HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

LINGUISTIC FEATURES OF MARK TWAIN'S WORKS A CORPUS-STYLISTIC STUDY OF VERNACULARITY, NARRATIVE VOICE, AND HUMOR

Safarova Durdona Kiyom kizi

Master degree, Asia international university

E-mail: marjonaturakulova79@gmail.com

Abstract. This study examines the linguistic features that make Mark Twain's prose recognizable across fiction, travel writing, and autobiographical narrative. It combines qualitative stylistic interpretation with a reproducible corpus-based analysis of five works: Adventures of Huckleberry Finn, The Adventures of Tom Sawyer, The Innocents Abroad, Roughing It, and Life on the Mississippi. The corpus contains 686,755 word tokens drawn from public-domain Project Gutenberg editions. Rule-based procedures were used to calculate mean sentence length, mean segmental type-token ratio, normalized frequencies of first- and second-person pronouns, contractions, selected vernacular spellings, speech-reporting verbs, question marks, exclamation marks, and dialogue-initial lines. The results show a consistent relationship between genre, narrative perspective, and linguistic texture. The two novels display markedly higher rates of contractions, vernacular markers, interrogatives, exclamatives, and dialogue than the travel and memoir texts. Adventures of Huckleberry Finn has the highest frequency of first-person reference and selected vernacular forms, confirming that its narrative voice is constructed through sustained oralization rather than occasional local color. The Adventures of Tom Sawyer records the highest exclamation rate and the greatest density of dialogue-initial lines, which supports its rapid scene construction and comic theatricality. By contrast, the travel narratives exhibit broader lexical diversity, longer sentences, and a more essayistic relation between narrator and audience. The analysis argues that Twain's style is not a single inventory of dialect spellings but a flexible system in which syntax, pronoun choice, dialogue management, evaluative punctuation, deadpan understatement, hyperbole, and shifts between elevated and colloquial registers work together. These features create character, organize social distance, and transform humor into a form of cultural criticism.

Keywords: Mark Twain; corpus stylistics; literary dialect; vernacular narration; humor; narrative voice; dialogue; American English.

1. INTRODUCTION

Mark Twain's contribution to American literature is inseparable from his treatment of language. His prose appears informal, conversational, and immediately accessible, yet this apparent simplicity is the product of controlled stylistic design. Twain moves between regional speech, standard written English, journalistic observation, mock-ceremonial diction, biblical cadence, technical description, and comic exaggeration. In his major works, language does more than represent place. It establishes social position, exposes pretension, regulates the distance between narrator and reader, and creates an ethical tension between what a character says and what the reader is expected to understand. The result is a prose style in which linguistic form becomes a central means of characterization and criticism.

Scholars have long identified the vernacular as a decisive element in Twain's development. Smith connects the writer's artistic growth with the discovery of an American idiom that could resist inherited literary decorum [10]. Blair likewise treats Huckleberry Finn as the point at which earlier experiments in humor, regionalism, and narrative voice achieve a particularly complex synthesis [12].

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Sewell's study of Twain's languages emphasizes that the writer's achievement cannot be reduced to a simple opposition between 'standard' and 'dialect'; his texts contain several competing discourses, each carrying assumptions about authority, education, race, class, and truth [7]. This multidiscursive quality is visible even within a single scene: a formal phrase may be interrupted by colloquial grammar, a solemn statement may collapse into bathos, or an apparently naive narrator may reveal the contradictions of respectable society.

The linguistic importance of Adventures of Huckleberry Finn is announced in the author's explanatory note, which states that several varieties of speech are differentiated deliberately rather than represented at random [1]. Carkeet's detailed analysis demonstrates both the ambition and the selectivity of this representation, showing that Twain's literary dialect is systematic without being a phonetic transcript of actual speech [6]. The distinction is important. Literary dialect is an artistic convention: it selects features that readers can recognize, distributes them unevenly among characters, and balances social specificity against readability. Consequently, an analysis of Twain's language must consider not only whether a form is historically accurate but also what narrative and ideological work it performs.

