Open Journal of Modern Linguistics, 2014, 4, 350-366
Published Online June 2014 in SciRes. http://www.scirp.org/journal/ojml
http://dx.doi.org/10.4236/ojml.2014.42028

Scientific Research

A Novel Method of Network Text Analysis

Starling Hunter

Carnegie Mellon University in Qatar, Pittsburgh, USA
Email: starling@qatar.cmu.edu

Received 10 December 2013; revised 6 March 2014; accepted 20 March 2014

Abstract

This paper describes a novel method of network text analysis, one that involves a new approach to 1) the selection of words from a text, 2) the aggregation of those words into higher-order concepts, 3) the kind of the relationship that establishes statements from pairs of concepts and 4) the extraction of meaning from the text network formed by these statements. After describing the method, I apply it to a sample of the seven most recent winners of the Academy Award for Best Original Screenplay—*Little Miss Sunshine*, *Juno*, *Milk*, *The Hurt Locker*, *The King's Speech*, *Midnight in Paris*, and *Django Unchained*. Consistent with prior research, I demonstrate that structure encodes meaning. Specifically, it is shown that statements associated with a text network's least constrained nodes are consistent with themes in the films' synopses found on Wikipedia, the International Movie Database, and Rotten Tomatoes.

Keywords

Network Analysis, Network Text Analysis, Content Analysis

1. Introduction

Diesner & Carley (2005: p. 83) employ the term *network text analysis* (NTA) to describe a wide variety of "computer supported solutions" that enable analysts to "extract networks of concepts" from texts and to discern the "meaning" represented or encoded therein. The key underlying assumption of such methods or solutions, they assert, is that the "language and knowledge" embodied in a text may be "modeled" as a network "of words *and the relations between them*" (ibid, emphasis added). A second important assumption is that the position of concepts within a text network provides insight into the meaning or prominent themes of the text as a whole.

Broadly considered, creating networks from texts has two basic steps: 1) the assignment of words and phrases to conceptual categories and 2) the assignment of links to pairs of concepts. Approaches to NTA differ with regard to how these steps are performed, as well as other dimensions like the level of automation or computer support, the linguistic unit of analysis (e.g. noun or verbs), and the degree and basis of concept generalization.

How to cite this paper: Hunter, S. (2014). A Novel Method of Network Text Analysis. *Open Journal of Modern Linguistics, 4,* 350-366. http://dx.doi.org/10.4236/ojml.2014.42028

That said, the four building blocks of text networks—*concept*, *relationship*, *statement*, and *map*—are the same. Specifically, a *concept* is "an ideational kernel" often represented by "a single word⋯ or phrase" while a *relationship* is "a tie that links two concepts together". A *statement* is "two *concepts* and the *relationship* between them" while a *map* is simply a "network formed from *statements*" (Carley & Palmquist, 1992: pp. 607-608, emphasis added). A simple example will help to clarify and illustrate these terms.

Figure 1, below, is adapted from Carley (1997) and it is typical of many network representations of texts. The network itself was constructed from the following two sentences: "Organizations use information systems to handle data. Information is processed by organizations who are interested in locating behavioral trends." Several things about the network are noteworthy. First, observe that there are seven *concepts* depicted as nodes in the network, each of which appears only once. They are "organizations", "information systems", "process", "information", "interested", "locating", and "trends". Secondly, see that there are also seven *statements*, i.e. pairs of concepts: 1) "information systems" and "process", 2) "information systems" and "organizations", 3) "process" and "information", 4) "process" and "organizations", 5) "interested" and "organizations", 6) "interested" and "locating", and 7) "locating" and "trends". Third, note that the *map* itself is comprised of the network formed by all seven *statements*. Typically, the analyst must read some or all of the *statements* in a *map* in order to extract the meaning of the text as a whole. In this regard, it is then notable that the seven *concepts* are implicated in varying numbers of *statements*. Specifically, the *concepts* labeled "organization" and "process" are found in three *statements* while all other *concepts* are found in either two or one. Differences of this kind have important implications for the extraction of meaning from a text network, especially in larger networks. However, as the number of *concepts* and *statements* increases, so does the complexity of that task (Popping & Roberts, 1997).

The remainder of this paper is organized as follows. In the next section, I review the relevant literature on network text analysis with an eye toward those methods that rely most heavily on network analytical methods for the extraction of meaning. In the third section, I describe a novel method for the creation and analysis of text networks. In the fourth section, I apply the method to a sample of seven Academy Award-winning screenplays—*Little Miss Sunshine*, *Juno*, *Milk*, *The Hurt Locker*, *The King's Speech*, *Midnight in* Paris, and *Django Unchained*.

2. Literature Review

As noted above, network text analysis (NTA) involves the conversion of words into concepts, the creation of linkages between pairs of those concepts, and the subsequent analysis of the resulting text network. In her doctoral thesis, Diesner (2012: pp. 90-91) compares 17 groups of methods for constructing networks of words along several dimensions, the most relevant four of which are—1) *automation*, i.e. "the ability to automatically collect one-mode and multi-mode network data", 2) *abstraction*, i.e. whether and how terms are abstracted to "concepts or higher level aggregates", 3) *generalization*, i.e. "the ability to identify new and unseen instances of entity classes and entity attributes", and 4) "*steps needed to reason about meaning of network data*". Among the approaches she reviewed are several that are similar to the method described later in this paper. Specifically, these include *hypertext* (Trigg & Weiser, 1986); *semantic webs* (Berners-Lee, 2001; van Atteveldt, 2008); *concept maps* (Novak & Gowin, 1984), *centering resonance analysis* (Corman et al., 2002; McPhee, 2002), *mental models*

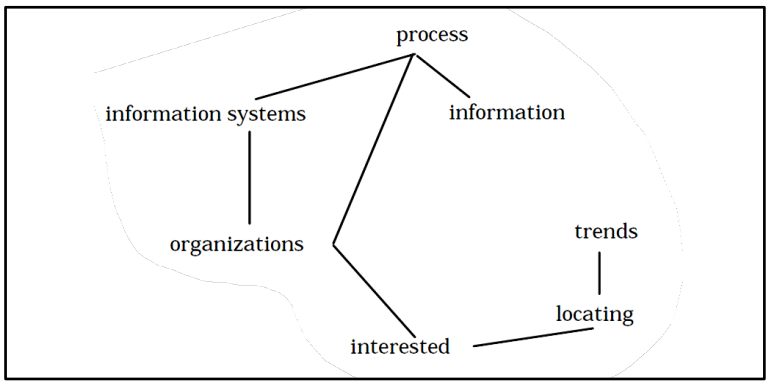


Figure 1. A simple text network (adapted from Carley, 1997).

(Collins & Loftus, 1975), *mind maps* (Buzan, 1974) *semantic grammars* (Roberts, 1997), *network text analysis in the social sciences* (Carley & Palmquist, 1992), *semantic networks in communication sciences* (Danowski, 1993), and *event coding in political science* (King & Lowe, 2003; Schrodt et al., 2008).

In **Table 1** below, a subset of the methods reviewed by Diesner is summarized along four dimensions that are relevant to this study—1) *selection*, i.e. the basis upon which words are selected for or excluded from the subsequent analysis, 2) *conceptualization*, i.e. to which conceptual categories are the selected words mapped, 3) *relationships*, i.e. the basis upon which pairs of concepts are linked to one another and 4) the approach taken to the *extraction of meaning*. Four groups of methods are summarized therein—Centering Resonance Analysis (CRA), Network Text Analysis in the social sciences, Semantic Networks, and Semantic Mapping. In the second column of the table are described the various approaches to selection employed in the studies under examination. Three distinctions are worthy of note. The first of these concerns parts of speech. CRA, for example, relies exclusively upon nouns and noun phraseswhich they define as "a noun plus zero or more additional nouns and/or adjectives" (Corman et al., 2002: p. 174). All methods exclude determiners (a, an, the), as well as some or all of the following—prepositions, conjunctions, transitive verbs, verbs of being, acronyms, place names and proper nouns, and other words that "distort the description of the text" (Doerfel & Barnett, 1999: p. 592). As for pronouns, some approaches convert them to their corresponding nouns or noun phrases (e.g. Carley, 1997) while others simply eliminate them from consideration altogether (Doerfel & Barnett, 1999). A second distinction concerning unit of analysis is the importance of word frequency as a basis for inclusion. Notably, all three of the "semantic" methods select words based on their frequency of occurrence, excluding of course the "stop words" just described.

The third column in the table indicates whether and how the resulting selected words and phrases are aggregated into higher-order conceptual categories. Two methods that do not aggregate are CRA (Corman et al., 2002) and semantic mapping (Leydesdorff & Welbers, 2011). The former applies "minimal affix stemming" which changes only plural forms of words to the singular, e.g. *drivers* → *driver*. The semantic mapping approach does not employ any kind of stemming and as such, its word networks include multiple related variants of key words and concepts, e.g. both *chemical* and *chemistry* or all of the following—*science*, *scientific*, *scientists*, and the abbreviation *Sci*. The other methods aggregate to broader categories of concepts. Carley (1997) aggregated words through the use of a conceptual thesaurus based upon two generalizations—one contextual and the other grammatical. In the former condition, concepts with similar meanings were recoded into the same concept, e.g. *information system* and *system* both were coded as *information system*. In the second instance, concepts with different tenses and endings were stemmed, e.g. *retrieved*, *retrieves*, *retrieving*, etc. all become *retrieve*. The semantic network studies appear to use only the second approach.

