2015

Center of Attention: A Network Text Analysis of American Sniper

Starling David Hunter
Carnegie Mellon University, starling@andrew.cmu.edu

Susan Smith
American University of Sharjah

Center of Attention: A Network Text Analysis of *American Sniper*

Starling Hunter
Carnegie Mellon University in Qatar

Susan Smith
American University of Sharjah

Abstract: Network Text Analysis (NTA) involves the creation of networks of words and/or concepts from linguistic data. Its key insight is that the position of words and concepts in a text network provides vital clues to the central and underlying themes of the text as a whole. Recent research has relied on inductive approaches to identify these themes. In this study we demonstrate a deductive approach that we apply to the screenplay of *American Sniper*, an Academy Award nominee for Best Adapted Screenplay in 2014. Specifically, we first use genre expectations theory to establish prior expectations as to the key themes associated with war films. We then empirically test whether words and concepts associated with the most influentially-positioned nodes are consistent with themes common to the war-film genre. As predicted, we find that words and concepts associated with the least constrained nodes in the text network were significantly more likely to be associated with the war, action, and biography genres and significantly less likely to be associated with the mystery, science-fiction, fantasy, and film-noir genres.

1. Introduction

Network text analysis (NTA) is a term used to describe a wide variety of "computer supported solutions" that model a text as a network "of words and the relations between them" (Diesner & Carley, 2005, p. 83). Constructing these networks is a four-step process and differences in how each is performed account for much of the variety to be found in approaches to NTA (Diesner, 2012). The first step involves the *selection* of which words are to be included or excluded in the analysis. The second step involves the *abstraction* of the included words to higher-order conceptual categories. In the third step, *connections* are established between pairs of related concepts. Subsequent analysis of the resulting network involves a fourth step—the identification or *extraction* of the key themes. Like other forms of content analysis, NTA explicitly assumes that structure encodes meaning (Fischer-Starcke, 2009). Where it differs from traditional content analytic approaches, i.e. those concerned with word-frequency, is that meaning is encoded in the structure of the network. Most specifically, in the extraction phase of NTA, prime importance is placed upon the position or role of concepts within the text network. In short, the more influential the network roles and position occupied by concepts, the greater the assumed thematic or semantic relevance they are assumed to have.

A number of recent studies have focused attention on the last step. In the last five years alone these include network text analyses of *abstracts* of academic journal articles (Beam, et al, 2014), medical school *mission statements* (Grbic, Hafferty, & Hafferty, 2013), *presidential inaugural addresses* (Light, 2014), *violent extremist propaganda* (Morris, 2014), *screenplays* and *novellas* (Hunter & Singh 2015; Hunter & Smith, 2014), *press releases* (Nerghes, Lee, Groenewegen, & Hellsten, 2014), *energy policy speeches* (Shim, Park, & Wilding, 2015), as well as newspaper articles about the *global financial crisis* (Nerghes, Groenewegen, & Hellsten, 2015), *mad cow disease* (Lim, Berry, & Lee, 2015), the *creationism debate* in the US (Shortell, 2011) and two major cities in *Afghanistan* (Martin, Pfeffer, & Carley, 2013). In these studies, measures of concept position within the text networks include *degree centrality* (Shortell, 2011; Grbic, Hafferty, & Hafferty, 2013; Martin, Pfeffer & Carley, 2013; Morris, 2014;) , *betweenness centrality* (Light, 2014; Shim, Park, & Wilding, 2015; Nerges, Groenewegen, & Hellsten, 2015), and network constraint (Hunter & Singh, 2015).

Despite the wide variety of corpora and research questions, the research designs share two important features. The first is the use of exploratory or inductive methods. Put another way, falsifiable hypotheses concerning both the content and positions of the themes are rarely if ever formulated and tested. Rather, the approach has been to select texts on a particular topic, generate semantic networks therefrom, identify the words/concepts occupying the most influential network positions or roles, and for the purposes of the ensuing analysis treat those words/concepts as indicators of the most important themes. For example, Shim, Park & Wilding

(2015) undertook to "explore and compare nuclear energy policy frames" in six countries – Japan, South Korea, USA, UK, France, and Germany—in the two years preceding and two years following the March 2011 Fukushima accident. They created semantic networks from "the speeches and addresses made by top policy makers" in each of the six countries. Their subsequent analysis found many important differences both with and across countries, as well as over time. But the analysis was exploratory and no falsifiable hypotheses or propositions were formulated concerning the differences in the centrality of concepts across the various analytical frames.

Similarly, Beam, Applebaum, Jack, et al (2014) undertook to map the semantic structure of "cognitive neuroscience" a field that links the "biological systems" investigated by neuroscientists to the "processing constructs" investigated by psychologists. Notably, their investigation revealed many significant instances of "negative" structure—what they termed "islands" and "gaps"—as well as "positive" structure—what they termed "hubs" and "branches" (Beam, Applebaum, Jack, et al, 2014, p. 1958). However, despite the clearly stated role of cognitive neuroscience as a "linking discipline", the authors offered no predictions concerning either which concepts would be central in each of the three domains or which concepts, if any, would serve as linchpins among them.

Two recently studies have undertaken an inductive approach concerning the network position of concepts. One of them is Grbic, Hafferty, & Hafferty (2013, p. 853) who studied the semantic structure of mission statements of 132 US medical schools. They divided the schools into four types—research-focused, social mission-focused, public, and private. Their hypothesis was that "key differences in institutional identity and purpose are projected through (medical) school mission statements, " differences that would "become apparent particularly under the lens of a network approach to text analysis." And as predicted they found that the top 10 most central themes—terms like *leader, biomedical, health, research,* and *community*—across the four hospitals differed. They did not, however, predict what any of the most central themes would be. Nor did they specify a priori *how* the most central concepts would differ across the four types of medical schools—only that they would differ.

Another study adopting a deductive approach is Hunter & Singh's (2015) analysis of the screenplay of the 1999 film *Fight Club* which starred Brad Pitt and Edward Norton Jr. They selected the film not only because of its status as a cult classic, but also because the film and the novel upon which it is based have been the subject of over 100 peer-reviewed journal articles with emphases on literary criticism, film studies, religion, philosophy, media and culture, race and ethnicity, gothic studies, psychotherapy, and the sociology of sport. And it was from the abstracts of such journal articles that they first identified thirteen prominent themes prominent in the academic discourse about the film. The four most frequently occurring in the sample of 52 abstracts were gender, social and individual identity, capitalism, and anarchism. As expected,

they found these themes to be clearly associated with the most central and least constrained nodes in the morpho-etymological network that they constructed from the screenplay's text. That said, it is important to note that neither Hunter & Singh (2015) nor Grbic, Hafferty, & Hafferty (2013) offered theoretically-grounded justifications for their predictions. And that is where the present study stands to contribute to the existing literature concerning the extraction of important themes. Our is the first of which we are aware that grounds our predictions in such a manner. As detailed in the next section, in this study we use genre expectations theory (Bignell, 2002; Altman, 1984; Eberwein, 2009) to determine *a priori* what themes should be prevalent in a war film in general and in an Iraq-war film in particular. We then test whether those themes are associated with centrally or influentially positioned nodes in the text network that we construct from the text of the screenplay. We then externally validate those results by having survey respondents attempt to determine the film's genre based only on their examination of words and concepts associated with the most influentially positioned nodes in the network.

