I Think, Therefore I Screenplay: A Corpus-Based Analysis of Epistemic Stance in Sorkin and Shyamalan Scripts

Jessica Arredondo

Portland State University

Ling 566: Culminating Workshop

Dr. John Hellermann

April 27, 2026

Introduction

Dialogue in screenwriting carries multiple responsibilities, which include shaping character identity, regulating narrative pacing, and modulating audience perception; as a result, it plays a central role within the genre. This study uses a corpus-based approach to examine stylistic differences in the dialogue written by M. Night Shyamalan and Aaron Sorkin. It addresses two research questions and one sub question. RQ1: How does the frequency of epistemic stance markers, such as *think*, *believe*, and *know* in each corpus, reflect differing stylistic priorities in their authorial voice, respectively. RQ2: How do part-of-speech (POS) patterns and grammatical structures differentiate the use of epistemic *think* in the dialogue of Shyamalan's and Sorkin's scripts. SubRQ3: To what extent do conventions of screenplay dialogue, such as story pacing and narrative function of dialogue, influence and interact with these epistemic stance markers.

Through frequency analysis, keyword extraction, and concordance line analysis, this project explores how language use contributes to a script writer's voice and style within written dialogue. The findings of this analysis suggest that while both writers rely on high-frequency verbs, Sorkin favors *think* in argumentative exchanges between characters, and Shyamalan favors *believe* by characters in dramatic, metaphysical scenes. These epistemic stance markers support character

identity and reflect writer-voice.

Literature Review

The present study investigates epistemic stance markers as stylistic features in screenwriting and as a potential vector for authorial voice and character positioning. This aligns with Bergqvist's (2015) definition of perspective as the relationship between speaker and propositional content. Stance, in this view, is not only epistemic but also a construction of identity, positioning characters and authors within social and dialogic contexts. Especially in scripted dialogue, such choices affect how characters express certainty, hedge claims, and align with others. Following Paul Kockelman (2004), stance reflects not only belief in truth and knowledge but also serves as a construct for identity within discourse. Kockelman briefly touches on authorial stance and how an author's beliefs and values are both reflected and compounded within their own texts.

Halliday's (1973, 1978, 1994) systemic-functional linguistics conceptualizes language as a semiotic system organized around ideational, interpersonal, and textual functions. Epistemic stance markers may serve this interpersonal function by shaping how speakers manage social roles and signal alignment or authority (Bergqvist, 2015). Because this analysis examines whether markers like *think* and *know* help construct speaker identity or negotiate relationships, this would make them relevant to authorial dialogue style in screenwriting.

To assist with the context for this analysis, this literature review looks to complementary frameworks in register, stance, and style. Biber (1995) and Biber

and Conrad (2019, p.6) provide a wider framework for analyzing register and genre variation, highlighting how linguistic features are shaped by communicative function. The parameters for epistemic stance markers set forth by Gray and Biber (2012), based on degree of certainty, attribution, and evidentiality, guide the selection of markers for keyword selection and analysis in this project. Gray and Biber's (2012) work on stance in academic discourse offers a useful model for analyzing sentence-level stance expressions, which can be adapted here to examine dialogue in screenwriting. Building on Biber and Finegan's (1989) exploration of stance styles, the analysis focuses on how grammatical and lexical combinations construct speaker positioning. Here, the combination of syntactic structures and epistemic stance markers behave similarly in shaping stance. In this framework, epistemic verbs such as think, know, and believe are selected because they represent high-frequency, prototypical stance markers that encode varying degrees of certainty and evidential commitment, making them suitable for systematic comparison between the two corpora.

To extend this framework into stylistic analysis, the study draws on work from Nils Enkvist. Enkvist's (1964) view of style, in which style emerges from consistent selections within a constrained communicative context, is applied here to the contexts of genre, character, and authorial intent. Enkvist notes that there are specific kinds of text, such as poetry or advertisements, that depart from conventions used by most "literate writing" (Enkvist, 1990, p. 167). By comparing how stance markers are used in Sorkin's and Shyamalan's screenplays, the analysis

follows Enkvist's methodological call for systematic stylistic comparison grounded in linguistic evidence.