The representation of African American speech requires especially careful interpretation. Fishkin argues that African American voices influenced the formation of Twain's vernacular art and the narrating voice associated with Huck [8]. At the same time, modern readers encounter racialized spellings and historically offensive vocabulary that cannot be separated from the power relations of nineteenth-century America. Jones's account of dialect literature shows that nonstandard writing may preserve excluded voices while also exposing them to stereotype and unequal evaluation [9]. A responsible linguistic analysis therefore avoids treating dialect forms as amusing errors. It asks how orthography, grammar, rhythm, and conversational positioning construct voice, how readers are invited to judge that voice, and how the text distributes intelligence and authority among speakers.

Twain's humor is equally linguistic. Cox describes humor as a major organizing force in the writer's career rather than an ornamental addition to serious content [11]. The humor often arises from incongruity between registers: a narrator adopts the vocabulary of science to describe nonsense, applies business logic to moral absurdity, or gives an inflated explanation for an ordinary failure. It also arises from timing. Repetition creates expectation; parallel syntax prepares a conclusion; a short final clause reverses the preceding argument. Punctuation, sentence length, and reported dialogue therefore contribute to comic effect as much as obviously humorous vocabulary. Budd's study of Twain's public personality further indicates that the speaking persona, with its calculated intimacy and apparent spontaneity, was central to how audiences understood the author [13]. The printed voice frequently retains features of oral performance.

Although the linguistic qualities of individual works have been widely discussed, comparison across genres can clarify which features belong to Twain's general repertoire and which are strongly conditioned by narrative form. A first-person vernacular novel, a third-person childhood narrative, a comic travel book, and a river memoir do not require the same distribution of pronouns, dialogue, contractions, or lexical variation. Corpus stylistics offers a useful way to connect local interpretation with patterned evidence. It does not replace close reading; rather, it identifies contrasts that deserve explanation and helps prevent a few famous passages from standing in for an entire work [16-20].

The present study addresses three questions. First, how do measurable features of colloquiality and narrative stance vary across five representative works by Twain? Second, which features most

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

clearly distinguish the two novels from the travel and autobiographical prose? Third, how can these quantitative patterns be interpreted in relation to vernacular voice, characterization, and humor? The working hypothesis is that Twain's linguistic identity is best described as a flexible system of register management. The novels are expected to show denser dialogue, more contractions, more direct address, and more vernacular forms, while the travel and memoir texts are expected to show greater lexical diversity and a stronger essayistic narrator. The study also expects Huckleberry Finn to stand apart because its first-person narration makes colloquial grammar a continuous principle of textual organization.

2. MATERIALS AND METHODS

2.1. Corpus design

The research corpus consists of five complete English-language texts by Mark Twain: Adventures of Huckleberry Finn [1], The Adventures of Tom Sawyer [2], The Innocents Abroad [3], Roughing It [4], and Life on the Mississippi [5]. The selected works represent two broad narrative groups. The first group contains the Mississippi novels, where dialogue and character interaction dominate. The second contains travel and autobiographical prose, where an adult narrator comments on places, institutions, memories, and cultural behavior. Together, the texts provide a cross-genre sample of 686,755 word tokens.

Plain-text UTF-8 editions were obtained from Project Gutenberg. The use of public-domain electronic texts supports transparency and reproducibility, but digital editions require preprocessing because front matter, tables of contents, illustration labels, and licensing notices can distort counts. The Gutenberg headers and footers were removed. The analytical text began with the first chapter of each work, and standalone image-file labels were excluded. Orthographic apostrophes and quotation marks were normalized so that equivalent forms would be counted consistently. Chapter headings remained in the corpus because they are part of the published textual architecture, but preliminary contents pages were omitted. The general principles follow corpus-linguistic recommendations concerning documentation, normalization, and the need to distinguish editorial material from the target text [17-21].

