute to Douglas Biber*, John Benjamins, Amsterdam–Philadelphia (2014), pp. 109–147.

Vickers 2002 Vickers, B. *“Counterfeiting” Shakespeare: Evidence, Authorship, and John Ford's Funerall Elegye*, Cambridge University Press, Cambridge (2002).

Vickers 2011 Vickers, B. *Authorship Attribution Studies and Shakespeare's Canon (A Review of Shakespeare, Computers, and the Mystery of Authorship*, *Shakespeare Quarterly* 62 (1) (2011): 106-142.

Zago 2016 Zago, R. *From Originals to Remakes. Colloquiality in English Film Dialogue Over Time*, Bonanno Editore, Acireale/Roma (2016).

Zago 2018 Zago, R. *Cross-Linguistic Affinities in Film Dialogue*, Sike, Roma (2018).