While corpus-based and register approaches identify patterns of stance, a literary stylistic perspective is useful to interpret how these patterns contribute to character and narrative meaning. Leech and Short's (2007 [1981]) work on literary style similarly emphasizes how linguistic forms contribute to identity and narrative function. While their focus is fiction, their insights potentially translate to screenwriting, where dialogue is often used for character construction and tone. Their view of stylistic choices as meaningful selections supports the treatment of epistemic markers as part of an author's stylistic representation.

Methodologically, the present analysis follows other corpus-based studies, using corpus analysis software to analyze recurring patterns. Wynne (2006) outlines the strengths of corpus-based approaches in showing stylistic patterns across texts, including the use of annotated corpora to analyze linguistic features beyond the word level. This is relevant here, as identifying patterns in epistemic stance requires examining recurring linguistic features in large bodies of dialogue such as in corpora. In particular, the present study uses POS tagging, qualitative annotation of syntactic and grammatical structures, frequency counts, and concordance line analysis to examine the use of epistemic stance markers.

In their book *Language, Usage, and Cognition* (2010), Joan Bybee explores how the repeated use of linguistic structures can be viewed as an emergent grammar, rather than a priori knowledge. This gives way to language being viewed

as a highly adaptive system, one that supports the argument that style emerges from usage patterns rather than discrete choices. In this context, recurrent use of epistemic verbs such as *think* and *know* can be understood as part of a writer's stylistic repertoire, observable through corpus-based frequency and patterning.

Despite previous work, there remains limited research applying corpus-based methods to screenplay dialogue. This register resides in a sort of liminal space between narrative fiction, entertainment, identity, and communicative function. While prior corpus stylistic studies, such as Wynne (2006) and Biber and Finegan (1989), demonstrate syntactical and lexical analysis for identifying stance, many do not explore specifically how epistemic stance markers such as *think, know,* and *believe,* function within screenplay dialogue and reflect authorial style.

Methods

Research Design and Approach

The study employs a corpus-based, descriptive, mixed-method design grounded in discourse analysis and linguistic feature analysis. The overarching approach is constructivist and interpretive, recognizing that language is co-constructed by social and stylistic conventions, especially in media scripts.

A custom corpus of six screenplays was built, totaling 72,431 tokens. Three scripts were written by M. Night Shyamalan (sub corpus A: 19,046 tokens) and three by Aaron Sorkin (sub corpus B: 53,385 tokens). The scripts used for Shyamalan's corpus were *Signs, Unbreakable,* and *The Sixth Sense;* scripts used for Sorkin's corpus were *Molly's Game, A Few Good Men,* and *Charlie Wilson's War*. All

scripts were sourced from the publicly available screenplay warehouse Script Slug and reflect the pre-filming draft versions, which often differ substantially from the final film. This distinction is crucial because directors and actors frequently alter scripts during production, and this study focuses exclusively on original scriptwriting style, rather than film interpretation, to accurately analyze each writer's style. Therefore, M. Night Shyamalan's dual role as writer and director is noted but not central to the analysis.

Data Preparation

All scripts were downloaded in PDF format and converted to UTF-8 encoded plain text (.txt) files using Notepad. Text files were then opened and cleaned in Notepad, with all non-dialogue content: stage directions, scene headings, action descriptions, and metadata, manually removed using the Find and Replace tool. Only character-spoken lines were retained, with white space preserved only between scene transitions.

M. Night Shyamalan's scripts often insert character action and emotional description directly between dialogue turns. Thus, these lines required manual inspection to distinguish dialogue from narration, since AntConc (Anthony, 2022) treats each line independently. Additionally, older scripts (e.g., The Sixth Sense) were affected by visual degradation, leading to conversion errors that had to be corrected manually to ensure accurate text representation.

For both sub corpora, boilerplate non-character speech, such as news report narration, was excluded to prevent skewing stylistic analysis. In cases where typos

were most likely due to poor optical character recognition quality rather than obvious author intent, corrections were made during the cleaning process.

Data Collection

To examine the use of epistemic stance verbs in the two corpora, frequency searches for epistemic verbs in AntConc were conducted. The verbs *think, know,* and *believe* were selected because they are high-frequency epistemic stance markers that align with the framework established by Gray and Biber (2012), as discussed in the literature review. Additionally, these words are from the list of epistemic stance markers described in Biber et al. (1999) and Brezina (2012). To address RQ3, high-frequency keywords alone can be susceptible to topical or genre-based conventions, whereas epistemic stance markers, although influenced by genre and task type, provide insight into how speaker attitudes are expressed within those constraints.