Table 1. Selected methods for generating and analyzing text networks.

Method/Approach	Selection	Conceptualization	Relationship	Extraction of Meaning
Centering resonance analysis (Corman, Dooley et al., 2002; Dooley, Corman et al., 2003; Kuhn & Corman, 2003)	Nouns and noun phrases	Yes: minimal affix stemming from plural to singular forms.	Links are created between sequentially-occurring and centered noun phrases in an utterance…	Qualitative (reading statements) Quantitative (correlation; resonance)
NTA in social sciences (Carley & Palmquist, 1992; Carley, 1997)	Frequently occurring "substantive" words and short phrases; pronouns converted to nouns	Yes: conceptual thesaurus for frequent substantive concepts	Causal, definitional, directional, and equivalence relations	Qualitative (reading statements) Quantitative (number of concepts and statements)
Semantic networks (Danowski, 1993; Doerfel & Barnett, 1999; Rice & Danowski, 1993)	Most frequently occurring words, minus "stop words" such as articles, prepositions, pronouns conjunctions, transitive verbs, acronyms, verbs of being, place names and other words that "distort"	Yes: stemming of plural forms	Co-occurrence of concepts within a researcher-defined window	Qualitative (reading statements), Quantitative (number of concepts and statements)
Semantic mapping (Leydesdorff & Welbers, 2011)	Words occurring more than twice, minus stop words, in the sample	No: uses terms verbatim.	Co-occurrence of words within a set of 195 documents	Qualitative (reading statements), Quantitative (correlation & factor analysis)

The fourth column of the table describes the nature of the relationship that links concepts. The most common of the approaches is co-occurrence of two or more concepts within some researcher-defined window. Typical of this approach is Carley (1997) who analyzed mental maps derived from answers to two open-ended questions by 41 undergraduate students enrolled in an information systems course. The questions were "What is an information system?" and "What leads to information systems success and failure?" After eliminating stop-words, maps were extracted using five windows—four, six, eight, ten, and 100 words. A linkage was established between pairs of words occurring within those windows. Leydesdorff & Welbers (2011) used a much larger window—their entire sample of the 195 documents—to establish co-occurrence and thus relationship. Neither Carley & Palmquist (1992) nor Corman et al. (2002) relied upon co-occurrence within windows of any kind. The former mapped relationships between concepts that were primarily causal or definitional in nature and had such attributes as strength, sign, direction, and meaning. For example, in response to questions such as "What is the purpose of research writing?" or "What steps would you follow in writing a research paper?" an answer such as "decide specifically what you are going to write-down" was coded as a relationship between the concepts "topic" and "writing". CRA, on the other hand, focused on essentially hierarchical relationships derived from the sequentially linking "noun phrases, which are potential centers in an utterance" (Corman et al., 2002).

As shown in the fifth column of **Table 1**, what all these methods have in common is their reliance upon a network analytic approach to extracting meaning from the textual data. More specifically, each uses a combination of qualitative and quantitative techniques to accomplish that end. In the former instance this involves, at a minimum, reading some or all of the statements and perhaps relating them back to parts of the text. For example, Corman et al. (2002) found that "chains of concepts" in their network graph corresponded directly to "specific chains of nouns in specific sentences" in the text (p. 199) and that different parts of the graph captured the "discursive division" between the two parties whose utterances comprised the map (p. 186). In contrast, quantitative analysis of the map usually involves counting the number of unique concepts and statements, especially when comparing maps associated with different individuals/groups (Carley, 1997). Enumeration is also key in determining the most influential concept in a network (Leydesdorff & Welbers, 2011). Other quantitative methods for discerning meaning within or across text networks include correlational (Corman et al., 2002) and factor analysis (Leydesdorff & Welbers, 2011). It is worth noting that CRA stands out among these approaches as the only method that developed its own metric—resonance—for measuring the "mutual relevance of two texts" (Corman et al., 2002: p. 189).

Although the sample is relatively small, it is possible to generalize broadly about these approaches to NTA. Firstly, all methods, select some words for further analysis while excluding others from further consideration. Articles, prepositions, and pronouns are among the words most often excluded. The remaining words then are aggregated to higher order conceptual categories, typically grammatical or logical in nature. Next, pairs of concepts are then related to one another, most frequently on the basis of their co-occurrence within some window of words. Finally, an analysis of the structure or pattern of these relationships—through network analysis and/or statistical methods like factor analysis—are employed to extract meaning.

Taken together, the four dimensions outlined above—selection, conceptualization, relation, and extraction of meaning—establish criteria upon which any claim to novelty in the NTA field must be understood. Considered in turn they are as follows:

- ***Selection*:** A new method might be novel because it includes or excludes different (kinds of) words and/or uses different rationales for doing so.
- ***Conceptualization*:** A new method might be novel if it employs a different approach to aggregating from words to concepts.
- ***Relation*:** A new method might novel for the kind of relationship it uses to link concepts.
- ***Extraction*:** Finally, a method might be novel for the techniques and/or metrics it uses in the extraction of meaning from the network data.

The next section of this paper details a new method of network text analysis that meets these four conditions.

3. A Novel Method

As suggested in the introduction, at the heart of the method described herein is etymology or word history, a term defined in the *Cambridge Advanced Learner's Dictionary* as "the study of the origin and history of words". Over the past several decades, academic etymologists have traced thousands of words in a wide variety of Eu-

ropean languages back to their Indo-European and other etymological roots. Among the most well-known contemporary sources in this field are the *American Heritage Dictionary of Indo-European Roots* (Watkins, 2011); the *Leiden Indo-European Etymological Dictionary* series (Lubtosky, 2006-2013) which is a compilation of etymological databases using Sergei Starostin's STARLING software technology; and the *Indogermanisches Etymologisches Woerterbuch* (Pokorny, 1959) which is an "updated and slimmed-down reworking" of Walde & Pokorny's (1927-32) three-volume *Vergleichendes Wörterbuch der Indogermanischen Sprachen* (Wikipedia, 2013).

Several aspects of contemporary research on etymology are potentially useful for network text analysis. First, there is the obvious fact of the network-like relationship that underlies the field itself. By definition, from every etymological root descends or originates at least one word, otherwise it is not a root. That relationship is genitive, i.e. a relational case typically expressing source, possession, or partition. It is hierarchical and directed—from the root (parent) to word (descendant). The name used to relationship of pair of words descending from the same root is "paronym". **Figure 2**, below depicts the Indo-European (IE) root **saewel—**which means "sun" and five words descending from it as indicated in the *American Heritage Dictionary of Indo-European Roots* (AHDIER) —*south*, *solar*, *solstice*, *helium*, and *sun*.

What is significant about **Figure 2** is that it represents a bridge between the second and third stages of the NTA process—*Conceptualization* and *Relation*. More specifically, while *conceptualization* involves assigning words to their etymological roots, that relationship is not just categorical: there is also a network-like relationship—as shown above. But that relationship is actually not sufficient to construct the kind of network map like other methods of NTA. The issue is that the text networks are built on relationships between concepts, not between words and concepts. The following example will explain how this hierarchical and directed relationship between words and higher-order concepts can be converted into one that links concepts.

Consider a text that contains the following nine words: the closed compounds *sunlamp*, *manpower*, *sunlight*, *lamplight*, and *gentleman*; the open compounds *solar power* and *native tongue*; the hyphenated compound *self-possessed*; and the proper noun *Secretary General*. As shown in **Table 2**, below, these words are all multi-morphemic compounds, each element of which descends from a different etymological root. And in that fact lies the solution to the linking of concepts. Recall that a *statement* in NTA is comprised of two concepts and the relationship that links them. In **Figure 3**, below, each of these words appears on the link between the two etymological roots—the concepts—that co-occur within the word. Put another way, the relationship is the co-occurrence of two different etymological roots in the same multi-morphemic compound or multi-word expression—co-occurrence in what is essentially a window of one word. For example, the etymological roots **gene—**(to give birth; beget) and **man-1** (man) are linked by their co-occurrence in the compound word *gentleman*. Taken together, that word and those two roots comprise a statement. And as shown below, it is possible to construct an entire map or network from these interconnected statements. That network is comprised of eight concepts—namely, the Indo-European roots ***dnghu***, ***gene***, ***man*1**, ***s*(*w*)*e***, ***poti***, ***sawel***, ***leuk***, and ***lap***—and nine statements which include the words *native tongue*, *gentleman*, *Secretary General*, *self-possessed*, *manpower*, *solar power*, *sunlight*, *sunlamp*, and *lamplight*, as well as the pair of roots to which each element belongs.