2. Literature Review & Hypotheses

The *American Heritage Dictionary of the English Language* defines the word *genre* both as "a type or class" and more specifically—in reference to the arts—as "a category of artistic composition…marked by a distinctive style, form, or content." In film studies the term "genre analysis" refers to the classification of films into recognizable groups and types—e.g. comedy, drama, science-fiction, and horror—as well as the study of the "codes and conventions" that define them (Bignell, 2002, p. 199). Many aspects of a film can convey its genre. These include, but are not limited to the story structure (plots, characters, issues, situations), locations and backdrops, props, the narrative style, dialog, lighting, emphasized camera shots, the musical score and sounds, lighting, etc. (Bignell, 2002; Altman, 1984). "Genres have characteristic features that are known to and recognized by audiences." For example, in a Western we see similar characters, situations, and settings, e.g. native Americans, settlers and homesteaders, horses and men on horseback, stagecoaches and covered wagons, guns and gunfights, corrals and ranches, wilderness and wide open spaces.

Taken together, all of these things—and more—"offer the audience a set of expectations" and allow the film producer a template upon which to base its marketing and promotional discourse. (Bignell, 2002, p. 199). There is currently no universally agreed-upon set of film genres. The *International Movie Database* (imdb.com) classifies films into 21 genres which have remained relatively constant over time (http://www.imdb.com/genre/). *Box Office Mojo* (BOM) , however, currently lists 217 genres and sub-genres (http://www.boxofficemojo.com/genres/). Notably, neither site claim that their genre categories are either mutually exclusive or cumulatively exhaustive, or scientifically precise (Bordwell & Thompson, 2009). In fact, BOM explicitly states that "more genres will be added over time" and it invites readers to indicate what new (sub-) genres should be added. On both sites, a very large number of films are assigned to

more than one genre but few appear to have more than four. Although definitions of genres vary, implicit in standard definitions is the notion that films can be classified into groups that have high similarity within the groups and low similarity across them. In other words, members of a genre share particular certain conventions of "content, e.g. themes or setting" with one another to a much greater degree than they do with those belong to other genres (David Chandler).

In the introduction of his book entitled *The Hollywood War Film*, Eberwein (2009) places the "conventions" that define the genre into two categories—stock "characters" and "basic narrative elements." As summarized in in the first column of **Figure 1**, below, the characters of three types—*males, females*, and "*youth/children/pets*." The former are further subdivided into (a) "the older, seasoned leader" (b) "young recruits" (c) "camp/platoon clown (d) "ladies man" [e] ("newly married or recent father" (f) "regional, ethnic, and racial types" and (g) "examples of different social classes." The basic narrative elements are of three kinds—(1) the "basic training" that characterizes the preparation for combat (2) activities characterizing the *specific branches of the armed services* at war (3) activities or *elements common to all branches* of the armed services and where appropriate (4) the *aftermath of war*.

Insert *Figure 1: Conventions & Codes that Define War Films* Here

The seventh chapter in Eberwein's book is entitled *The Iraq Wars on Film* and it details the specific ways in which these conventions define the sub-genre of films about the first (the "Gulf" War) and second Iraq wars. These include "cramped doorways and narrow, almost impassable streets", "endless checkpoint confrontations" indicating the inability of soldiers to determine friend from foe—"suicide bombers' cars that explode" and "the gunfire that rains down from snipers above" (p. 134). In addition, "atrocities appear with frightening regularity" and crucial information about them, including images, are frequently found on "cell phones." In general, Eberwein continues, the tone of the films is "despairing" and "veterans and their families and survivors find little if any solace." Soldiers return home with drinking problems, have difficulty adjusting to prostheses, and survivors.

As noted previously, prior studies in semantic network analysis operate on the assumption that the most influentially positioned words and concepts in text network embody or illustrate the source texts' most important themes or meaning. Prior research of an exploratory kind on screenplays has already demonstrated that thematically-relevant words are associated with the least constrained and most central nodes in text networks (Hunter, 2014). One of those was *The Hurt Locker*, an Iraq War film about a bomb-disposal team working in and around Baghdad. In the text network of the screenplay, words associated with the least constrained nodes in that network included— *HUMVEE, IED (improvised explosive device), machine gun, shell-shocked, suicide bomber, army-issue, body armor, fireball, gunfire, gunshot, , UN (United Nations),* and *USA*. All of these words, and others, were illustrative of the conventions that Eberwein (2009) and others have identified as defining war films, in general, and Iraq war films in particular. As such, our first prediction is that

H1: *In a text network constructed from the screenplay of a war film, words associated with the most influentially-positioned will embody or illustrate the codes and conventions of the war genre.*

To the best of our knowledge, no prior semantic network analysis has compared the words associated with the most influentially positioned nodes with those associated with the *least* influential. However, if network position matters for meaning then differences should be evident among words associated found in positions of varying influence. As such, our second prediction is that

H2: *Among the least influentially positioned nodes in a text network constructed from the screenplay of a war film, words associated with the most influentially-positioned nodes will more accurately embody or illustrate the conventions of the war genre than do words associated with the least influentially positioned nodes.*

3. Methods and Data

The war film which we chose to analyze in this paper is *American Sniper,* an autobiographical drama based on the book *American Sniper: The Autobiography of the Most Lethal Sniper in US Military History* (Kyle & McEwen, 2013). The screenplay (Hall, 2013) was written by Jason Hall whose prior screenwriting credits include the 2009 dramedy *Spread* and the 2007 dramatic thriller *Paranoia*. The film itself was directed by Clint Eastwood and starring Bradley Cooper as Chris Kyle. The film premiered on November 11th, 2014—Veteran's Day—at the *American Film Festival Institute*. A limited release followed on Christmas Day while the wide theatrical release was on January 16th, 2015. The film has been an enormous commercial success. According to Box office Mojo, as of the weekend ending March 15th, 2015 the film had earned $342 million domestically and another $175 internationally, making it the highest-grossing film released in 2014. According to *Box Office Mojo* (2015), when adjusted for inflation *American Sniper's* box office revenues are second only to Stephen Spielberg's *Saving Private Ryan* in the war genre. The copy of the screenplay used in this paper was the latest (second) of two drafts available for purchase from online screenplay seller *Scriptfly*. Dated July 17th, 2013, the second draft is 141 pages, about 20 pages longer than the industry standard. Interestingly, a little more than two weeks after the completion of the second draft, the originally-intended director, Stephen Spielberg, withdrew from the project—supposedly due to budgetary constraints—and was replaced by Eastwood within days. Critical response to the film has been largely positive. Thirty-three of forty reviews by "top critics" collected by *Rotten Tomatoes* (2015) are classified as "fresh" with an average rating of 7.3 out of 10. The film also received six Award nominations—Best Motion Picture, Best Performance by an Actor in a Leading Role, Best Writing (Adapted Screenplay), Best Achievement in Film Editing, Best Achievement in

Sound Mixing, and Best Achievement in Sound Editing—but won only one, the latter. The screenplay itself won the *British Academy of Film & Television Arts* (BAFTA) award from Best Adapted Screenplay and was nominated for best screenplay by the *Denver Society of Film Critics*, the *Phoenix Film Critics Society*, the *Satellite Awards*, and the *Writers Guild of America*.

Recall that the four steps involved in a semantic network analysis, as detailed by Diesner (2012, pp. 90-1), are (1) *selection*, the determination of which words are to be included and excluded from consideration (2) *abstraction*, i.e. assigning the retained words to higher-level conceptual categories (3) *connection*, establishing a relationship for connecting pairs of conceptual categories and (4) and extraction, i.e. extracting or inferring meaning and key themes from the completed network. Our choices for these four steps were consistent with prior research on the semantic networks of screenplays (Hunter, 2014a, b; Hunter & Singh, 2015). In those studies, the only words that are selected are multi-morphemic compounds (MMCs), i.e. *hyphenated* and *closed compounds*, e.g. heavy-handed or shotgun; *acronyms* and *abbreviations*, e.g. *NATO, radar, laser*; *blend words*, e.g. guesstimate (guess + estimate) and motel (moter + hotel); *clipped words*, e.g. internet(work), e(lectronic)-mail; *multi-word compounds*, e.g. son-in-law, over-the-top; *copulative compounds*, e.g. actor/model, attorney-client; and *open compounds*, e.g. trade secret and post office. The text contained 5,812 unique words repeated 28,361 times. The text contained 453 multi-morphemic compounds, about 7.8% of the number of unique words.