Searches were conducted using exact word forms e.g. *think, know, believe,* rather than lemmas. This means that conjugated forms such as *thought, thinks,* or *believed* were not systematically included in the initial frequency counts. As a result, the analysis mostly reflects present-tense forms of these verbs, which likely affects the part-of-speech patterns observed, especially the high frequency of present tense verb (VBP). While this approach allowed for a controlled and comparable dataset for both corpora, it may also partially reflect how the search process was structured rather than the full range of verb usage in the scripts.

For each verb, I recorded the total number of occurrences as well as the number used specifically with epistemic meaning. To do this, I determined

epistemic usage through concordance line analysis of the surrounding context to ensure the verb expressed a speaker's belief, knowledge claim, or degree of certainty, rather than non-epistemic usage such as idiomatic or nominal usage e.g. *he's a believer*. Because the corpora differ in size (53,385 tokens for Sorkin and 19,046 for Shyamalan), I calculated normalized frequencies per 10,000 tokens for relative comparisons between the two datasets.

While a lemma-based search including conjugated forms may provide a more comprehensive representation of stance marking in future work, the present analysis focuses on high-frequency base forms to identify consistencies in author-style between the two writers' dialogue.

Data Analysis

AntConc was used to perform Keyword, Frequency, and Keyword in Context (KWIC) concordance analyses. AntTag (Anthony, 2022) was used to tag POS to aid in grammatical patterns across analysis of the two subcorpora. Google Sheets was used to organize lists, compare results between the subcorpora, normalize frequencies, run statistical tests, and produce charts.

Selected high-frequency epistemic stance markers i.e. *think, know, believe* were further examined using AntConc's KWIC Tool. This allowed qualitative inspection of surrounding contexts to determine how each writer used such markers to convey doubt, belief, or assertion. Each concordance line was analyzed and classified as either epistemic or non-epistemic.

Both descriptive and inferential statistics were used to analyze patterns in epistemic stance verb usage. Descriptive statistics, such as raw and normalized frequency counts, were used to summarize how often *think, know,* and *believe,* appeared in each scriptwriter's corpus. To determine whether these differences were statistically meaningful beyond the sample, inferential statistics were applied in the form of log-likelihood tests.

Findings

To compare the use of epistemic stance verbs across the two corpora, the frequency was analyzed in Excel, then normalized; given the difference in corpus size, normalized frequencies per 10,000 words are included to compensate. The keyness of *know, think,* and *believe* was confirmed through AntConc's Keyword Tool, which identified them as salient in both sub corpora. As stated earlier, these verbs are drawn from the list of epistemic stance markers described in Biber et al. (1999) and Brezina (2012).