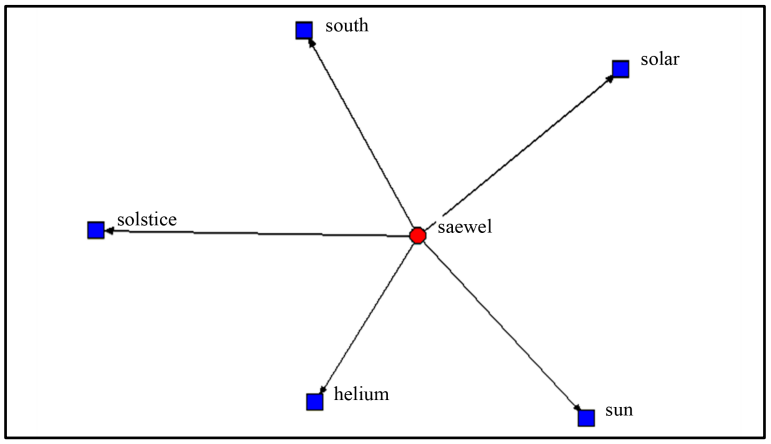


Figure 2. An etymological root and its derivatives.

Table 2. Selected Indo-European roots and their derivatives.

Roots (definition)	Paronyms
lap (to light, burn)	lamp, lantern, and eclampsia
dnghu-(tongue)	tongue, language, lingo, linguine, and linguist
leuk (light, brightness)	light, lightning, lux, lucent, Lucifer, illumination, lunar, luster, illustrate, lucid, and lynx
mei-2	meiosis, menu, mince, minute, minor, minus, diminish, minimum, minister, ministry, and administration
man-1	man, manikin, mannequin, Manu, and mensch
poti (powerful; lord)	possess, power, possible, potent, and pasha
saewel-(the sun)	sun, south, southern, solar, and solstice
gene (to give birth, beget)	gender, general, generate, generic, genre, gene, genocide, genius, engine, genuine, gonad, kin, kind, gentle, progeny, pregnant, nascent, natal, nation, native, innate, puny, and renaissance.
s(w)e-(self)	self, gossip, suicide, swami, secede, seclude, secret, secure, seduce, segregate, select, separate, sever, sure, sober, sole, solo, solitary, idiom, idiot, and Sinn Fein.

Note: **Source:** American heritage dictionary of Indo-European roots.

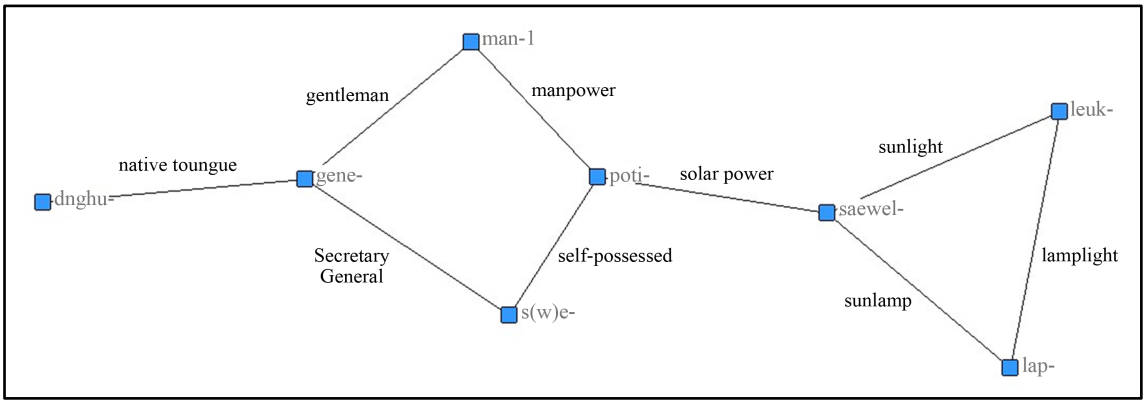


Figure 3. A text network based on etymological relationships among selected words contained in **Table 2**.

Once we recognize that networks can be constructed in this manner, it becomes apparent that compounds words are not the only ones that can or should be included. For example, one can include abbreviations and acronyms, e.g. *USA* (United States of America), *FBI* (Federal Bureau of Investigation) and *CNN* (Cable News Network). Also relevant are portmanteau or blend words which are formed by joining the beginning of one word and the end of another. Examples of blend words include *brunch* which is a blend of **br**eakfast and l**unch** and *motel* which is a blend of **mo**tor and ho**tel**. A third category would include certain clipped words, which are defined in the *Random House Dictionary* as ones "formed by dropping one or more syllables from a longer ***word*** or phrase with no change in ***meaning***". Examples of relevant clipped words include *internet* which is a clipped form of *internetwork* and *slo-mo* which is a clipped form of slow-motion.

And though it may seem otherwise, these groups of words are no random assortment. Rather, they comprise a well-defined, inter-related set that is extensively-studied in the field of morphology. Specifically, they all belong to the branch of morphology known as word-formation, the study of creation of new words principally through changes in their form (Wisniewski, 2007). Morphologists described several kinds of word-formation (Stekauer, 2000; Plag, 2003, Lieber & Stekauer, 2009), all of which are hierarchically related to one another as shown in **Figure 3**, below (Plag, 2003: p. 17).

As the diagram suggests, the two major branches of morphology are *word-formation* ("the ways in which new complex words are built on the basis of other words of morphemes") (Plag, 2003: p. 13) and *inflection* (the modification of words to express a broad range of grammatical categories, e.g. tense, person, number, etc.) *Word-formation* can be further sub-divided along two lines—*derivation* (forming new words through changes in form)

and *compounding* (the creation of novel words through the combination of existing ones). *Derivation* is also of two varieties—*affixation* (changes in form through the addition of affixes) and *non-affixation* (changes in form through means other than affixes). The former involves the addition of *prefixes*, *suffixes*, and *infixes* while the latter includes the creation of *abbreviations*, *acronyms*, *blend* words, truncated or *clipped* words, as well as the process *conversion*. **Table 3**, below, provides examples of each type of word-formation described in **Figure 4**. Word formation is, then, the basis for the *selection* of words in this method of NTA.

To summarize briefly, the method described above 1) *filters* words based on their morphological properties—only multi-morphemic compounds are analyzed, 2) maps them to higher-order *concept*s on the basis of etymology—elements of each compound are mapped back to their respective etymological root, and 3) *relates* those concepts on the basis of their co-occurrence within the same word. As far as we are aware, no other method of NTA performs these three basic steps in this way. However, there remains to be considered the question of the extraction of meaning from the overall network. In short, this method achieves that goal through an approach not unlike other methods—through reading and through enumeration of concepts and statements. There are, however, important differences—differences detailed in the following section.

4. Methods and Data

In this Section I demonstrate the application of the method described in the previous section to a selection of Academy Award-winning screenplays, namely the seven that received the Oscar for Best Original Screenplay between 2006 and 2012. As shown in **Table 4**, below, the seven winners of the Academy Award for Best Original Screenplay in the years 2006-2012 were *Little Miss Sunshine* (Arndt, 2006), *Juno* (Busey-Mario, 2007), *Milk* (Black, 2008), *The Hurt Locker* (Boal, 2007), *The King's Speech* (Seidler, 2009), *Midnight in Paris* (Allen, 2011), and *Django Unchained* (Tarantino, 2012). Electronic copies of these seven screenplays were found in

Table 3. Examples of seven types of "novel" word-forms in the sample.

Type	Examples
Word Formation > Compounding > Closed Compounds	Armpit, briefcase, cowboy, deadline, earring, forehead, goldfish, handcuffs, inmate, jailhouse, voicemail, wallpaper, yearbook, zookeeper
Word Formation > Compounding > Copulative Compounds	Attorney-client, actor/model
Word Formation > Compounding > Hyphenated Compounds	African-American, open-minded, panic-stricken, rear-view, tree-lined, under-aged, voice-mail, wide-eyed
Word Formation > Compounding > Multi-Word Compounds	Over-the-top, jack-in-the-box, sister-in-law, ten-year-old, all-you-can-eat, head-to-head, back-and-forth, ball-and-chain
Word Formation > Compounding > Open Compounds	Anxiety attack, bachelor pad, candy bar, end zone, fast food, nose dive, olive branch, palm tree, quarter mile, rag doll, school bus
Word Formation > Derivation > Affixation > Prefix	Understand, overdrive, overhand, underhanded, overstate, under-appreciated
Word Formation > Derivation > Affixation > Suffix	Awesome, handsome, bothersome, hardware, software, clockwise
Word Formation > Derivation > Affixation > Infix	Unbloodybelievable (bloody + unbelievable), fanbloomingtastic (blooming + fantastic)
Word Formation > Derivation > Non-Affixation > Abbreviations, Acronyms, and Initialisms, Anacronyms	AOL (America Online), DMV (Dept. of Motor Vehicles), mph (miles per hour), MTV (music television), VCR (video cassette recorder), Yuppie (Young Urban Professional), radar (**ra**dio **d**etecting **a**nd **r**anging)
Word Formation > Derivation > Non-Affixation > Blend Words	Medevac (medical + evacuation), motel (motor hotel), guesstimate (guess + estimate), camcorder (camera + recorder), helipad (helicopter + pad)
Word Formation > Derivation > Non-Affixation > Clipped Words	Internet (internetwork), hi-fi (high-fidelity), email (electronic mail), slo-mo (slow-motion), vid-com (video + communication)
Word Formation > Derivation > Non-Affixation > Conversion	*Eyeball* (as noun or as verb)

Table 4. Seven academy-award winning, best original screenplays, 2006-2012.