Abstraction involved the assignment of each element of each MMC to a category defined as its etymological root. The source used to determine these roots was the 3rd edition of the Watkins' (2011) *American Heritage Dictionary of Indo-European Roots* (AHDIER). which traces more than 13,000 English words back to over 1,300 Indo-European roots. For example, word *shotgun* is comprised of two morphemes—*shot* and *gun*. According to the AHDIER, the former descends from the IE root ***skeud-*** which means " to shoot, chase, throw" (p. 81). The latter descends from the IE root ***gwhen-*** which means "to strike, kill" (p. 36). When Indo-European (IE) roots were not identified, then Greek, Latin, Semitic, or other roots are used, as provided in words' etymology in the *American Heritage Dictionary of the English Language* (AHDEL). Because no software exists that etymological stems words in this fashion, the mapping had to be performed manually. At the conclusion of this process the 453 MMCs were traced back to 402 unique roots, 302 of which were Indo-European.

The relationship between a word and its etymological root is genetic in that it suggests that the former descends from the latter. The choice of relationship used to connect these 403 roots was the co-occurrence of descendants of two or more roots within the same MMC. As shown above, the word *shot* descends from *skeud-* and gun from *gwhen-*. In semantic network of the screenplay of American Sniper, these two roots are linked or connected because their descends co-occur within the same MMCs—both *shotgun* and *gunshot*. And because roots have many descendants that may co-occur with many other descendants of other roots, the result is a semantic network where the nodes are etymological roots and the linkages representing the MMCs in which the roots co-occur. The 402 nodes in the semantic network for the screenplay of

American Sniper were connected by 485 links. As shown in **Figure 2**, the main component of the network—i.e. the largest group of interconnected or mutually-reachable nodes—contained 309 nodes connected by 404 linkages.

Insert Figure 2 Here: Main Component

4.1 Identifying Key Themes

As discussed above, prior research in semantic network analysis has relied upon a variety of node-level measures to identify the most influentially-positioned concepts in a text network, degree and/or betweenness centrality being the most common (Nerges, Groenewegen, & Hellsten, 2015). Following other semantic analyses of screenplays, we rely upon network-level measure of constraint (Burt, 2000) which captures the degree to which a node serves to link otherwise disconnected segments of a network. But we also use this measure in a way not seen in prior work. Recall that in all other semantic network analyses reviewed above—and all of which we are more generally aware—words are nodes in the networks whereas in the morpho-etymological approach the nodes are etymological roots and the words are associated with the edges. As such, while the determination of the most influentially-positioned nodes is a relatively straight-forward process, the only contingency or element of uncertainty being which of many measures of influence to use. However, in a method where words are on the edges, the matter is less straight-forward. In this study we define the most influentially-positioned MMCs as being those that link two or more influentially positioned nodes, i.e. two or more of the least constrained nodes. More specifically, we classify as most important the MMCs that appear on the linkages between nodes forming the main component of the sub-network formed only by nodes with the lowest constraint scores. We defined low constraint as scores in the bottom 10% of the sample . In practice this amounted to 54 nodes with constraint values of less than or equal to 0.25. As shown in Figure 3, below, the main component this sub-network had 43 nodes.

Insert *Figure 3* Here

Excluding prepositions and pronouns, these nodes were linked by the following 62 MMCs:

> *air-raid, asshole , back-and-forth, backseat, back-up, bonfire, bull's-eye, bullshit, daylight, dog-ass, downrange, eyeball, firelight, football, forehead, GPS (Global Positioning System), gunfire, handstand, hand-to-hand, hard-headed, headlights, head-on, headshots, Hellfire, HUMVEE, JDAM (Joint Direct Attack Munition), JTAC (Joint Terminal Attack Controller), off-eye, on-board, outer-hallway, outpost, outrank, overhead, overweight, poster-boy, ranch-hand, Ranger-One, ringside, roadside, roadway, ROEs (rules of engagement), roll-back, settling-in, set-up, shithole, shotgun, sideways, stand-down, Sunday, sunlight, sunset, sun-up, today, underfoot, understand, uprange, upright, upset, upside, US (United States), white-board,* and *white-side*

Conversely, we classified as less important the words associated with isolates in the network formed by words with the *highest* constraint values. We defined high constraint as nodes with values equal to 1.0, the theoretical maximum. Notably, exactly 50% of the 402 nodes in the semantic network that assumed this value. Note that because the nature of the relationship used to link concepts, there are two kinds of isolates—reflexive and non-reflexive. two-node. The former occur when both elements of an MMC descend from the same etymological root. An example is pronoun *anyone*. Both *any* and *one* descend from the same IE root, ***oi-no-*** which means "one, unique" (Watkins, 2011, p. 61). The other instance is found when descendants of a pair of nodes occur only once. The hyphenated compound *middle-east* is a typical example. The word *middle* descends from the IE root *medhyo-* which means "middle" (Watkins, 2011, p. 53) while the word *east* descends from the IE root *aus-1* which means "to shine" (p. 6). The words *middle* and *east* are the only descendants of these roots found in any MMC in the entire screenplay. As such, the roots ***aus-1*** and ***medhyo-*** are connected to one another but to no others. Thus, they form an isolated pair. As shown in Figure 4, below, there were 31 such isolated pairs of nodes. The 31 MMCs associated them are:

> *baby-crib, breastfeeding, chest-full, cob-nosed, concertina-wire, cornhusker, dead-sprint, duct-taped, eardrums, fingernail-sized, flack-jacket, hash-marks, horse-shoe, ill-at-ease, middle-east, mind-melting, mini-van, now-naked, otherworldy, pepper-flake, pinpricks, playbook, plywood, rattlesnake, rifle-barrel, rush-hour, taxi-cab, trigger-slack, voicemail, warfare, well-worn, whisper-mic(rophone)*

Insert Figure 4 Here

Several MMCs from both networks typify several of the codes and conventions of war films defined by Eberwein (2009). For example, among the low-constraint MMCs, the terms *cornhusker* and *ranch-hand* were referenced with regard to the "regional, ethnic, racial type" of "male character" while *downrange* appeared in the screenplay in the context of "demanding exercises, drills" under the "Pre-Combat: Basic Training" category. The hyphenated compound *settling-in* is an example of one of the described as while *outer-hallway* is an example of the "cramped doorways and narrow, almost impassable streets" that Eberwein said uniquely distinguish the Iraq-era war films from those about Viet Nam and World War II. Another low-constraint MMC was *uprange* which occurs in the context of combat and in "Post-Combat: Aftermath of War" categories. In the latter case, the term is employed in scenes where Kyle spends time with injured fellow veterans at a shooting range, thus typifying the "recovery/rehab for physical (or) psychological injuries" convention. Finally, the terms *Range-One* (the name for a detachment of Army Rangers), *JDAM* (Joint Direct Attack Munition), *JTAC* (Joint Terminal Attack Controller), *Hellfire* (missile), *GPS*, *headshot*, *HUMVEE*, *ROE*s (rules of engagement), *settling* in, off-*eye* (the eye of the sniper that is not looking through the rifle scope), *gunfire*, *stand-down*, *hand-to-hand*, and *on-board* are all low-constraint MMCs that typify or signify one or more first three of Eberwein's "Combat" conventions : "water landings, patrols, ambushes,

raids, digging in" and other maneuvering; combat in the desert; and "tanks, grenades, and flamethrowers" and other vehicles and weapons of war.