Table 1. Composition of the Mark Twain research corpus

Work	Original publication	Analytical category	Word tokens
Adventures of Huckleberry Finn	1884-1885	First-person novel	111,070
The Adventures of Tom Sawyer	1876	Third-person novel	70,889
The Innocents Abroad	1869	Travel narrative	193,681
Roughing It	1872	Travel/autobiographical narrative	166,201
Life on the Mississippi	1883	Memoir/travel narrative	144,914

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Work	Original publication	Analytical category	Word tokens
Total	**-**	**Five-work corpus**	**686,755**

2.2. Analytical measures

Tokenization was performed with a rule-based expression that identifies alphabetic word forms and preserves internal apostrophes. Sentence boundaries were approximated from terminal punctuation followed by a new orthographic sentence. Because nineteenth-century abbreviations and editorial punctuation complicate automatic segmentation, mean sentence length is interpreted comparatively rather than as an absolute grammatical count. Lexical diversity was measured through mean segmental type-token ratio (MSTTR) in consecutive 1,000-word segments. Unlike a raw type-token ratio, MSTTR reduces the direct effect of unequal text length and is therefore more suitable for comparison among complete works of different sizes.

The analysis calculated normalized frequencies per 1,000 words for first-person pronouns, second-person pronouns, contractions, selected vernacular forms, speech-reporting verbs, question marks, and exclamation marks. The contraction category included forms ending in n't and common apostrophe contractions such as I'm, we'll, they'd, and you've, as well as historically relevant forms such as ain't, hain't, and warn't. Possessive apostrophe-s was excluded because it cannot be reliably separated from verbal 'is' or 'has' without full syntactic parsing.

The vernacular category was intentionally narrow and transparent. It included recurrent items such as ain't, hain't, warn't, gwine/gwyne, dat, dis, dem, wuz, wid, chile, kase, reckon/reckoned, yonder, git, and sivilize/sivilized. This list does not define a dialect and is not used to evaluate correctness. It functions as an index of conspicuous literary-vernacular marking. Some items, especially reckon and yonder, also occur in regional standard usage; consequently, the measure is interpreted as a stylistic tendency rather than a sociolinguistic classification.

Dialogue density was estimated as the percentage of non-empty text lines beginning with a quotation mark. Gutenberg editions do not use quotation conventions uniformly, and some extended speeches continue across lines without a new mark. The measure therefore provides a conservative indicator of speech presentation. Qualitative close reading was then used to interpret the quantitative contrasts. Attention was given to orthographic representation, clause coordination, repetition, register shift, hyperbole, understatement, narrator-reader address, and the relationship between speech form and social evaluation. This combined procedure reflects the central principle of corpus stylistics: numerical patterns become meaningful only when returned to their textual and cultural contexts [16, 19, 20].

2.3. Scope and limitations

The study is descriptive rather than inferential. The five texts are not independent samples from a population in the statistical sense; they are complete works selected to represent contrasting genres. Normalized frequencies reveal textual differences, but they do not by themselves prove authorial intention. The analysis also does not attempt phonological reconstruction. Printed spellings are evidence of literary representation, not direct recordings of nineteenth-century speakers.

A further limitation concerns historical and racial language. Some forms in Huckleberry Finn belong to conventions through which white print culture represented African American speech. They must be examined within the history of enslavement, segregation, minstrelsy, and unequal literacy

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

access. The article therefore reports aggregate patterns without reproducing racial slurs and distinguishes linguistic description from endorsement. The ethical purpose is to clarify how textual authority is constructed and contested, not to naturalize derogatory representation.

3. RESULTS

3.1. General stylistic profile

Table 2 summarizes the principal measures. The first clear result is the difference between the two novels and the three travel or autobiographical texts. The novels average 21.96 contractions per 1,000 words, while the travel and memoir group averages only 2.84. The same contrast appears in selected vernacular markers: 8.11 per 1,000 words in the novels compared with 0.51 in the travel group. Questions, exclamations, speech-reporting verbs, second-person pronouns, and dialogue-initial lines are likewise more frequent in the novels. The pattern indicates that colloquiality in Twain is strongly connected with represented interaction. It is not simply an author-wide preference for informal vocabulary.