Title (Year)	Author	Total Nominations	Length (Pages)	Genres
Little Miss Sunshine (2006)	Michael Arndt	0	110	Drama/Comedy
Juno (2007)	Diablo Cody	0	101	Drama/Comedy
Milk (2008)	Dustin Black	0	104	Drama
The Hurt Locker (2009)	Mark Boal	2	118	Drama
The King's Speech (2010)	David Seidler	0	92	Drama
Midnight in Paris (2011)	Woody Allen	15	89	Comedy
Django Unchained (2012)	Quentin Tarantino	3	167	Drama

Note: Source: Oscar.com, IMDb.com.

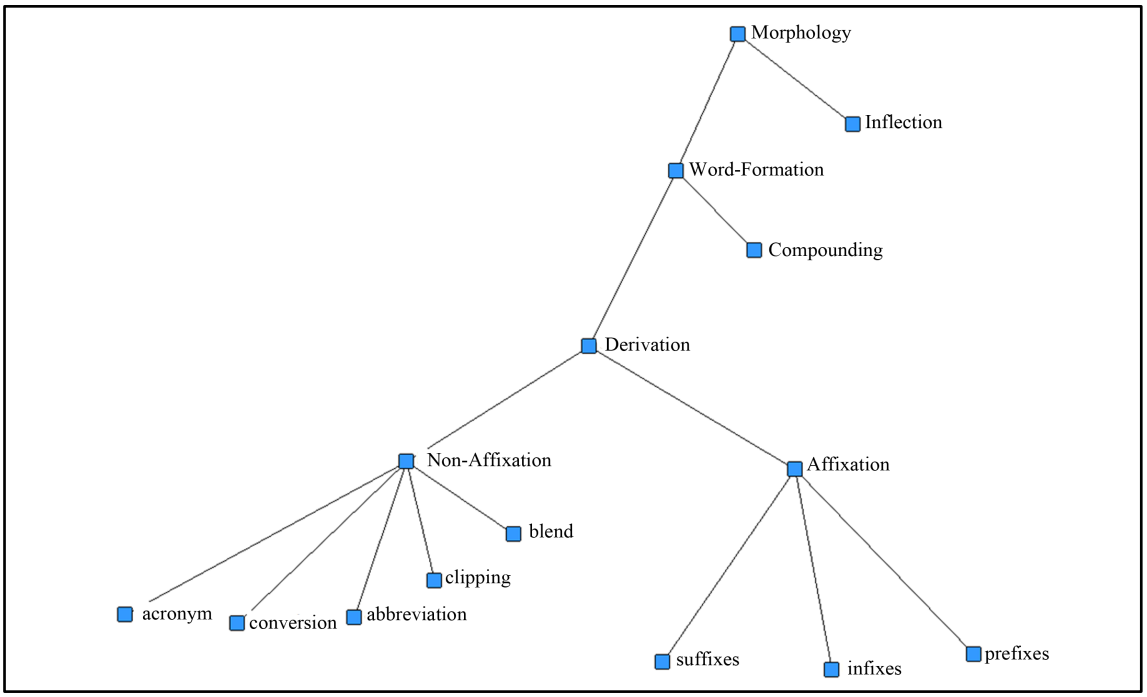


Figure 4. Tree diagram of morphology and its domains.

numerous places online and where more than one version existed, the latest publication date was selected. All but one of the screenplays was machine readable—Woody Allen's *Midnight in Paris*. It was also the only of the seven to be categorized by the International Movie Database (IMDb, 2013) as comedy and not a drama or drama-comedy. Notably, all seven screenplays were solo-authored and only one author is a woman—Brook Busey-Mario, the author of *Juno*. Three of the authors were nominated for the Academy Award for Best Original Screenplay either in years before or during the time periods under study. They are Mark Boal, who was also nominated for *Zero Dark Thirty* in 2012; Quentin Tarantino, who was also nominated for *The Inglorious Basterds* in 2009 and *Pulp Fiction* in 1994; and Woody Allen who has received a total of 15 nominations. Finally, note that the page range varied substantially around the 111 page average with the longest *Django Unchained* (167 pages) being almost twice the length as *Midnight in Paris* (89 pages) and *The King's Speech* (92 pages).

The first step in the text analysis was to generate a list of all relevant words and phrases for each screenplay. This was accomplished with the help of a software program called Automap 3.0.10 (Carley, 2001-2013). The process was as follows. First the screenplay was converted to a text file and uploaded to Automap. After removing single letters, extra spaces, and spurious characters, four routines were run within Automap—*Extract Hy-*

phenated Words, *Suggest N-Grams*, *Identify Possible Acronyms* and *Concept List*. As their names suggest, the first routine extracted all words containing hyphens, many of which turned out to be hyphenated compounds. The second routine selected from the screenplay all multi-word phrases and combinations that were not named entities but which could be open compounds, e.g. *post office* or *parking lot*. The third routine identified and extracted all words that were capitalized. Several of these turned out to acronyms or abbreviations, e.g. CD (compact disk) and HUMVEE. The fourth routine used was *Concept List* (*Per Text*) which generated a list of all unique words in the text. Because Allen's *Midnight in Paris* was not machine-readable, I read and coded the screenplay manually.

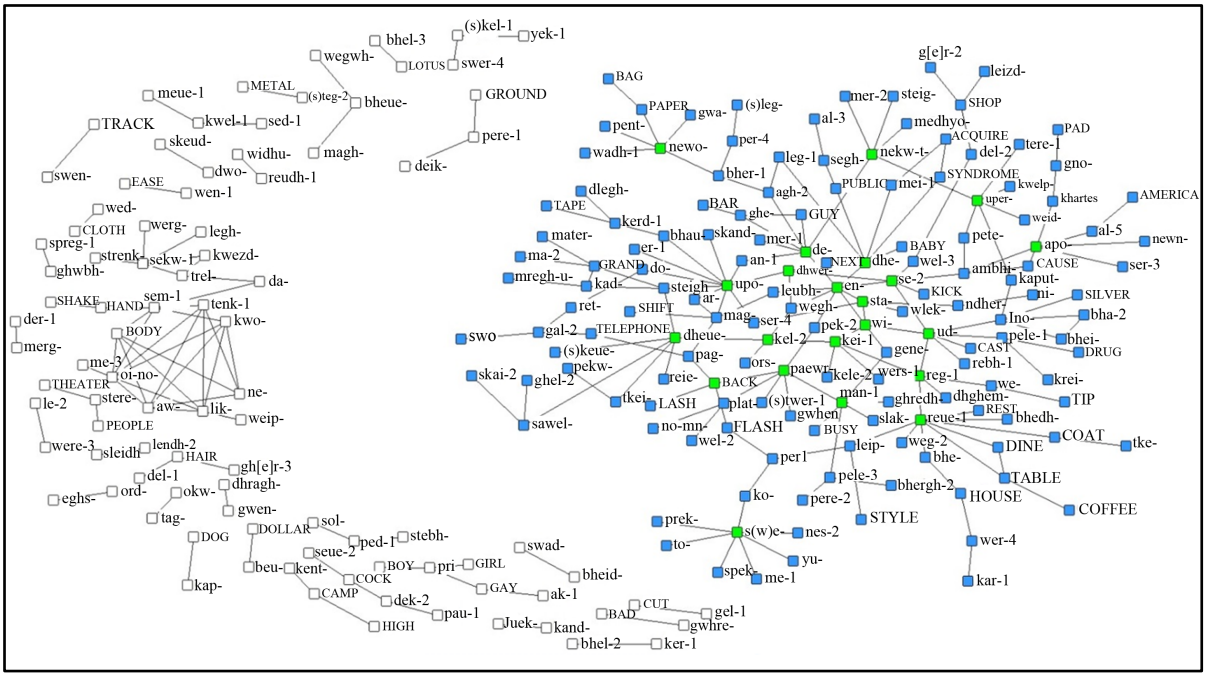
The output of each of the four routines was consolidated into a single list for each screenplay and then scanned by the author to identify all multi-morphemic compounds—described in **Table 3**—acronyms, abbreviations, anacronyms, blend words, clipped words, compounds, and pseudo compounds. Excluded from consideration were all proper nouns (Green Zone, Hollywood), place and organization names (South Pole, Scotland Yard, Burger King), product names (Land Rover), holidays (New Year's Eve, Christmas Eve), as well as any other word or phrase connoting a specific person, place, or thing through capitalization. Also eliminated were all references to screenplay and film jargon, e.g. ECU (extreme close-up), off-screen, VO (voice-over) and POV (point of view). In order for suggested n-grams to be included in the word list, the word combination had to appear in one of seven on-line dictionaries—the *American Heritage Dictionary of the English Language*, the *Cambridge Advanced Learner's Dictionary*, the *Cambridge Dictionary of American English*, *Collins English Dictionary*, the *Compact Oxford English Dictionary*, the *MacMillan Dictionary*, or *Merriam Webster's Online Dictionary*. Finally, multi-word exclamations and interjections such as *good night*, *goodbye*, *OMG* were also eliminated.