Although there are many fewer of them, there were also a number of highly-constrained MMCs that—when taken together—could convey the sense of combat, armed conflict, and or the armed forces more generally. One of these was the closed compound *warfare* which appeared in the screenplay as part of the proper noun *Naval Special Warfare Center*. Notable, had that name's acronym been used instead, the resulting MMC would have been much less highly constrained than was *warfare* itself. Other highly-constrained words were *chest-full* (which appeared in the phrase "a chest-full of medals"), *concertina-wire* (a type of barbed or razor wire formed in large coils and commonly used around prisons and military installations), *duct-taped, flack-jacket, middle-east, ill-at-ease, rifle-barrel, trigger-slack,* and *hash marks* (a service stripe on the sleeve of an enlisted person's uniform). Finally, two other highly-constrained MMCs were *baby-crib* and *breast-feeding* which were relevant in the context of the "newly-married or recent father" (male character) and the "loyal wife" (female character).

4.2 Survey Results

In addition to the qualitative analysis described above—an analysis which is typical of the semantic network analyses earlier reviewed—we opted to further validate our coding with an approach not previously undertaken in any study of which we are aware. Specifically, developed a survey that would allow us to directly compare how well the two sets of MMCs convey not just the codes and conventions of the war genre, but also the other genres to which the film belongs and does not belong.

The first step was to divide the 93 words into three groups—one comprised of 31 randomly-selected low-constraint MMCs, another consisting of the remaining low-constraint MMCs, and a third group consisting of the highly-constrained MMCS. Those lists appear in the appendix along with screenshot of the survey as it appeared on *Survey Monkey*. We then recruited a sample of respondents from Amazon.com's *mTurk* e-worker service (mTurk.com). All survey respondents were located in the USA, had previously completed at least 5000 human intelligence tasks (HITs) and had a 98% or better approval rates from other employers. Respondents were told in the introduction to the survey that they would be matching keywords extracted from the screenplay of a film to types or genres of films. After viewing just one of the three groups of keywords, respondents were asked to answer a series of 20 questions, each of which provided a definition of a genre and which required the respondent to rate on a 1-10 scale the likelihood that the resulting film belonged to that genre. The question specific to the war genre read as follows: "How likely is it that this list of words was taken from a WAR film, i.e. one that contains numerous scenes and/or a narrative that pertains to a real war (i.e., past or current)." The question regard westerns was worded similarly: "How likely is it that this list of words was taken from a WESTERN film, i.e. a film that contains numerous scenes and/or a narrative that portrays frontier life in the American West during 1600s to contemporary times." The eighteen other genres about which the respondents provided opinions were ACTION, ADVENTURE,

ANIMATION, BIOGRAPHY, COMEDY, CRIME, DRAMA, FAMILY, FANTASY, FILM-NOIR, HISTORY, HORROR, MUSIC/MUSICAL, MYSTERY, ROMANCE, SCIENCE-FICTION, SPORT, and THRILLER. Their definitions embedded in the questions were taken directly from the *International Movie Database*.[1] Consistent with our qualitative analysis, we found the predicted relationship between the network position of concepts and the film's genre. Table 1 contains the results of our survey.

Insert ***Table 1*** Here

Columns 2-5 contain the average scores given by respondents to the question concerning the film's membership in a given genre. The asterisks indicate the significance value of the β-coefficients of OLS regressions where the dependent variable, $SCORE_{Genre}$ is the score given (on a 1-10 scale) by respondents to the question of whether film belongs to a given genre; where LOWCON1 is a dummy variable equal to 1 if the respondent viewed the first set of low-constraint keywords and 0 otherwise; where LOWCON2 is a dummy variable whose value is 1 of the respondent viewed the second set of low-constraint keywords and 0 otherwise

$$(1)\quad SCORE_{Genre} = \alpha + \beta_1 * LOWCON1 + \beta_2 * LOWCON2 + \varepsilon$$

Descriptive statistics and correlations are presented in Table 2.[2]

Insert *Table 2: Descriptive Statistics and Correlation Matrix* Here

In the first row of Table 1 we see that when asked if the film belonged to the WAR genre, responses from who viewed the "high-constraint" words averaged 6.11 points (on a scale of 1-10). Responses of those whose viewed the first and second groups "low-constraint" words averaged 7.37 and 9.00, respectively. The former score was significantly higher—at the $p < 0.05$ level, 2-tailed—than 6.11 while the latter score was even more significantly higher ($p < 0.0001$, 2-tailed). Similarly, the average score given to the ACTION genre by those reviewing the high-constraint words was 6.80 while averages by those viewing these low-constraint words were 8.06 and 8.41, both of which were significantly higher.

But whereas respondents who saw either group of low-constraint MMCs were far better able to identify the film as belonging to the WAR and ACTION genres, they were not able to better identify the other two genres that IMDb assigned to the film—BIOGRAPHY and DRAMA. Respondents viewing either low-constraint group were also better able to determine the genres to which the film did *not* belong. Specifically, they were much better able to tell that the film did not belong to the FAMILY, WESTERN, ROMANCE, MYSTERY, FANTASY, NOIR, HORROR, MUSIC, COMEDY, ANIMATION, or SPORTS genres. They were slightly better

[1] http://www.imdb.com/help/search?domain=helpdesk_faq&index=2&file=genres
[2] Remark on VIF, model R2 and F.

able to tell that the film did not belong to SCI-FI, CRIME, or HISTORY genres. Finally, they were no better able to determine that the film did not belong to THRILLER or ADVENTURE genres.

4. CONCLUSION

As noted in the introductory section of this paper, the extraction of meaning from text or semantic networks involves an examination of the most influentially positioned nodes in the network. That said, we are aware only one other study wherein the themes associated with a text network's most influential nodes were specified *a priori*. Thus, the present study is distinguished from most of the prior literature in this regard. Where the present study is most distinctive concerns the theory we used to generate our falsifiable two hypotheses. Specifically, it was genre theory within the broader film studies literature, and research on the war-film genre that we applied to the screenplay of *American Sniper*. As discussed in the preceding section, our results supported both hypotheses, i.e. that the words associated with the most influentially-positioned nodes would embody the codes and conventions of the war film genre and that they would so more accurately than words associated with the least influentially-positioned nodes. Recall that we used network constraint (Burt, 2000) as our measure of positional influence and found that several words associated with the subset of least constrained—and thus most influentially-positioned—nodes were clearly associated with the codes and conventions of war films, words like *air-raid, bull's-eye, gunfire, hand-to-hand, Hellfire, HUMVEE, JDAM (Joint Direct Attack Munition), JTAC (Joint Terminal Attack Controller), outrank, ROEs (rules of engagement), shotgun,* and *stand-down.* In marked contrast, our reading of the words associated with the most constrained—and thus least influentially-positioned—nodes revealed that many fewer evoked the codes and conventions of the war genre. But unlike prior studies of this kind, we also externally validated our impressions of these two sets of words. Specifically, we administered surveys to over 100 participants recruited through Amazon's mTurk service and, as discussed above, we found further and stronger support for both hypotheses. To the best of our knowledge, ours is the only network text analysis that has validated its findings in such a manner.