Mean sentence length varies less dramatically, but the direction is consistent. The two novels average approximately 22.66 words per sentence under the rule-based segmentation, whereas the three prose narratives average 25.20. The difference is not large enough to support a simplistic claim that Twain's fiction is uniformly 'short-sentenced.' Instead, the novels combine many short speech turns with long coordinated narrative sentences. Travel prose contains fewer turns but more extended exposition, catalogues, comparisons, and retrospective commentary.

Lexical diversity shows the reverse distribution. The MSTTR values are 0.37 for Huckleberry Finn and 0.44 for Tom Sawyer, compared with 0.44-0.46 in the travel and memoir texts. Repetition in the novels is partly functional: character names, reporting verbs, pronouns, conversational formulas, and recurrent action vocabulary are repeated across scenes. The adult travel narrator ranges across geography, architecture, history, transport, religion, and national customs, producing a broader vocabulary within each 1,000-word segment. Lexical diversity here reflects topic range and discourse mode as much as individual stylistic richness.

Table 2. Selected corpus-stylistic indicators

Work	Mean sent. length	MSTTR	1st-person /1k	Contractions /1k	Vernacular /1k	Dialogue lines %	Exclamations /1k
Huckleberry Finn	25.43	0.37	50.24	26.78	12.72	14.77	4.63
Tom Sawyer	19.90	0.44	18.11	17.13	3.50	19.38	9.08
Innocents Abroad	24.41	0.46	29.19	1.18	0.22	2.93	3.76
Roughing It	25.96	0.45	29.93	2.82	0.48	4.69	4.10

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Work	Mean sent. length	MSTTR	1st-person /1k	Contractions /1k	Vernacular /1k	Dialogue lines %	Exclamations /1k
Life on the Mississippi	25.22	0.44	29.06	4.51	0.83	6.30	3.48

Note. Frequencies are normalized per 1,000 words; dialogue lines represent the percentage of non-empty lines beginning with a quotation mark.

3.2. Vernacularity and oralization

Adventures of Huckleberry Finn is the strongest outlier in the corpus. Its 26.78 contractions and 12.72 selected vernacular markers per 1,000 words are substantially higher than the values in every other work. This finding confirms that the novel's oral effect is continuous. Huck's narration does not alternate between a neutral authorial voice and occasional dialect dialogue; the narrator's grammar, vocabulary, and rhythm carry colloquial features through description, reflection, and action. The reader therefore encounters the story from within a socially positioned consciousness.

The 50.24 first-person pronouns per 1,000 words in Huckleberry Finn also exceed the values of the other works. First-person narration alone cannot explain the whole difference, because the three travel narratives are also strongly personal. The higher rate reflects the narrow experiential frame of Huck's account: events are repeatedly organized around what he sees, knows, decides, fears, or fails to understand. This repeated self-reference creates immediacy, but it also establishes dramatic irony. Huck often reports facts accurately while interpreting them through moral assumptions inherited from his community. The linguistic limitation of the narrator thus becomes a method for exposing the larger moral limitation of society.

Tom Sawyer is less heavily vernacularized because the narrative frame is third-person and more mobile. Its 17.13 contractions and 3.50 selected vernacular markers per 1,000 words remain much higher than those of the travel texts, but the features concentrate in dialogue rather than controlling every sentence. The narrator can shift between affectionate irony, mock-heroic description, and children's speech. This layered structure allows the text to portray Tom both from within his imaginative world and from the perspective of an adult observer who recognizes its theatricality.