The components of each of the remaining multi-morphemic compounds were then mapped to their corresponding etymological roots, over 95% of which were roots found in the *American Heritage Dictionary of Indo-European Roots* (AHDIER). Next, all pairs of roots in a given screenplay were converted into a symmetrical matrix which was then uploaded into version 6.487 of the UCINet software program (Borgatti, Everett, & Freeman, 2002). Text networks were then generated using version 2.118 of the Net Draw software program embedded in UCINet. As shown in **Table 5** below, there was considerable variation across the seven screenplays in terms of the number of unique concepts (etymological roots) and the number of statements contained in their text networks. In **Figure 5** is depicted the entire text network for Dustin Black's *Milk*. It was selected because among the seven screenplays in this sample, its text network is closest to the average in terms of the number of concepts (248 vs an average of 251), number of statements (252 vs and average 258), the number of concepts in the network's main component (154 vs an average of 142), and the number of statements in the main component (176 vs an average of 176).

As noted previously, the extraction of meaning from text data is one of the primary concerns of NTA. As also mentioned, there is not a consensus about how this is to be done. One thing that all reviewed methods have in common, however, is reading statements, particularly those associated with the most central or influential nodes. In network analysis there are dozens of measures of centrality and influence, the majority of which are based either on the number of connections a node has and/or the position and related role it occupies within the network.

Table 5. Structural characteristics of seven academy award winning, best original screenplays.

Title (Release Year)	Pages	Number of Concepts	Number of Statements	Nodes in Main Component	Statements in Main Component
Little Miss Sunshine (2006)	110	307	282	165	191
Juno (2007)	101	296	301	188	235
Milk (2008)	104	248	252	154	176
The Hurt Locker (2009)	118	273	276	148	191
The King's Speech (2010)	92	209	194	118	140
Midnight in Paris (2011)	89	124	194	14	46
Django Unchained (2012)	167	299	305	209	252
Average	111	251	258	142	176



Nodes colored in green represent concepts—etymological roots—whose constraint values are less than or equal to 0.25. Blue nodes have constraint values greater than 0.25. The unfilled or white nodes also have constraint scores above 0.25 and are notable for not being mutually reachable in relation to the blue and green nodes (the network's main component), as well as in relation to most other white node.

Figure 5. Text network map of Dustin Lance's *Milk*.

The method adopted here is to rely on the position/role, namely a measure of structural holes known as constraint. In short, nodes that have low constraint in a network are those that bridge or connect its otherwise disconnected parts. In social network analysis this role is known as brokerage which, in management and organizational studies has been linked to superior firm-level, business-unit, managerial, and individual performance (Burt, 2005; Cross & Cummings, 2004; Hunter, 2011; Hunter & Chinta, 2013).

Table 6, below, shows the statements associated with the least constrained nodes in the text network for the screenplay of *Milk*. As calculated in UCINet, constraint values range from 0 (unconstrained) to 1.0 (completely constrained). In this study "least constrained" was defined as values less than or equal to 0.25. There were 22 nodes or concepts with values in this range, just under 9% of the 248 concepts in screenplay's entire text network. Notably, all of the most influential nodes are in the network's main component, i.e. the largest grouping of nodes that are mutually reachable. Also noteworthy is that 21 of the 22 most central nodes are concepts traced to Indo-European roots found in the AHDIER.

The synopses provided by IMDb for the film *Milk* reads as follows: "The story of Harvey Milk, and his struggles as an American gay activist who fought for gay rights and became California's first openly gay elected official." Wikipedia describes *Milk* as "a 2008 American biographical film based on the life of gay rights activist and politician Harvey Milk, who was the first openly gay person to be elected to public office in California, as a member of the San Francisco Board of Supervisors". The site's summary of the film's plot provides information on the difficult relationship that existed between Milk and a fellow supervisor, Dan White, who eventually shot and killed Milk in 1978:

> After two unsuccessful political campaigns in 1973 and 1975 to become a city supervisor and a third in 1976 for the California State Assembly, Milk finally wins a seat on the San Francisco Board of Supervisors in 1977 for District 5. His victory makes him the first openly gay man to be voted into major public office in California and in the top three in the entire US. Milk subsequently meets fellow Supervisor Dan White, a Vietnam veteran and former police officer and firefighter. White, who is politically and socially conservative, has a difficult relationship with Milk, and develops a growing resentment for Milk when he opposes projects that White proposes.

Table 6. Statements associated with the 22 least constrained nodes in the text network of the screenplay for Dustin Black's *Milk*.

Node	Constraint	Statements
apo-	0.217	after all, afternoon, postcard, sort-of, because
BACK	0.250	backfire, backlash, payback
de-	0.184	today, tomorrow, tonight, go-to guy, into, onto
dhe-	0.167	AIDS, baby-faced, elected official, indeed, public office
dheue-	0.167	downstairs, hometown, rundown, showdown, sundown, town hall
dhwer-	0.250	doorway, next door, open door, side door
en-	0.147	side entrance, inside, entryway, indeed, infighting, inside, instead, into, love-in
kei-1	0.200	city council, city hall, city-wide, civil rights, civil war
kel-2	0.250	hallway, town hall, asshole, city hall
man-1	0.167	manslaughter, policeman, businessman, congressman, fireman, gentleman
nekw-t-	0.200	tonight, midnight, nightmare, nightsticks, overnight
newo-	0.200	newborn, newcomer, newfound, newlywed, newspaper
paewr-	0.167	backfire, firefighters, fireman, fireplace, firestorm
reg-1	0.200	human rights, outright, right wing, civil rights, dressing room
reue-1	0.136	rest room, waiting room, bathroom, bedroom, coatroom, dining room, dressing room, living room
s(w)e-	0.185	self-respect, self-deprecating, themselves, yourself, herself, himself, itself, myself, yourselves
se-2	0.184	outside, side door, side entrance, sidekick, sidewalk, besides, inside
sta-	0.227	statewide, storefront, understand, instead, liquor store
ud-	0.143	without, full out, outcast, outline, outrage, outright, outside
uper-	0.167	overhead, overnight, overpass, overturning, overwhelmed, supervisor
upo-	0.132	upbeat, updates, upfront, upon, uprising, upscale, upstairs, make-up artist, open door
wi-	0.227	within, without, citywide, nationwide, statewide

A comparison of the keywords in Table 6 with the information provided by the plot excerpt and synopses reveals large areas of overlap. Specifically, we have descending from the IE root **dhe** (to set, put) are the words *elected official* and *public office*. From **dheue** (to close, finish, come full circle) descend the words *showdown* and *town hall*. From the root **kei-1** (to lie; bed, couch; beloved, dear) descend *city council*, *city hall*, *city-wide*, *civil rights*, and *civil war*. The root **kel-2** (to cover, conceal, save) gives rise to *city hall* and *town hall*. From the root **man-1** (man) we have several words related to Dan White. Specifically there are *policeman* and *fireman* (White's professions prior to entering politics), as well as *manslaughter* (the crime of which White was convicted for shooting Milk). From the root **reg-1** (to move in a straight line; to direct in straight line, lead, rule) descend the words *human rights*, *right wing*, and *civil rights*. Finally, descending from the root **wi** (apart, in half) descend the words *citywide*, *statewide*, and *nationwide*.

Taken together these fourteen words—or better yet statements because they represent the relationship existing between two concepts/roots—neatly encapsulate the story. Notably, several of them appear in the film synopses provided by IMDb and Wikipedia. Specifically, the single-sentence synopsis from IMDb contains the phrases "gay rights" and "elected official" while Wikipedia's contains "gay rights" and "public office". Notably the phrase "gay rights" does not appear in Table 6. Even though it appears several times in the screenplay, the two-word combination was not found in any of the seven online dictionaries. If the list of sources had included Wikipedia or Wiktionary or Dictionary.com, it would have been included and, of course, been part of a very central statement.

The same approach was applied to the six other screenplays in the sample and largely similar results were obtained, as summarized in **Table 7**, below. Specifically, in four of the other six screenplays the statements associated with the least constrained nodes clearly overlap with the synopses provided in the right hand column. A typical example is *The Hurt Locker*, a film about the exploits of US Army bomb squad deployed to Iraq. Several of the statements associated with the least constrained nodes evoke war and explosives. These include *army-issue*, *body armor*, *fireball*, *first aide*, *gunfire*, *gunshot*, *half-destroyed*, *headquarters*, *HUMVEE*, *IED* (Improvised Explosive Device), *machine gun*, *shell-shocked*, *state-of-the-art*, *suicide bomber*, and *USA*, as well as a robot called *TALON* (Threat and Local Observation Notice).