We should note that there are several limitations to this study that should be explicitly recognized, limitations that may place bounds around on the generalizability of the result. The first concerns the nature of the text analyzed. Whereas all other network text analyses of which we are aware use non-fiction, we used a screenplay, one adapted from an autobiography. This is important because contemporary screenplays adhere to very well-defined set of story-telling conventions, plot devices, and narrative structures (Field, 2005; Snyder, 2005) and are characterized by a level or thematic and lexical repetition not common to texts in other domains (Hunter& Smith, 2013). Secondly, we analyzed a screenplay from a film genre that is long-standing and well-defined in the minds of American movie-goers and media consumers. It's possible that widespread familiarity with war films enhanced survey respondents' ability to

correctly identify the genre and it's an open question as to whether a romantic comedy or a family drama would be so easily identified through a similar approach. Third, we should note that the method upon which we based our findings relies on a single source for tracing multi-morphemic compounds back to their etymological roots—the *American Heritage Dictionary of Indo-European Roots* (Watkins, 2011). Recall that approximately 75% of the individual morphemes in the network model were traced back to that source. It is possible that some of the remaining 25% could have been traced to common roots—Indo-European or otherwise—described and defined in other well-known sources, e.g. the *Barnhart Concise Dictionary of Etymology* (Barnhart, 1995) or *The Concise Oxford Dictionary of English Etymology* (Hoad, 1993),. Finally, we should note that it is unclear whether and to what degree the approach described here can be applied to languages other than English. While it is widely accepted in comparative linguistics that compounding is a common to all languages, the same compound words are not always used across languages to describe the same thing. For example, the English compound *butterfly* is *papillon* in French. This suggests that when working with texts translated from other languages, a one-to-one correspondence between MMCs will not be achieved and that, as such, the network structures of the "same" text will differ accordingly.

REFERENCES

Altman, R. (1984). A semantic/syntactic approach to film genre. *Cinema Journal*, 6-18.

Barnhart, R. K. (Ed.). (1995). *The Barnhart concise dictionary of etymology*. HarperCollins.

Beam, E., Appelbaum, L. G., Jack, J., Moody, J., & Huettel, S. A. (2014). Mapping the semantic structure of cognitive neuroscience. *Journal of cognitive neuroscience*, *26*(9), 1949-1965.

Bignell, J. (2002). *Media semiotics: An introduction*. Manchester University Press.

Box Office Mojo (2015), Retrieved April 23, 2015, from http://www.boxofficemojo.com/movies/?id=savingprivateryan.htm

Burt, R. S. (2000). The network structure of social capital. *Research in organizational behavior*, *22*, 345-423.

Diesner, J., & Carley, K. M. (2005). Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. *Causal mapping for information systems and technology research: Approaches, advances, and illustrations*, 81-108.

Diesner, J. (2012). *Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts*. Carnegie-Mellon University, Pittsburgh PA. Institute of Software Research International.

Eberwein, R. (2009). *The Hollywood war film* (Vol. 13). John Wiley & Sons.

Field, S. (2005). Screenplay: The foundations of screenwriting. New York: Delta.

Fischer-Starcke, B. (2009). Keywords and frequent phrases of Jane Austen's Pride and Prejudice A corpus-stylistic analysis. *International Journal of Corpus Linguistics*, *14*(4), 492-523.

Grbic, D., Hafferty, F. W., & Hafferty, P. K. (2013). Medical school mission statements as reflections of institutional identity and educational purpose: A network text analysis. *Academic Medicine*, *88*(6), 852-860.

Hoad, T. F. (Ed.). (1993). *The concise Oxford dictionary of English etymology*(p. 210). Oxford: Oxford University Press.

Hall, J. (2013). *American Sniper.* Retrieved on 04 Feb. 2015 from http://pdl.warnerbros.com/wbmovies/awards2014/pdf/as.pdf

Janszen, K. (2000). *A Walk to Remember*. Retrieved from http://www.imsdb.com/scripts/Walk-to-Remember,-A.html

Hunter, S. (2014). A Novel Method of Network Text Analysis. *Open Journal of Modern Linguistics*, *4*(2), 350-66.

Hunter, S., & Singh, S. (2015). A Network Text Analysis of Fight Club. *Theory and Practice in Language Studies*, *5*(4), 737-749.

Hunter, S., & Smith, S. (2013). Thematic and Lexical Repetition in a Contemporary Screenplay. *Open Journal of Modern Linguistics*, *3*(01), 9-19.

Hunter, S., & Smith, S. (2014). A Network Text Analysis of Conrad's Heart of Darkness. *English Linguistics Research*, *3*(2), p39.

Kyle, C., & McEwen, S. (2013). *American Sniper: The Autobiography of the Most Lethal Sniper in US Military History*. alima.

Light, R. (2014). From Words to Networks and Back: Digital Text, Computational Social Science, and the Case of Presidential Inaugural Addresses. *Social Currents*, 2329496514524543.

Lim, S., Berry, F. S., & Lee, K. H. (2015). Stakeholders in the Same Bed with Different Dreams: Semantic Network Analysis of Issue Interpretation in Risk Policy Related to Mad Cow Disease. *Journal of Public Administration Research and Theory*, in press.

Martin, M. K., Pfeffer, J., & Carley, K. M. (2013). Network text analysis of conceptual overlap in interviews, newspaper articles and keywords. *Social Network Analysis and Mining*, *3*(4), 1165-1177.

Morris, T. (2014). Networking vehement frames: neo-Nazi and violent jihadi demagoguery. *Behavioral Sciences of Terrorism and Political Aggression*, *6*(3), 163-182.

Nerghes, A., Hellsten, I., & Groenewegen, P. (2015). A Toxic Crisis: Metaphorizing the Financial Crisis. *International Journal of Communication*, *9*, 27.

Nerghes, A., Lee, J., Groenewegen, P., & Hellsten, I. (2014). The shifting discourse of the European Central Bank: Exploring structural space in semantic networks. *Proceedings of SITIS*, 447-455.

Rotten Tomatoes (2015). Retrieved April 23, 2015, from http://www.rottentomatoes.com /m/american_sniper/

Shim, J., Park, C., & Wilding, M. (2015). Identifying policy frames through semantic network analysis: an examination of nuclear energy policy across six countries. *Policy Sciences*, *48*(1), 51-83.

Shortell, T. (2011). The conflict over origins: A discourse analysis of the creationism controversy in American newspapers. *Mass Communication and Society*, *14*(4), 431-453.

Snyder, B. (2005). Save the cat! The last book on screenwriting you'll ever need. Studio City, CA: M. Wiese Productions.

Watkins, C. (Ed.). (2011). *The American Heritage Dictionary of Indo-European Roots*. Houghton Mifflin Harcourt.

Figure 1: Codes and Conventions that Define the War Film Genre—Adapted from Eberwein, (2009)

MALE CHARACTERS

- Older, seasoned leader
- Young recruits
- Camp/platoon clown
- Ladies' man
- Newly married or recent father
- Regional, ethnic, & racial types
- Different social classes

FEMALE CHARACTERS

- Loyal wife, girlfriend, nurse
- Prostitute, floozie
- Wise, sustaining mother

YOUTH, CHILDREN, & PETS

- Eager brothers, boys
- Younger sisters
- Endangered or killed child
- Animals (dogs, cats)

PRE-COMBAT: BASIC TRAINING

- Tyrannical squad leader
- Demanding exercises, drills
- Bonding, pranks
- Weekend passes
- Sexual initiation
- Successful graduation, completion of training

COMBAT: Army/Infantry/Marines

- Water landings, patrols, ambushes, raids, digging in
- Combat in jungles, deserts, mountains
- Tanks, grenades, flamethrowers
- Dealing with heat/cold (or elements)