Figure 1 makes the genre contrast visible. The two novels form a separate cluster, with Huckleberry Finn showing the most sustained combination of contractions and marked vernacular items. Life on the Mississippi occupies an intermediate position. Its river anecdotes and remembered pilot speech introduce colloquial elements, but explanatory and historical chapters reduce their overall density. The Innocents Abroad has the lowest values, despite its reputation for comic informality. Its humor is less dependent on written dialect than on an urbane but deliberately irreverent narrator who applies American skepticism to European cultural authority.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

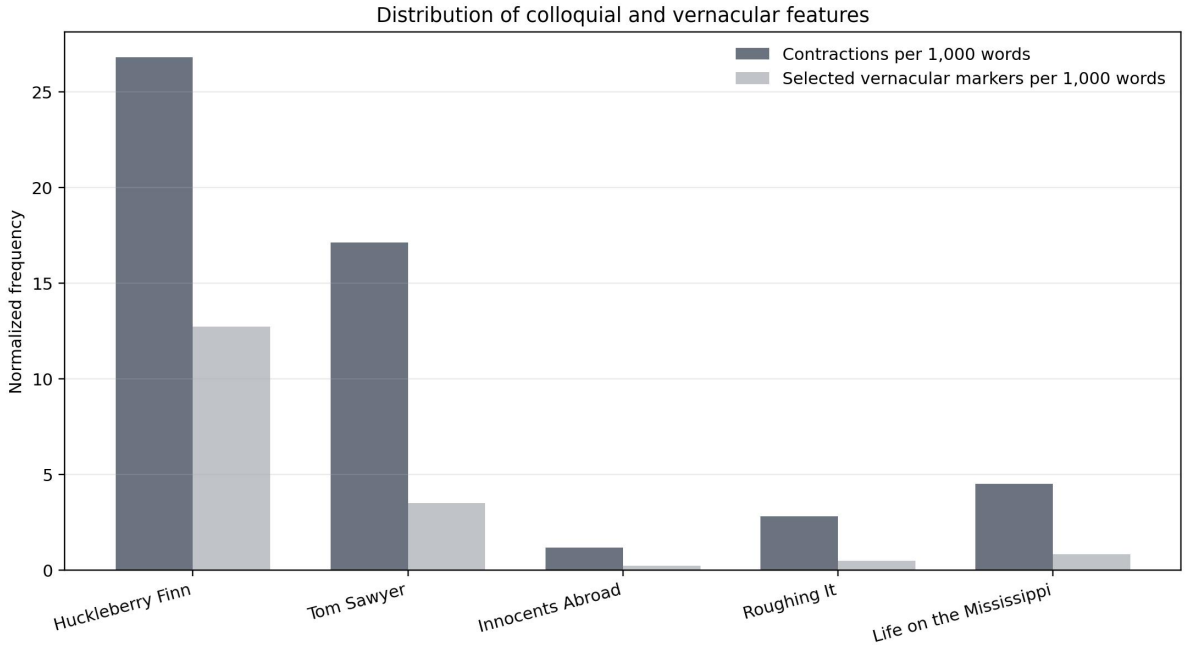


Figure 1. Distribution of contractions and selected vernacular markers across the five works.

3.3. Dialogue, interpersonal orientation, and punctuation

Dialogue-initial lines account for 19.38% of non-empty lines in Tom Sawyer and 14.77% in Huckleberry Finn. The corresponding figures are below 7% in each of the other works. Tom Sawyer's particularly high value reflects its scene-based design. Episodes such as the whitewashing of the fence, classroom performances, games, arguments, and court testimony develop through rapid exchanges. The narrator frequently withdraws and lets speech organize tempo. This produces a dramatic quality: characters reveal desire, pride, fear, and manipulation through the timing of their responses.

Huckleberry Finn contains more speech-reporting verbs than Tom Sawyer (12.33 versus 7.14 per 1,000 words), although Tom Sawyer has a higher proportion of dialogue-initial lines. The difference suggests distinct reporting strategies. Huck repeatedly anchors speech in the sequence of his personal recollection, using simple verbs to mark who spoke and what happened next. Tom Sawyer more often presents exchanges as parts of broader scenes, with narrative description carrying some of the turn organization. In both works, the simplicity of reporting clauses helps foreground the spoken content and preserves the impression of oral storytelling.