Table 7. Selected statements associated with the least constrained nodes in six academy award-winning, best original screenplays.

Title and Synopses	Selected Statements Associated with the Least Constrained Nodes
The Hurt Locker: "During the Iraq War, a Sergeant recently assigned to an army bomb squad is put at odds with his squad mates due to his maverick way of handling his work" (Wikipedia, 2013). "Forced to play a dangerous game of cat-and-mouse in the chaos of war, an elite Army bomb squad unit must come together in a city where everyone is a potential enemy and every object could be a deadly bomb" (IMDb, 2013).	army-issue, body armor, bulls-eye, fireball, fire-extinguisher, first aide, gunfire, gunshot, half-destroyed, headquarters, headset, HUMVEE, IED (Improvised Explosive Device), machine gun, shell-shocked, standard-issue, state-of-the-art, suicide bomber, TALON (Threat and Local Observation Notice), UN (United Nations), and USA.
Juno: "The title character (is) an independent-minded teenager confronting an unplanned pregnancy and the subsequent events that put pressures of adult life onto her" (Wikipedia, 2013). "Faced with an unplanned pregnancy, an offbeat young woman makes an unusual decision regarding her unborn child" (IMDb, 2013).	backpack, band-aid, bedroom, bedside, birthday, birthing room, blackboard, burnout, dime bag, dropouts, drugstore, handsome, high school, high-fives, knocked-up, make-out, runaways, sweat bands, teenage, teenwave, time-suck, waistband, waiting room, weight loss.
Little Miss Sunshine: "A family determined to get their young daughter into the finals of a beauty pageant take a cross-country trip in their VW bus" (IMDb, 2013). "When a pudgy, bespectacled seven-year-old, Olive voices her desire to take home the coveted Little Miss Sunshine crown at an upcoming beauty pageant, her wildly dysfunctional family sets out on an interstate road trip to ensure her a clear shot at realizing her dreams" (RT, 2012).	back seat, backstage, butt-wagging, convenience store, dressing room, gas station, get-up, half-dancing, hatchback, highway, make-up, newsstand, on-ramp, onstage, outfit, over-coiffed, over-dressed, runners-up, runway, seat belt, service station, side door, stagehand, SUV (Sport Utility Vehicle), turn-off, under-revving, upbeat, US (United States)
The King's Speech: "The story of King George VI of the United Kingdom of Great Britain and Northern Ireland, his impromptu ascension to the throne and the speech therapist who helped the unsure monarch become worthy of it" (IMDb, 2013). "After the death of his father King George V and the scandalous abdication of King Edward VIII, Bertie who has suffered from a debilitating speech impediment all his life, is suddenly crowned King George VI of England. With his country on the brink of war and in desperate need of a leader, his wife, Elizabeth, the future Queen Mother, arranges for her husband to see an eccentric speech therapist, Lionel Logue" (RT, 2012).	ballroom, common man, control room, prime minister, state room, teatime, wartime, WWI (World War I), pub (public house), onstage, goose-step, naval officer, poker-faced, public affairs, tea time,
Django Unchained: "With the help of a German bounty hunter, a freed slave sets out to rescue his wife from a brutal Mississippi plantation owner" (IMDb, 2013). "Set in the South two years before the Civil War, Django Unchained (is)⋯ a slave whose brutal history with his former owners lands him face-to-face with German-born bounty hunter Dr. King Schultz. Schultz is on the trail of the murderous Brittle brothers, and only Django can lead him to his bounty. Honing vital hunting skills, Django remains focused on one goal: finding and rescuing Broomhilda, the wife he lost to the slave trade" (RT, 2012).	back porch, backside, bare ass, bare chest, bareback, barefoot, barn yard, Big-House, blacksmith, Boss Man, bowler hat, bullet hole, bullwhip, bunkhouse, campfire, country road, countryside, courthouse, dirt road, family unit, farmhouse, firewood, fishing pole, free-man, graveyard, gunbelt, gunfight, gunman, gunpowder, hangman, hill top, hillbilly, hillside, horseback, inbreed, incoming, lawman, lifetime, livestock, lumber yard, marksmanship, menfolk, mountain man, outlaw, outhouse, overseer, peace officer, peacemaker, pocket knife, runaway, shotgun, showdown, showmanship, stable boy, stagecoach, townspeople, track-down, wagon wheels.

Legend: IMDb = International Movie Database, http://www.imdb.com/; RT = Rotten Tomatoes, http://rottentomatoes.com.

Little Miss Sunshine, the winner in 2006, is about a dysfunctional family that sets out on an interstate road trip to get their youngest member, seven-year-old Olive, to the finals of the eponymous beauty pageant. Several of the multi-morphemic words associated with the least constrained nodes suggest the family's vehicle (*back seat*, *hatchback*, *seat belt*, *side door*, *SUV*, and *under-revving*), the trip that they undertake together including the several stops along the way (*convenience store*, *gas station*, *service station*, *newsstand*, *state trooper*) and the road itself (*US Highway*, *turn-off*). Another large set of statements reference the beauty pageant itself and/or the participants (*back-stage*, *butt-wagging*, *half-dancing*, *make-up*, *outfit*, *over-coiffed*, *over-dressed*, *runners-up*, *runway*, *and upbeat*).

Juno, the 2007 winner, a story about "offbeat" and "an independent-minded teenager confronting an unplanned pregnancy" (Wikipedia, 2013). Among the statements associated with the least constrained nodes are ones concerned with conception (*bedroom*, *bedside*, *make-out*), pregnancy and childbirth (*birthing room*, *birthday*, *knocked-up*, *waiting room*), high school (*backpack*, *blackboard*, *dropouts*), recreational drug-use (*drugstore*, *dime bag*, *burnout*) and disaffected young people (*teenage*, *teenwave*, *runaways*), some of whom are not making the most productive use of their time and ability (*burnout*, *dropouts*, *time suck*).

The winner in 2012 for best original screenplay was *Django Unchained*, Quentin Tarantino's Civil-War-era western about a freed slave's efforts to rescue his wife from a brutal plantation owner. Among the statements associated with the least constrained nodes are several commonly associated with slavery and slaves (*free man*, *boss man*, *overseer*, *bullwhip*, *bare chest*, *bare feet*, *runaway*), as well as plantations and the work performed thereon (*family unit*, *big house*, *farmhouse*, *blacksmith*, *barnyard*, *bunkhouse*, *livestock*, *stableboy*). There are also several statements associated with Django's cooperation with a bounty hunter who is pursuing dangerous fugitives from justice (*courthouse*, *outlaws*, *track-down*, *peace officer*), the means of transportation typical of the times (*wagon wheels*, *stagecoach*, *horseback*, *country road*), and, of course, the violent means to which each resorts to obtain their ends (*bullet hole*, *gunfight*, *gunbelt*, *gunman*, *gun powder*, *hangman*, *lawman*, *marksmanship*, *shotgun*, *showdown*).

The two screenplays whose statements that do not correspond as well to the film's synopses are *The King's Speech* and *Midnight in Paris.* The problem with the latter is a lack of data. Recall this screenplay, as well as *The King's Speech* were outliers regarding the number of statements in the network map, both having 194 where the average was 258. *Midnight in Paris* also has a highly fragments screenplay with only 14 of its 124 (11.2%) concepts being mutually reachable. By comparison, every other screenplay has at least 50% of its concepts mutually reachable. In fact, so fragmented is the text network for *Midnight in Paris* that none of its constraint values are less than 0.25, the threshold set for assessing statements in this study.

As for *The King's Speech*, there are a few statements that evoke the period in which the story takes place, i.e. the period after first World War and build-up to the second. These include *goose-step* (a reference to Nazi soldiers), *wartime*, *WWI* (World War I), and perhaps *naval officer*. There are also a few statements that, when taken together, place the story in Great Britain or some part of the United Kingdom, e.g. *tea time*, *prime minister*, and *pub* which is clipped word for the open compound *public house.* But noticeably absent from the rather limited set of statements is anything to do with speech, language, public speaking, stage fright, or speech impediments and speech therapy. Nor is there anything that suggests that the story takes place in and among members of Britain's Royal Family. That said, if the rules used to include words had allowed for proper nouns, then the several of the following statements might have figured more prominently in the resulting text network—*Whitehall* (both a wide thoroughfare in London and the name for the British Civil Service), *Englishman*, *Great Seal*, *Great War*, *Scotland Yard*. But again, there are no terms among these related to speech and speaking. One conclusion that one might draw from this omission is that the screenplay's author did not have a well-developed mental model of the subject. Specifically, his lack of use of either lexicon or terminology specific to the field and/or less morphologically complex words might suggest a commensurate lack of understanding. Alternatively, the author may have known but consciously decided not to use this language. Interestingly, Wikipedia provides some interesting information about the matter. It states that the author first⋯

> ⋯ read about George VI's life after overcoming a stuttering condition he endured during his youth. He started writing about the relationship between the monarch and his therapist as early as the 1980s, but at the request of the King's widow, Queen Elizabeth The Queen Mother, postponed work until her death in 2002. He later rewrote his screenplay for the stage to focus on the essential relationship between the two protagonists. Nine weeks before filming began, Logue's notebooks were discovered and quotations from them were incorporated into the script.