ELEMENTS COMMON TO ALL BRANCHES

- Writing letters; receiving mail from home (typically birth announcements and "Dear John" letters.
- Sharing and observing photographs
- Listening to the radio
- Spontaneous and improvised play to alleviate tension and boredom
- Singing; prayers/church service; communion
- Burials with short, moving eulogies & tributes
- Leaves and R&R
- Reflections on the nature of the enemy

POST-COMBAT: AFTERMATH OF WAR

- Recovery/rehab for physical/psychological injuries
- Difficulty adjusting to civilian life
- Reunion with wife, girl, family, friends

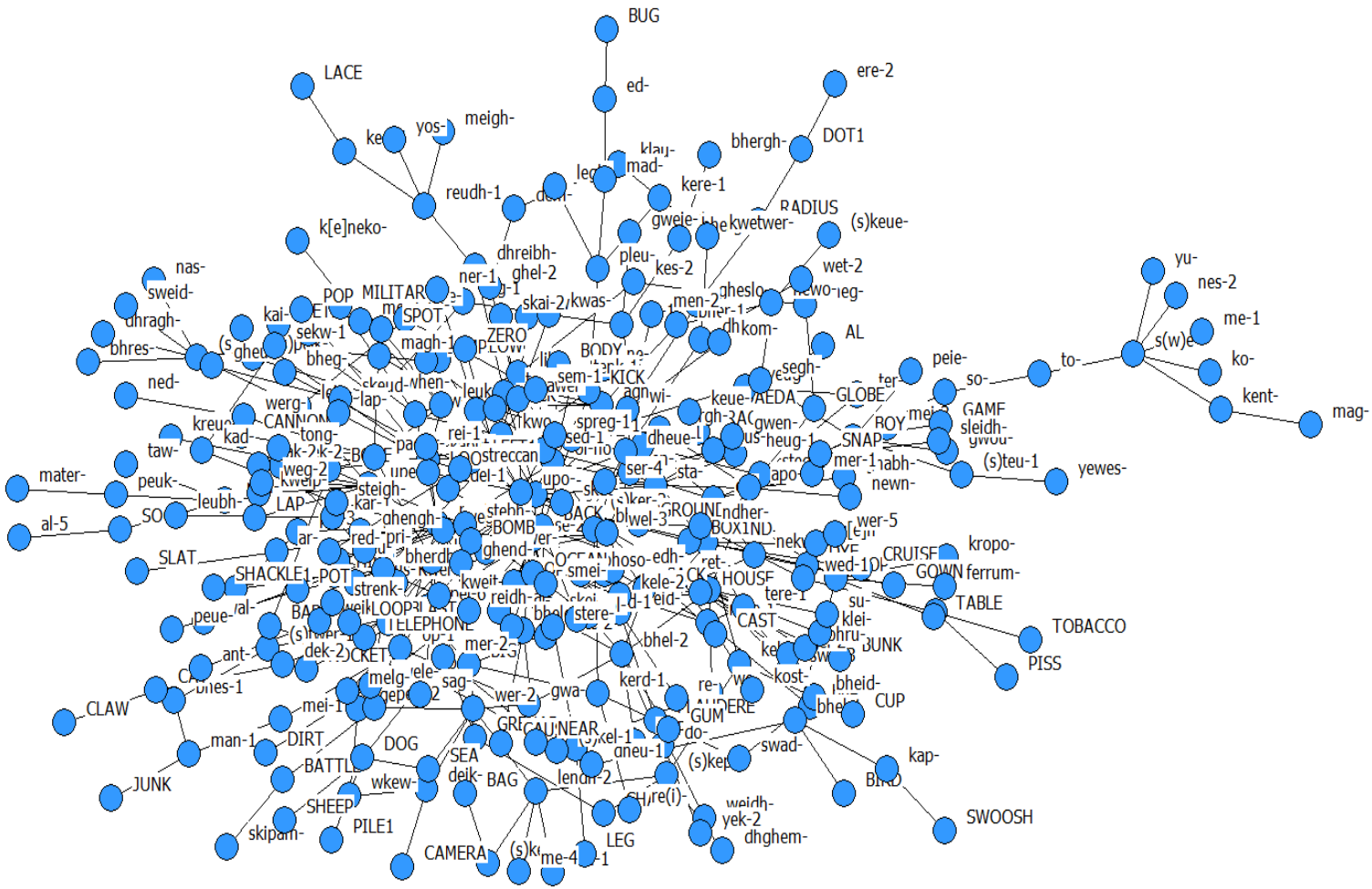


Figure 2: Main Component of the Morpho-Etymological Network of the Screenplay of American Sniper

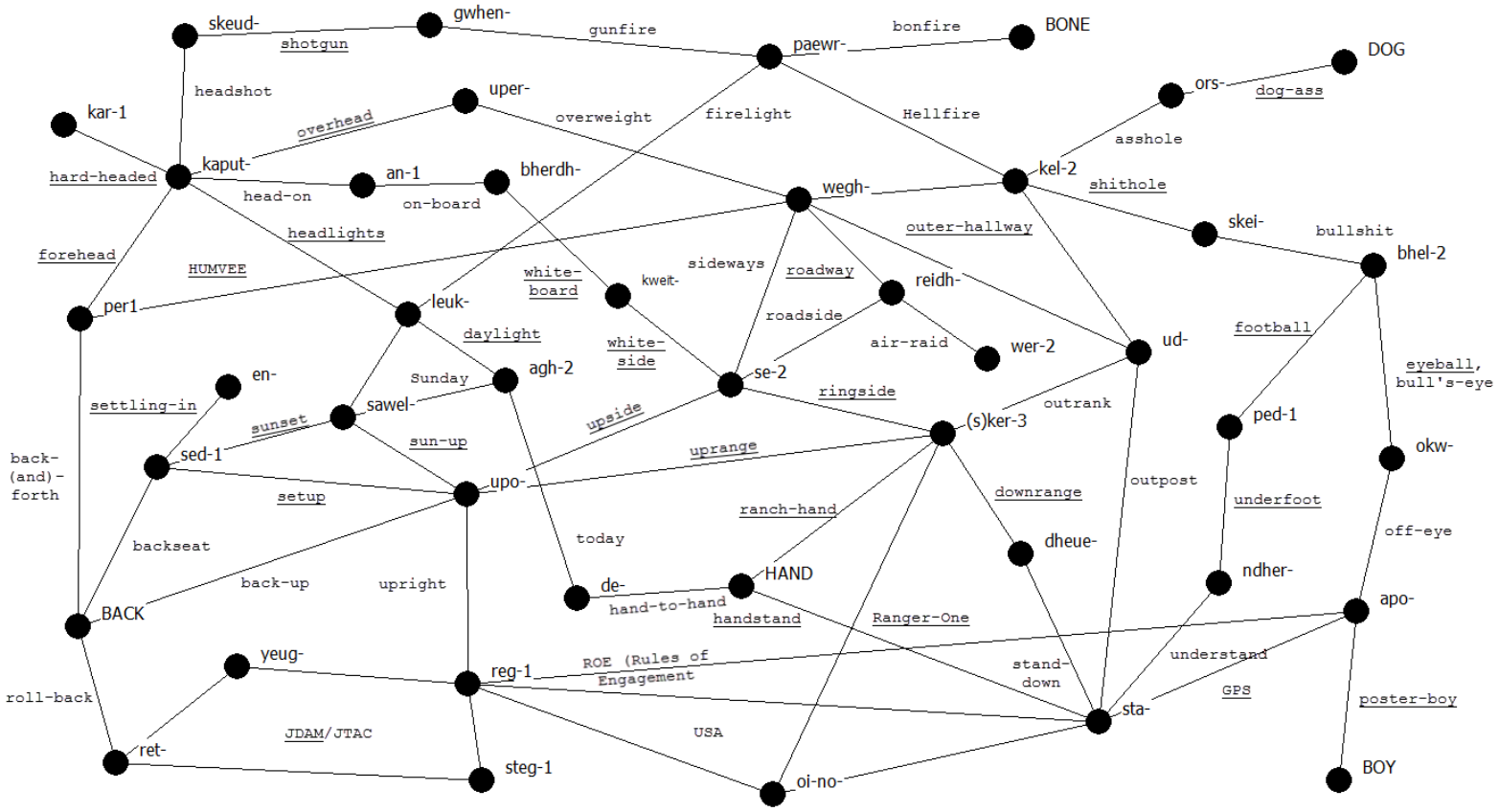


Fig.3 Main Component of Network Formed by the Nodes in the Lowest Decile of Network Constraint ($c \leq 0.25$) and the Multi-Morphemic Compounds (MMCs) Associated with Them