Second-person pronouns average 13.06 per 1,000 words in the novels and 5.83 in the travel group. Much of this difference comes from dialogue, but direct narrator-reader address also matters. Twain's narrators often create an implied conversational partner. A phrase such as 'you could see' or 'you may imagine' converts observation into shared experience. In travel prose, this device brings the lecture platform into print: the narrator performs familiarity, anticipates disbelief, and invites the reader to enjoy an evaluative stance. The address is therefore not merely informal; it establishes a temporary community of judgment.

Punctuation reinforces the same contrast. Tom Sawyer has 9.08 exclamation marks per 1,000 words, almost twice the Huckleberry Finn rate and more than twice the rate in any travel work. The result corresponds to the novel's emotional and performative energy. Characters shout, celebrate,

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

threaten, pray, and dramatize ordinary events. Huckleberry Finn, by comparison, often produces humor through restraint. Extreme events are narrated in a relatively level tone, allowing understatement to conflict with danger or absurdity. The travel books occupy a middle position: exclamations occur in comic climaxes, but extended descriptive and reflective passages lower their normalized frequency.

Question marks show a similar pattern: approximately 6.5 per 1,000 words in both novels, compared with an average of 2.0 in the travel and memoir group. Questions are central to child-centered dialogue because characters negotiate knowledge, plans, rules, and status. They also support Twain's comic logic. A speaker asks a reasonable question within an unreasonable system, or asks an absurd question with complete confidence. Interrogative form thereby becomes a device for exposing gaps between institutional language and practical sense.

3.4. Sentence structure and lexical patterning

The mean sentence-length figures require interpretation at the level of syntax. Huckleberry Finn's average of 25.43 words is higher than that of Tom Sawyer and close to the travel narratives, despite its reputation for plain speech. The explanation lies in coordination. Huck frequently links actions and perceptions with and, but, so, and then. The resulting sentences may be long in word count while remaining cognitively transparent. They reproduce the forward motion of oral narrative: one event is added to another without the hierarchical subordination associated with formal expository prose.

Tom Sawyer's lower mean of 19.90 words reflects brisker alternation among dialogue, action, and narrator comment. Short paragraphs and quick changes of speaker contribute to its comic pacing. The travel books, especially Roughing It, more often build elaborate comparisons or accumulate details before delivering a reversal. A long sentence can establish a serious descriptive frame, while its final phrase produces comic deflation. Thus sentence length is not itself a marker of formality. Its stylistic function depends on clause relation, rhythm, and placement.

The higher lexical diversity of The Innocents Abroad and Roughing It is consistent with the encyclopedic range of travel narrative. Twain names unfamiliar objects, institutions, costumes, routes, historical figures, and local practices. Yet this lexical breadth is repeatedly destabilized by colloquial evaluation. Specialized or elevated nouns are followed by blunt adjectives, practical comparisons, or American idioms. The stylistic effect comes from register collision: the narrator temporarily adopts the language of cultural reverence and then measures it against bodily discomfort, expense, boredom, or common sense.

Repetition, by contrast, is a major strength of the novels. Recurrent words create character signatures and comic routines. A character repeats a formula until it becomes predictable; the next variation then carries humor or emotional weight. Repetition also supports the representation of obsession, superstition, and rule-governed play. Tom's language repeatedly converts ordinary life into the vocabulary of adventure, while Huck's repeated practical verbs keep attention on movement, concealment, food, weather, and immediate risk. These patterns show why lexical diversity should not be treated as a direct scale of literary quality. A lower diversity value can indicate deliberate cohesion and voice consistency.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

4. DISCUSSION

4.1. Twain's style as register management

The results support the hypothesis that Twain's linguistic identity lies in register management rather than in a fixed set of dialect spellings. Across the five works, he repeatedly brings unlike forms of discourse into contact. Colloquial narration may frame a learned quotation; technical language may describe an unreliable process; sentimental rhetoric may be interrupted by physical detail; and a socially prestigious speaker may expose ignorance through grammatical completeness. The humor and criticism emerge from the relation among registers.