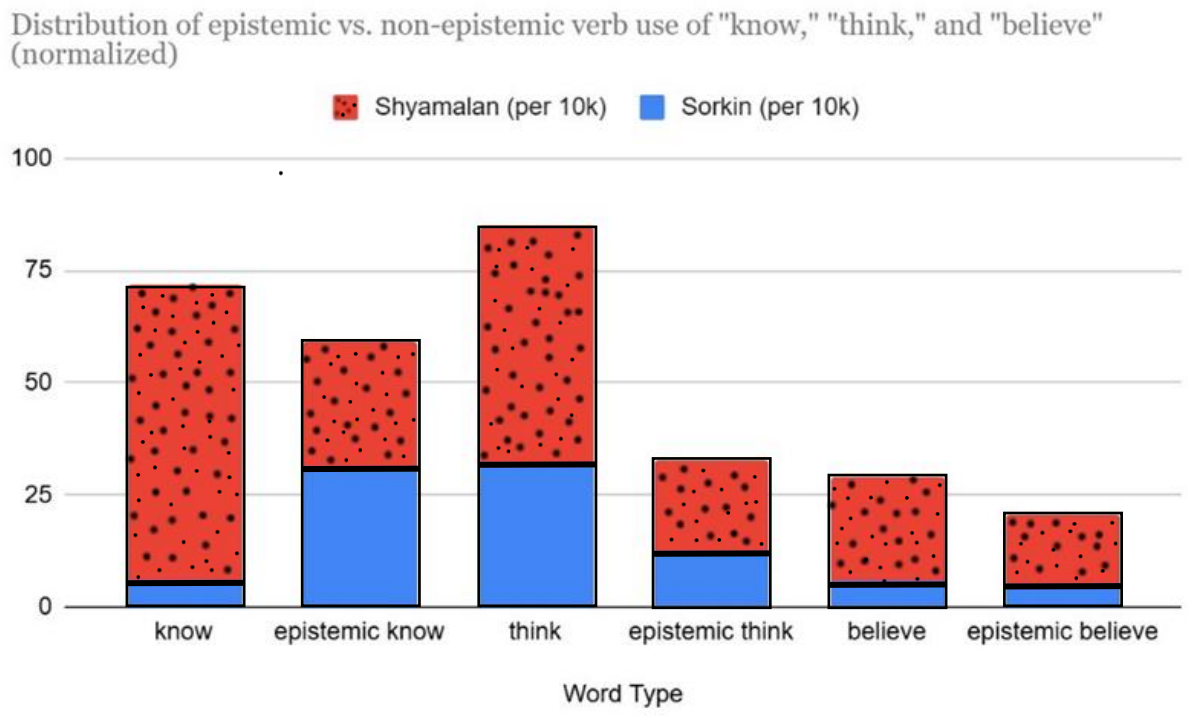
Table 1 presents the total number of hits for each verb in both Sorkin's and Shyamalan's scripts, alongside the subset of those hits used epistemically. As shown in Table 1, Sorkin uses epistemic *know* more frequently (164 vs. 54), however post-normalization reveals frequencies are closer to 30.7 vs. 28.3, which is reflected in Figure 1.

Table 1.

Screenwriter	Verb	Total Hits	Epistemic Use	Corpus Size
A. Sorkin	*know*	301	164	53,385
M. N. Shyamalan	*know*	126	54	19,046

A. Sorkin	*think*	169	66	53,385
M. N. Shyamalan	*think*	101	39	19,046
A. Sorkin	*believe*	30	26	53,385
M. N. Shyamalan	*believe*	45	30	19,046

Figure 1.



While the present study does not aim to generalize findings to all of Shyamalan's or Sorkin's writing, statistical indicators (log-likelihood, normalized frequency) are used to identify features that stand out in each corpus. These results are interpreted qualitatively to explore stylistic tendencies and thematic contrasts between the two writers.

To assess whether observed differences in epistemic verb usage between the two sub corpora were statistically significant; a log-likelihood test was performed on

the normalized frequencies of these verbs. The log-likelihood analysis revealed that epistemic *believe* was significantly more associated with Shyamalan's corpus (LL = 18.67, $p < 0.001$), indicating that expressions of belief are a stronger stylistic feature in his dialogue. The epistemic verb *think* also showed a statistically significant difference (LL = 5.92; p is approx. 0.015), more commonly used by Sorkin. Conversely, epistemic *know* was not statistically significant between the two sub corpora (LL = 0.26, $p = 0.61$).

Table 2.

Log-likelihood test results comparing epistemic uses of *think*, *know*, and *believe*

Verb (epistemic)	Frequency in Sorkin (normalized)	Frequency in Shyamalan (normalized)	Log-Likelihood	p-value
believe	26	30	18.67	< 0.001
think	66	39	5.92	≈ 0.015
know	164	54	0.26	0.61

Epistemic Verbs and POS Frequency

The verbs *think*, *believe*, and *know* appeared with notable frequency in both corpora, but their syntactic and pragmatic uses differed. The relative prominence of *think* and *know* in both corpora points to their centrality as a flexible stance device, yet their usage patterns offer insight into genre-specific rhetorical strategies (Kärkkäinen, 2003).

A focused part-of-speech (POS) analysis was conducted on epistemic *think* in the Shyamalan and Sorkin corpora; figure 2 provides an example from the Shyamalan corpus. In both corpora, the most common syntactic structure was a

present-tense verb form (VBP) followed by a personal pronoun (PRP), typically in constructions such as *I think* + clause. This pattern suggests that epistemic *think* frequently functions as a clause-initial stance marker used to frame subjective evaluation. However, because the analysis was based on base forms rather than lemmas, the prominence of present-tense patterns may be partially influenced by the search structure rather than fully representing all verb forms.