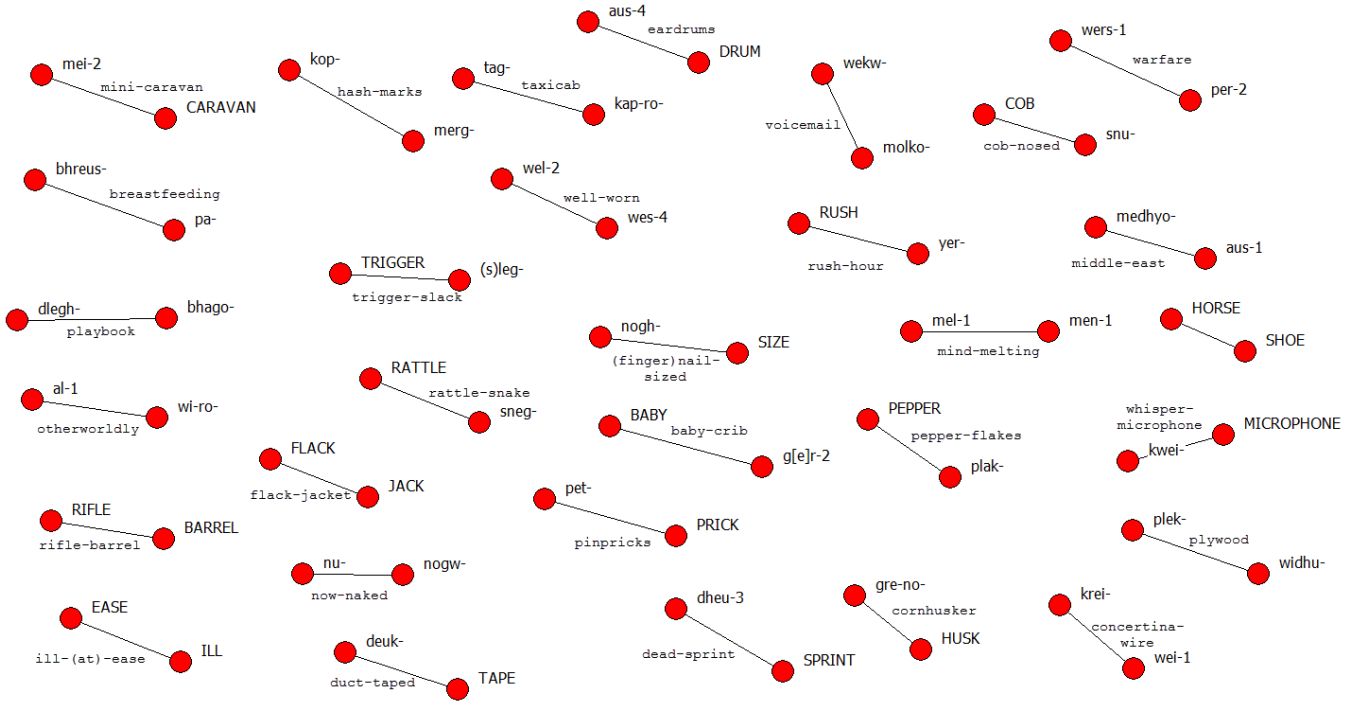


Fig.4 Network Formed by the Nodes with the Top 10% Scores for Network Constraint (c = 1.0)

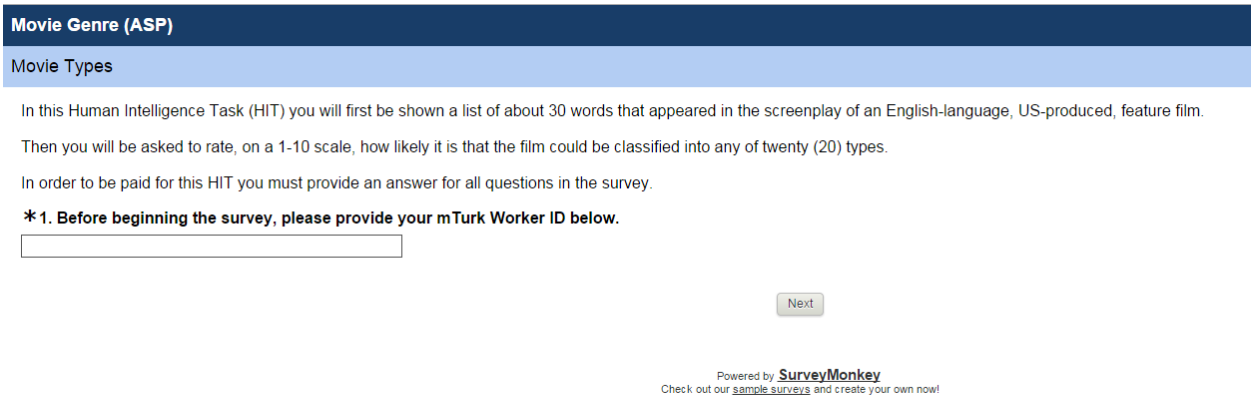
Table 1 Results of Survey & Regression Model

Genre	High Constraint	Low Con1	Low Con2	Model F-statistic	Model adj-R^2
WAR	6.11	7.37*	9.00****	7.74***	16.7%
ACTION	6.80	8.06**	8.41***	6.15***	13.3%
BIOGRAPHY	3.51	5.14***	4.50*	4.37**	9.1%
DRAMA	5.94	5.06	5.53	1.01	0.0%
Mystery	4.91	2.89****	2.72****	11.69****	24.1%
Sports	2.66	3.26	1.72	4.18**	8.6%
Romance	2.54	1.74**	1.69**	3.35*	6.5%
Crime	6.00	4.31**	5.09	3.72*	7.5%
Western	4.14	2.66*	3.19	2.34	3.8%
Film noir	5.03	3.03***	3.69*	4.56**	9.6%
Family	2.43	1.63	1.38*	2.28	3.7%
Sci-fi	3.51	2.94	4.69**	3.96**	8.1%
Fantasy	3.51	2.03***	2.63*	3.96**	8.1%
Horror	4.63	3.69	3.28*	2.31	3.7%
Thriller	6.51	5.17*	5.16*	3.44*	6.8%
Comedy	3.49	3.08	2.44*	2.05	3.0%
Animated	2.51	1.94	1.94	1.91	2.6%
History	4.80	4.89	6.19*	2.28	3.7%
Music	1.97	1.69	1.41*	1.60	1.7%
Adventure	6.31	6.06	5.66	1.17	0.5%