What all this tells us is that the writing of the screenplay was carried out over a period of more than 20 years. The long period of inactivity was presumably due to unspecified sensitivities on the part of the King's widow. We also learn that what began as a screenplay became a stage play and before becoming a screenplay once more. Finally, we learn that just nine weeks before filming began in the fall of 2009, new material concerning the work of the speech therapist Logue was rapidly incorporated into the script. These conditions are highly irregular and contradict much of the conventional wisdom concerning screenwriting. As such, it is not surprising that the resulting text network appears under-developed and highly fragmented. Nor is it surprising that terminology and jargon on speech therapy is absent and/or poorly integrated with other themes. That said, that screenplay's network is not nearly as fragmented as that of another winner, *Midnight in Paris*, a screenplay whose numerous eccentricities will have to await explanation in another time and place.

5. Discussion

The objective for this paper has been to describe and demonstrate a novel method of network text analysis. The basis for claiming novelty, it has been argued, rests on four pillars. The first of these concerns the unit of analysis, particularly the rationales for including or excluding the words that are to be further analyzed. In this method the only words that are included are multi-morphemic compounds, e.g. abbreviations, acronyms, blends or portmanteau, clipped words, and several varieties of compounds including pseudo-compounds. As was discussed at some length, this is by no means a random assortment. Rather, these categories represent the majority of the processes that morphologists refer to as "word-formation", the creation of new or novel and complex word forms from existing words. To my knowledge, no existing method of NTA uses morphology as a primary or secondary basis for the selection process. And precisely because this criterion is so novel, it is possible to anticipate an important objection: it allows for the inclusion of pronouns and prepositions and many other words typically excluded by almost every other form of text analysis, whether network or frequency-based. This justifyable concern is addressed in the discussion of the three remaining bases for claiming novelty—aggregation, relationship or linkage, and the extraction of meaning.

As noted previously, after selecting words from among the many appearing in a text, the analyst typically assigns or aggregates them to high-order categories. That process may be as simple as stemming words back to their base forms or may involve the use of synonyms or other conceptually-based categories and relationship. The method described in this paper relies upon etymology. Specifically, each element of the multi-morphemic compound is aggregated to its corresponding etymological root, most often an Indo-European (IE) root. For example, for the compound word *firewood* the modifier, *fire*, was assigned to the IE root **paewr** which means "fire" while the head of the compound, *wood*, was assigned to another IE root **widhu** which means "tree, wood". The result of this aggregation process was the creation of a linkage between the two roots—**paewr** and **widhu**—due to the fact that they both co-occur within the same word, *firewood*. As far as I am aware, no existing method of network text analysis aggregates words to their etymological roots and/or creates linkages among concept based on the co-occurrence of the descendants of etymological roots, either within or between words.

The final and arguably most important stage in NTA involves the extraction of meaning from the text network. The most common approach to extraction involves reading some or all of the statements of which the entire network is comprised. Recall that statements are the two concepts or nodes and the relation between them. In the method outlined above, only a subset of statements were used to extract meaning—those associated with the least constrained nodes or concepts in the network. And as was shown in some detail, in five of seven screenplays analyzed this small subset of statements mapped fairly well—in some cases very well—onto the major themes of the narrative. One import of this approach was that while prepositions or pronouns do exist in the network as statements, they do no distort the meaning or adversely affect the extraction process. In short, there is no compelling reason that statements containing prepositions or pronouns even be read, let alone treated as substantive. As such, these words and related statements are retained for both completeness' sake and because of the possibility that they may function as bridges between otherwise disconnected segments of the overall network.

6. Conclusion

In light of the above, two distinct but related avenues of future research are immediately and compellingly evident. The first concerns the abstraction of meaning. The second concerns the question of inter-text analysis. Concerning the former, recall that the extraction of meaning from the textual data involved the reading of a sub-

set of the statements that collectively comprise the network. The selection of that subset was based on one network property—constraint. In particular, the statements used in extraction of meaning were limited only to those associated with nodes having constraint values less than 0.25. Although constraint is widely used in network analytic studies, there is no reason that it should or must be the only one employed at this stage. It is entirely possible that other measures of influence in networks, e.g. centrality, betweenness, closeness, etc., could provide similar or even better results. Future research should seek to determine which measures are better suited, if any, to the extraction of meaning from nodes in text networks like the ones generated here.

As for inter-text analysis, it's important to know that the network approach outlined here can also be used to compare texts of different authors or the different texts by the same author. Such an application is not at all novel. Corman et al. (2002), for example, expressly intended for centering resonance analysis to be used in this way, i.e. to compare similarities and differences in themes present across texts. Carley (1997) compared the mental models of eight project teams, each with 4 - 6 members, enrolled in an information systems project course at a private university. Each team was required to "analyze a client's need and then design and build an information system to meet that need within one semester." Five of these teams were eventually deemed successful and three were not. At three points during the semester, each teams were required to provide responses to two open-ended questions—"What is an information system?" and "What leads to information system success or failure?" Their answers were coded and used as data. Carley reported that, on average, the "cognitive maps" of the members of successful groups had significantly more concepts and more statements compared to maps by members of non-successful groups. Future research should attempt to replicate and extend findings such as these, perhaps for example, using a sample of screenplays written by both experts and amateurs.

Before concluding this section, one broader point concerning network statements in these screenplays is in order. It concerns the matter of lexicon. One definition of that word is *lexicon* which is defined in the *American Heritage Dictionary of the English Language* (AHDEL) as "the morphemes of a language considered as a group." But that is only one of the AHDEL's three definitions of the term. The other two are "a dictionary" and the "stock of terms used in a particular profession, subject, or style; a vocabulary." In light of the analysis above, the third definition is also relevant, and highly so. In his best-selling book on screenwriting entitled *Story*, Robert McKee, argues that not only do films fall into widely-recognized genres, but audiences have "a complex set of expectations" for each one (McKee, 2010: p. 80). These expectations play out at multiple levels—primarily in terms of "subject, setting, role, event, and values". In short, horror films must scare; westerns are expected to have outlaws and lawmen and men on horseback; wars films must depict battles and battlefields, death and dying, bravery or heroism, and grace or cowardice under pressure; films about sports must depict them being played; love stories need lovers and heartbreak, make-up and break-up; epics need quests; legal thrillers need courtrooms, attorneys-at-law, evidentiary rules, prosecutors, judges, and opposing counsel; hospital dramas need doctors and nurses, EMTs and x-rays, vital signs and wheelchairs, and so on. Audiences know whether their expectations are met or not based on what they see on the screen or hear coming out of the mouths of the actors. But before that can happen, screenwriters must, as McKee says, use words and only words to "put a film in the reader's head" (p. 394). That, he says, requires words and language that are "vivid" and "descriptive". And although he doesn't say it directly, a lot of the advice and examples that McKee gives in this matter concern precisely the kinds of words that this method selects and analyzes. Moreover, he emphasizes that the words used must be appropriate to the topic. It's hard to imagine, for example, a movie about basketball that doesn't use the words backboard, front court, half-time, back court, slam-dunk, jump shot, power forward, point guard, etc. Those words comprise some—but not all—of the conceptual vocabulary or lexicon of basketball. What McKee and other writers emphasize is that all film genres and sub-genres have such a lexicon and that to fulfill an audience's genre expectations, able screenwriters are expected to know the genre's lexicon. As an aim to informing screenwriting theory and practice, future research with this method should compare and contrast the lexicon of screenplays within and across film genres, and especially those screenplays that have achieved recognition and awards.

Finally, when compared with existing approaches to *network text analysis* (NTA) the method described in this study is novel, in part, because of its reliance upon *morphology* as a basis for filtering/selecting words. As mentioned previously, a sizable proportion of the remaining words—and arguably the most substantive—is *lexicon*, bothin the *morphemic* sense of the word and the conceptual vocabulary or genre-specific sense, as well. The method is also novel, in part, because of its reliance upon both *morphology* and *etymology* as the bases for aggregating the words into broader conceptual categories and for creating linkages among them. As such, the most

fitting acronym for this method is MENTAL, i.e. ***m**orpho-**e**tymological **n**etwork **t**ext **a**nalysis of **l**exicon.*

References

Allen, W. (2010). Midnight in Paris. http://www.pages.drexel.edu/~ina22/splaylib/Screenplay-Midnight_in_Paris.pdf

Arndt, M. (2006). Little Miss Sunshine. http://www.dailyscript.com/scripts/LITTLE_MISS_SUNSHINE.pdf

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284,* 34-43. http://dx.doi.org/10.1038/scientificamerican0501-34

Boal, M. (2007). The Hurt Locker. http://www.roteirodecinema.com.br/scripts/reader/files/the_hurt_locker_by_mark_boal.pdf

Black, D. L. (2008). Milk. http://www.filminfocus.com/awards08/download.php?filename=milk.pdf

Busey-Mario, B. (aka Cody, D.) (2007). Juno. http://www.scribd.com/doc/2083474/Juno-Final-Script

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for Windows: Software for Social Network Analysis.* Harvard, MA: Analytic Technologies.