Qualitative analysis of concordance lines further revealed variation between epistemic and non-epistemic uses. Epistemic constructions e.g. *Because I think you faked it*, express speaker belief or uncertainty, while non-epistemic or idiomatic uses e.g. t*hink about what you want*, function as pragmatic routines or directives rather than markers of knowledge or belief. This distinction shows the multifunctionality of *think* and stresses the importance of contextual analysis in identifying stance.

Figure 2.

Example of POS and epistemic verb analysis from M. Night Shyamalan corpus

Syntactic form	POS tags	Function
I think he's sick . I_PRP think_ VBP he_PRP 's_VBZ sick_JJ ._.	PRP + VBP + PRP + VBZ + JJ	Epistemic stance
I don't think so dad_NNP ._. __SP I_PRP do_VBP n't_RB think_ VB so_RB ._	NNP + SP + PRP + NEG + VB + RB	Non-epistemic; Idiomatic Expression; not about knowledge
Think about what you want you_PRP do_VB something_NN for_IN me_PRP ?_. Think_ VB about_IN what_WP you_PRP want_VBP	PRP + VB + NN + + IN + PRP + VB + IN + WP + PRP + VBP	Non-epistemic stance; Idiomatic expression; suggestion

Because I think you faked it P Because_IN I_PRP think_ VBP you_PRP faked_VBD it_PRP ._.	IN + PRP + VBP + PRP + VBD + PRP	Epistemic stance; speaker's subjective belief

Discussion and conclusion

The findings suggest that while both writers rely on high-frequency epistemic verbs such as *think, know,* and *believe,* the quantitative and qualitative patterns of use reflect differing stylistic strategies used by Sorkin and Shyamalan. Quantitatively, normalized frequencies and log-likelihood tests showed that *believe* was significantly more associated with Shyamalan's written dialogue, while *think* occurred more frequently in Sorkin's (LL = 5.92; $p \approx 0.015$). *Know*, however, did not differ significantly between corpora. This suggests that Shyamalan relies more on belief-oriented constructions, potentially indicative of emotional subjectivity and reflective of the supernatural genre their work often resides within. Sorkin, alternately, uses logic, reason, and argumentative posturing through the verb use of *think.*

From a qualitative perspective, Sorkin's scripts use epistemic verbs in more regimented, confrontational contexts. Concordance lines '*I think I'm entitled to it',* and '*what makes you think you know and I don't?',* show a dialogic style grounded in this kind of adversarial exchange, reflecting what Kärkkäinen (2003) describes as epistemic negotiation. Here, *think* functions not as an epistemic marker but also as a tool to assert and challenge. This aligns with Halliday's (1978, 1994) framework of language as socially semiotic, where interpersonal meta functions express power relations and judgments. Sorkin's main characters are assertive, confident,

poignant, and steadfast in opinion. Often, their utterances are lexically dense, which mirrors their knowledge and intelligence. In Sorkin's dialogue, epistemic verbs operate within a register of institutional discourse, such as legal, military, and political debate, reflecting what Biber and Finegan (1989) might attribute to speaker authority.

Shyamalan's usage of epistemic verbs often appears emotional and supernatural in use and context. Shyamalan's main characters are reticent, brooding, and observant. Their utterances are often lexically sparse which reflects Shyamalan's common protagonist archetype. Concordance lines 'I *think God did it',* and '*I believe in signs, Morgan'* reflect, personal belief, otherworldly subjectivity, and metaphysical reflection. Kockelman (2004) speaks of stance as a projection of subjectivity; Shyamalan's stance marker *believe* was not only statistically significant but occurred with lexical items related to faith and general existential mystery. This, perhaps, is reflective of Shyamalan's character development.

The frequency analysis reinforces Bybee and Hopper's (2001) view of grammar and style as emergent from use, rather than discrete structural conventions. The frequent use of specific epistemic verbs by each writer, such as Sorkin's preference for epistemic *think* or Shyamalan's preference for epistemic *believe*, demonstrates how stance can become part of routine usage patterns used within a writer's discourse, forming part of their style through habitual usage. The statistical significance in usage of these verbs shows a preference from these

scriptwriters to use one epistemic verb over the others; whether this pattern reflects genre, authorial voice, or both remains unclear and requires further study.