Table 2: Descriptive Statistics and Correlation Matrix

Genre	Average	Range	St. Dev.	Correlation with LOWCON1	Correlation with LOWCON2
WAR	7.45	1 - 10	2.71	-0.021	0.389****
ACTION	7.74	2 - 10	1.86	0.126	0.246*
BIOGRAPHY	4.38	1 - 9	2.02	0.273**	0.040
DRAMA	5.51	1 - 10	2.24	-0.147	0.007
Mystery	3.53	1 - 9	2.18	-0.214*	-0.252*
Sports	2.57	1 - 8	1.97	0.254**	-0.293**
Romance	2.00	1 - 6	1.34	-0.140	-0.159
Crime	5.14	1 - 10	2.47	-0.242*	-0.012
Western	3.33	1 - 10	2.49	-0.197*	-0.040
Film noir	3.92	1 - 10	2.53	-0.257**	-0.063
Family	5.27	1 - 10	2.53	-0.109	-0.210*
Sci-fi	3.69	1 - 10	2.32	-0.233*	0.294**
Fantasy	2.73	1 - 10	1.92	-0.264**	-0.036
Horror	3.88	1 - 10	2.27	-0.063	-0.180
Thriller	5.63	1 - 10	2.36	-0.140	-0.136
Comedy	3.02	1 - 8	1.98	0.024	-0.200*
Animated	2.14	1 - 10	1.50	-0.094	-0.090
History	5.27	1 - 10	2.53	-0.109	0.248*
Music	1.70	1 - 6	1.07	-0.007	-0.184
Adventure	6.02	1 - 10	2.36	0.012	-0.104

Appendix: Survey Instrument



Movie Genre (ASP)

Movie Types

In this Human Intelligence Task (HIT) you will first be shown a list of about 30 words that appeared in the screenplay of an English-language, US-produced, feature film.

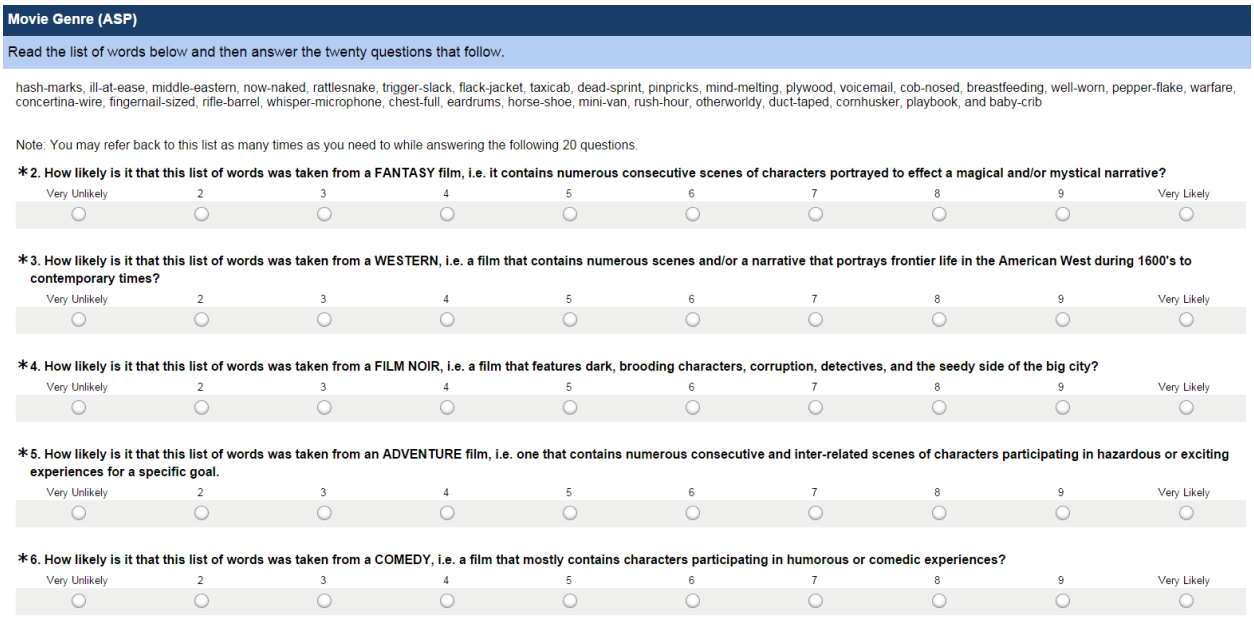
Then you will be asked to rate, on a 1-10 scale, how likely it is that the film could be classified into any of twenty (20) types.

In order to be paid for this HIT you must provide an answer for all questions in the survey.

***1. Before beginning the survey, please provide your mTurk Worker ID below.**

Next

Powered by SurveyMonkey
Check out our sample surveys and create your own now!



Movie Genre (ASP)

Read the list of words below and then answer the twenty questions that follow.

hash-marks, ill-at-ease, middle-eastern, now-naked, rattlesnake, trigger-slack, flack-jacket, taxicab, dead-sprint, pinpricks, mind-melting, plywood, voicemail, cob-nosed, breastfeeding, well-worn, pepper-flake, warfare, concertina-wire, fingernail-sized, rifle-barrel, whisper-microphone, chest-full, eardrums, horse-shoe, mini-van, rush-hour, otherworldy, duct-taped, cornhusker, playbook, and baby-crib

Note: You may refer back to this list as many times as you need to while answering the following 20 questions.

***2. How likely is it that this list of words was taken from a FANTASY film, i.e. it contains numerous consecutive scenes of characters portrayed to effect a magical and/or mystical narrative?**

Very Unlikely 2 3 4 5 6 7 8 9 Very Likely

***3. How likely is it that this list of words was taken from a WESTERN, i.e. a film that contains numerous scenes and/or a narrative that portrays frontier life in the American West during 1600's to contemporary times?**

Very Unlikely 2 3 4 5 6 7 8 9 Very Likely

***4. How likely is it that this list of words was taken from a FILM NOIR, i.e. a film that features dark, brooding characters, corruption, detectives, and the seedy side of the big city?**

Very Unlikely 2 3 4 5 6 7 8 9 Very Likely

***5. How likely is it that this list of words was taken from an ADVENTURE film, i.e. one that contains numerous consecutive and inter-related scenes of characters participating in hazardous or exciting experiences for a specific goal.**

Very Unlikely 2 3 4 5 6 7 8 9 Very Likely

***6. How likely is it that this list of words was taken from a COMEDY, i.e. a film that mostly contains characters participating in humorous or comedic experiences?**

Very Unlikely 2 3 4 5 6 7 8 9 Very Likely

Low Constraint 1: *HUMVEE, headlights, outer-hallway, shithole, JDAM (Joint Direct Attack Munition), underfoot, Ranger-One, downrange, settling-in, handstand, ranch-hand, sun-up, hard-headed, forehead, overhead, daylight, uprange, roadway, dog-ass, sunset, GPS (Global Positioning System), ringside, poster-boy, football, shotgun, upset, eyeball, upside, set-up, white-board, white-side*

Low Constraint 2: *headshots, today, air-raid, outpost, ROEs (rules of engagement), outrank, back-and-forth, bulls-eye, US (United States), Hellfire, upright, overweight, JTAC (Joint Terminal Attack Controller), back-up, hand-to-hand, on-board, off-eye, bullshit, understand, bonfire, firelight, head-on, sunlight, backseat, gunfire, sideways, stand-down, Sunday, roadside, roll-back, asshole*

High Constraint: *baby-crib, breastfeeding, chest-full, cob-nosed, concertina-wire, cornhusker, dead-sprint, duct-taped, eardrums, fingernail-sized, flack-jacket, hash-marks, horse-shoe, ill-at-ease, middle-east, mind-melting, mini-van, now-naked, otherworldy, pepper-flake, pinpricks, playbook, plywood, rattlesnake, rifle-barrel, rush-hour, taxi-cab, trigger-slack, voicemail, warfare, well-worn, whisper-mic(rophone)*