This interpretation extends Sewell's claim that Twain's texts stage conflicts among languages and discourses [7]. 'Language' in this sense includes more than regional grammar. It includes the speech of schools, churches, courts, newspapers, tourism, commerce, science, and romance literature. Characters borrow these institutional idioms to acquire authority, but the borrowed language often fails when tested against experience. Tom Sawyer imitates literary adventure codes; travelers reproduce conventional admiration; townspeople repeat moral formulas. Twain's narrators reveal the gap between such formulas and observable conduct.

The corpus contrasts show that the same general strategy is realized differently by genre. In the novels, social languages are embodied in characters and negotiated through dialogue. In the travel narratives, they are quoted, paraphrased, or parodied by a mobile narrator. The linguistic surface therefore changes: contractions and dialogue markers decline, while lexical variety and expository syntax increase. Nevertheless, the underlying method remains recognizable. A stable serious register is rarely allowed to remain unchallenged for long.

4.2. Vernacular voice, realism, and social meaning

The exceptional profile of Huckleberry Finn clarifies why its language has had such influence and generated such controversy. Its vernacular features are not detachable decoration. They shape the narrator's categories of perception and the novel's structure of knowledge. Huck's voice appears limited from the perspective of formal education, yet it is often more attentive to concrete evidence than the language of respectable adults. The contrast complicates the equation of standard grammar with intelligence or morality.

Carkeet's analysis remains important because it demonstrates that Twain differentiates speech through selective patterns rather than complete phonetic consistency [6]. Literary realism does not require a transcript. It requires a credible distribution of salient features, rhythm, vocabulary, and interactional behavior. Twain's explanatory claim of careful shading should therefore be read as an artistic program. The novel constructs several voices close enough to be mutually intelligible but distinct enough to indicate locality, race, age, and social experience.

At the same time, the representation is inseparable from nineteenth-century conventions that gave white authors and publishers disproportionate control over how marginalized speech appeared in print. Fishkin's account of African American influence challenges any narrative in which American vernacular literature is imagined as exclusively white [8]. Jones, however, reminds us that dialect writing can both resist and reproduce hierarchy [9]. These perspectives are not mutually exclusive. A form may preserve the expressive force of a voice and still be received through stereotypes created by unequal power.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

For this reason, linguistic analysis should distinguish at least three levels. The first is the historical variety that speakers may have used. The second is Twain's literary selection and spelling of features. The third is the reader's social interpretation of those spellings. Confusing these levels leads either to uncritical celebration of authenticity or to the assumption that every nonstandard form is merely caricature. The corpus data contribute by showing distribution: marked forms are concentrated in particular narrative environments, while the qualitative analysis asks whose knowledge is validated within those environments.

4.3. Linguistic mechanisms of humor

The findings also illuminate several mechanisms of Twain's humor. First, oralization creates timing. Contractions, short speech turns, repeated reporting clauses, and coordination make the reader process events at the pace of a speaker. This pace supports delayed revelation: the narrator may provide several apparently neutral details before a final clause changes their significance. Second, interrogatives create pseudo-reasoning. Characters ask questions that follow local rules but reveal the absurdity of those rules. Third, exclamatives intensify theatrical behavior, particularly in Tom Sawyer, where performance is a central mode of childhood identity.

Fourth, Twain uses disproportion. Hyperbole gives trivial discomfort epic scale, while understatement gives violence or deception an ordinary grammatical form. The two techniques may occur in neighboring paragraphs. The humor depends on the reader's recognition that linguistic magnitude and ethical magnitude do not match. Fifth, repetition creates both predictability and variation. A formula becomes familiar, then a small lexical change produces the comic result. These mechanisms help explain why Twain's humor survives outside individual jokes: it is built into syntax, discourse sequence, and narrative perspective.

The travel books demonstrate that a low rate of written dialect does not mean a weak oral persona. The Innocents Abroad, for example, constructs informality through direct judgment, reader address, comic comparison, and the refusal of obligatory reverence. Its narrator's language is often grammatically standard, but pragmatically rebellious. This distinction is useful for stylistics: colloquiality can be grammatical, lexical, interactional, or attitudinal. Counting contractions captures only one dimension. The broader pattern emerges when quantitative measures are interpreted together.