Syntactic analysis supports distinctions found between the corpora; the most common epistemic construction was *think* followed by a personal pronoun or clause. However, idiomatic and non-epistemic uses were more prevalent in Shyamalan's corpus showing a wide range of use. It would have been both interesting and beneficial to examine scripts written by the same writer across different genres to better isolate whether observed syntactic patterns stem from genre conventions or from the writer's authorial style. However, this type of comparison is constrained by the fact that many screenwriters operate primarily within a single genre, often tied to their creative identity and market niche. This limitation may, in fact, highlight the presence of stylistic consistency at the authorial level, lending further support to Enkvist's (1990) claim that *style* is revealed through patterned deployment of syntactic and pragmatic features (rather than genre). Thus, what initially appears as a methodological constraint may actually reflect the salience of individual writer style as a factor distinct from genre.

Another limitation is that the search process focused on base forms rather than lemmas, so conjugated forms like *thought* or *thinks* were not accounted for. Because of this, the prominence of present tense patterns may be, at least in part, a product of the search structure rather than a complete reflection of how these verbs function in all their forms between the two corpora. Future work could address this by incorporating lemma-based searches. Additionally, future work could expand

this analysis to a broader range of stance markers or incorporate a multimodal analysis, while including a wider variety of scripts could also enhance the generalizability of the findings. Implications for this work include the application of corpus-based methodology in genre, style, and usage-based approaches and designs. Work such as this can serve as groundwork for corpus-based script-analyses, acting as part of a larger review for further tangential study.

Reference

Anthony, L. (2022). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Bergqvist, H. (2015). *Epistemic marking and multiple perspective: An introduction*. *Sprachtypologie und Universalienforschung (STUF), 68*(2), 123–141. https://doi.org/10.1515/stuf-2015-0007

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press.

Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text, 9*(1), 93–124.

Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge University Press.

Enkvist, N. E. (1964). On defining style. In N. E. Enkvist, J. Spencer, & M. Gregory (Eds.), *Linguistics and style* (pp. 1–56). Oxford University Press.

Enkvist, N. E. (1990). Discourse comprehension, text strategies and style. *A.U.M.L.A., 73*, 166–180.

Gablasova, D., Harding, L., Brezina, V., & Dunlea, J. (2024). Expressions of epistemic stance in computer-mediated L2 speaking assessment: A corpus-based approach. *International Journal of Learner Corpus Research, 10*(1), 183–215. https://doi.org/10.1075/ijlcr.00044.gab

Gray, B., & Biber, D. (2012). Current conceptions of stance. In K. Hyland & C. S. Guinda (Eds.), *Stance and voice in written academic genres* (pp. 15–33). Palgrave Macmillan.

Halliday, M. A. K. (1973). *Explorations in the functions of language*. Edward Arnold.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. Edward Arnold.

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). Edward Arnold.

Jucker, A. H., & Ziv, Y. (1998). Discourse markers: Descriptions and theory. In *Discourse markers*. John Benjamins Publishing Company.

Kärkkäinen, E. (2003). *Epistemic stance in English conversation: A description of its interactional functions, with a focus on "I think"*. John Benjamins.

Kockelman, P. (2004). Stance and subjectivity. *Journal of Linguistic Anthropology, 14*(2), 127–150. https://doi.org/10.1525/jlin.2004.14.2.127

Leech, G. N., & Short, M. (2007). *Style in fiction* (Original work published 1981). Longman.

McArthur, T. (1981). *Longman lexicon of contemporary English*. Longman.

Script Slug. (n.d.). *A Few Good Men screenplay*. https://www.scriptslug.com/script/a-few-good-men-1992

Script Slug. (n.d.). *Charlie Wilson's War screenplay*. https://www.scriptslug.com/script/charlie-wilsons-war-2007

Script Slug. (n.d.). *Molly's Game screenplay*. https://www.scriptslug.com/script/mollys-game-2017

Script Slug. (n.d.). *Signs screenplay*. https://www.scriptslug.com/script/signs-2002

Script Slug. (n.d.). *The Sixth Sense screenplay*. https://www.scriptslug.com/script/the-sixth-sense-1999

Script Slug. (n.d.). *Unbreakable screenplay*. https://www.scriptslug.com/script/unbreakable-2000

Wynne, M. (2006). Stylistics: Corpus approaches. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (2nd ed., pp. 223–225). Elsevier.

Xiao, R., & McEnery, T. (2005). Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics, 33*(1), 62–82.