Burt, R. (2005). *Brokerage and Closure: An Introduction to Social Capital*. Oxford: Oxford University Press.

Buzan, T. (1974). *Using Both Sides of the Brain*. New York: Dutton.

Carley, K. M., & Palmquist, M. (1992). Extracting, Representing, and Analyzing Mental Models. *Social Forces, 70,* 601-636. http://dx.doi.org/10.1093/sf/70.3.601

Carley, K. M. (1997) Extracting Team Mental Models Through Textual Analysis. *Journal of Organizational Behavior, 18,* 533-538. http://dx.doi.org/10.1002/(SICI)1099-1379(199711)18:1+<533::AID-JOB906>3.3.CO;2-V

Carley, K. M. (2001-13). *Automap 3.0.10*. Pittsburgh, PA: Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research International (ISRI), School of Computer Science, Carnegie Mellon University.

Collins, A., & Loftus, E. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review, 82,* 407-428. http://dx.doi.org/10.1037/0033-295X.82.6.407

Corman, S. R., Kuhn, T., McPhee, R. D., & Dooley, K. J. (2002). Studying Complex Discursive Systems: Centering Resonance Analysis of Communication. *Human Communication Research, 28,* 157-206.

Cross, R., & Cummings, J. N. (2004). Tie and Network Correlates of Individual Performance in Knowledge-Intensive work. *Academy of Management Journal, 47,* 928-937. http://dx.doi.org/10.2307/20159632

Danowski, J. A. (1993). Network Analysis of Message Content. *Progress in Communication Sciences, 12,* 198-221.

Diesner, J., & Carley, K. M. (2005). Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a Novel Method for Network Text Analysis. In V. K. Narayanan, & D. J. Armstrong (Eds.), *Causal Mapping for Research in Information Technology* (pp. 81-108). Harrisburg, PA: Idea Group Publishing.

Diesner, J. (2012). *Uncovering and Managing the Impact of Methodological Choices for the Computational Construction of Socio-Technical Networks from Texts.* Dissertations, Paper 194.

Doerfel, M., & Barnett, G. (1999). A Semantic Network Analysis of the International Communication Association. *Human Communication Research, 25,* 589-603. http://dx.doi.org/10.1111/j.1468-2958.1999.tb00463.x

Dooley, K. J., Corman, S. R., McPhee, R. D., & Kuhn, T. (2003). Modeling High-Resolution Broadband Discourse in Complex Adaptive Systems. *Nonlinear Dynamics, Psychology and Life Sciences, 7,* 61-85. http://dx.doi.org/10.1023/A:1020414109458

Hunter, S. (2011). Pricing Banner Advertisements in a Social Network of Political Weblogs. *Journal of Information Technology Theory and Application, 12,* 5-24.

Hunter, S. (2013). Word-Formation in Mark Boal's *The Hurt Locker*. *Open Journal of Modern Linguistics, 3,* 20-29. http://dx.doi.org/10.4236/ojml.2013.31003

Hunter, S., & Chinta, R. (2013). Structural Holes and Banner-Ad Click-Throughs. *Technology and Investment, 4,* 30-44. http://dx.doi.org/10.4236/ti.2013.41005

Hunter, S., & Smith, S. (2013). Thematic and Lexical Repetition in a Contemporary Screenplay. *Open Journal of Modern Linguistics, 3,* 9-19. http://dx.doi.org/10.4236/ojml.2013.31002

International Movie Database (IMDb) (2013). Django Unchained. http://www.imdb.com/title/tt1853728/

International Movie Database (IMDb) (2013). Juno. http://www.imdb.com/title/tt0467406/

International Movie Database (IMDb) (2013). Little Miss Sunshine. http://www.imdb.com/title/tt0449059/

International Movie Database (IMDb) (2013). The Hurt Locker. http://www.imdb.com/title/tt0887912/

International Movie Database (IMDb) (2013). The King's Speech. http://www.imdb.com/title/tt1504320/

King, G., & Lowe, W. (2003). An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization, 57,* 617-642. http://dx.doi.org/10.1017/S0020818303573064

Kuhn, T., & Corman, S. (2003). The Emergence of Homogeneity and Heterogeneity in Knowledge Structures during a Planned Organizational Change. *Communication Monographs, 70,* 198-229. http://dx.doi.org/10.1080/0363775032000167406

Lieber, R., & Stekauer, P. (2009). Status and Definition of Compounding. In R. Leiber, & P. Stekauer (Eds.), *The Oxford Handbook of Compounding* (pp. 3-18). Oxford: Oxford University Press.

Leydesdorff, L., & Welbers, K. (2011). The Semantic Mapping of Words and Co-Words in Contexts. *Journal of Infometrics, 5,* 469-475. http://dx.doi.org/10.1016/j.joi.2011.01.008

Lubtosky, A. (Ed.) (2006-2013). *Leiden Indo-European Etymological Dictionary Series.* Leiden: Brill.

McKee, R. (2010). *Story: Substance, Structure, Style and the Principles of Screenwriting*. New York: Harper Collins.

McPhee, R., Corman, S., & Dooley, K. (2002). Organizational Knowledge Expression and Management. *Management Communication Quarterly, 16,* 274-281. http://dx.doi.org/10.1177/089331802237241

Novak, J., & Gowin, D. (1984). *Learning How to Learn*. New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139173469

Plag, I. (2003). *Word-Formation in English*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511841323

Pokorny, J. (1959). *Indogermanisches Etymologisches Wörterbuch.* Bern and Munchen: Francke.

Popping, R., & Roberts, C. W. (1997). Network Approaches in Text Analysis. In R. Klar, & O. Opitz (Eds.), *Classification and Knowledge Organization* (pp. 381-898). Berlin, New York: Springer.

Rice, R., & Danowski, J. (1993). Is it Really Just Like a Fancy Answering Machine? Comparing Semantic Networks of Different Types of Voicemail Users. *International Journal of Business Communication, 30,* 369-397. http://dx.doi.org/10.1177/002194369303000401

Roberts, C. W. (1997). A Generic Semantic Grammar for Quantitative Text Analysis: Applications to East and West Berlin Radio News Content from 1979. *Sociological Methodology, 27,* 89-129. http://dx.doi.org/10.1111/1467-9531.271020

Rotten Tomatoes (RT) (2012). Django Unchained. http://www.rottentomatoes.com/m/django_unchained_2012/

Rotten Tomatoes (RT) (2012). Little Miss Sunshine. http://www.rottentomatoes.com/m/little_miss_sunshine/

Rotten Tomatoes (RT) (2012). The King's Speech. http://www.rottentomatoes.com/m/the_kings_speech/

Seidler, D. (2009). The King's Speech. http://www.pages.drexel.edu/~ina22/splaylib/Screenplay-Kings_Speech_The.pdf

Stekauer, P. (2000) *English Word-Formation: A History of Research, 1960-1995*. Tubingen: Gunter NarrVerlag.

Tarantino, Q. (2012). Django Unchained. http://pdfcast.org/pdf/django-unchained-script-by-quentintarantino-de

Trigg, R., & Weiser, M. (1986). TEXTNET: A Network-Based Approach to Text Handling. *ACM Transactions on Information Systems (TOIS), 4,* 1-23. http://dx.doi.org/10.1145/5401.5402

Van Atteveldt, W. (2008). *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content*. Charleston, SC: BookSurge.

Walde, A., & Pokorny, J. (1927-1932). *Vergleichendes Wörterbuch der Indogermanischen Sprachen*. BD, I-III, Berlin: de Gruyter.

Watkins, C. (2011). *The American Heritage Dictionary of Indo-European Roots* (3rd ed.). Boston, MA: Houghtfon Mifflin Harcourt.

Wikipedia (2013). Milk (Film). http://en.wikipedia.org/wiki/Milk_(film)

Wikipedia (2013). The King's Speech. http://en.wikipedia.org/wiki/The_King's_Speech

Wikipedia (2013). The Hurt Locker. http://en.wikipedia.org/wiki/The_Hurt_Locker

Wikipedia (2013). Juno (Film). http://en.wikipedia.org/wiki/Juno_(film)

Wikipedia (2013). Indogermanisches etymologisches Wörterbuch. http://en.wikipedia.org/wiki/Indogermanisches_etymologisches_W%C3%B6rterbuch

Wisniewski, K. (2007). Word Formation. http://www.tlumaczenia-angielski.info/linguistics/word-formation.htm