4.4. Implications for translation and literary education

The results have practical implications for translation. A translator who replaces every marked form with a generalized nonstandard spelling may erase distinctions among characters and produce an unintended comic or stigmatizing effect. Conversely, translating all speech into neutral standard language removes a major part of characterization and narrative ethics. Functional differentiation is therefore more important than mechanical imitation. Translators can distribute colloquial syntax, contractions, discourse particles, regional vocabulary, and levels of formality so that relationships among voices remain perceptible in the target language.

The analysis also supports a multi-level approach in literary education. Students may first identify visible features such as eye dialect, contractions, repetition, and punctuation. They can then examine grammar and discourse: coordination, turn-taking, reporting clauses, direct address, and register shift. Finally, they can discuss social interpretation: who is allowed to speak, whose speech is quoted, which voice controls the narrative, and how standardness influences judgment. This progression prevents dialect from being treated as a list of amusing spellings and connects linguistic form with history and ethics.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Corpus evidence is especially useful in the classroom because it can test impressions. Readers often remember Huckleberry Finn as universally short and simple, yet the mean sentence length is comparable to that of the travel books. The simplicity lies less in sentence size than in coordination, familiar vocabulary, chronological sequence, and a stable narrating perspective. Similarly, The Innocents Abroad can sound conversational despite a very low contraction rate because its narrator creates intimacy through evaluation and implied dialogue with the audience. Quantitative results therefore encourage more precise stylistic claims.

5. CONCLUSION

The corpus-stylistic comparison demonstrates that Mark Twain's linguistic features vary systematically with genre and narrative perspective. The novels contain much higher frequencies of contractions, selected vernacular forms, dialogue-initial lines, questions, exclamations, second-person pronouns, and speech-reporting verbs than the travel and memoir texts. Adventures of Huckleberry Finn is distinguished by the strongest first-person orientation and the most sustained vernacular marking. The Adventures of Tom Sawyer is distinguished by dense dialogue and emphatic punctuation. The travel works display broader lexical diversity, longer expository structures, and a narrator whose conversational force depends more on evaluation and register collision than on nonstandard grammar.

These results confirm that Twain's style cannot be defined by dialect spelling alone. Its essential features include oralized syntax, coordination, repetition, direct address, strategic punctuation, shifts between elevated and colloquial diction, hyperbole, understatement, and the management of competing social languages. Together, these devices make speech a form of action: characters claim authority, hide motives, imitate institutions, and reveal the limits of their knowledge through the way they speak.

The study also underlines the need for historically and ethically informed analysis. Literary dialect is a constructed representation shaped by artistic selection and unequal social conditions. Its interpretation must consider both expressive achievement and the possibility of stereotype. Future research could extend the corpus to Twain's short stories, speeches, letters, and later fiction; compare characters through tagged dialogue; or examine translations in Uzbek, Russian, and other languages. Such work would further clarify how Twain's linguistic art travels across genres, media, and cultures.

DECLARATIONS

Conflict of interest. The author declares no conflict of interest.

Data availability. The primary texts are public-domain Project Gutenberg editions; the normalized values reported in Tables 1-2 were calculated from the five-work corpus described in the Methods section.

Funding. No external funding was received for this study.

REFERENCES

1. Twain, M. Adventures of Huckleberry Finn. Project Gutenberg eBook No. 76. Original work published 1884-1885. https://www.gutenberg.org/ebooks/76
2. Twain, M. The Adventures of Tom Sawyer. Project Gutenberg eBook No. 74. Original work published 1876. https://www.gutenberg.org/ebooks/74

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

3. Twain, M. The Innocents Abroad. Project Gutenberg eBook No. 3176. Original work published 1869. https://www.gutenberg.org/ebooks/3176
4. Twain, M. Roughing It. Project Gutenberg eBook No. 3177. Original work published 1872. https://www.gutenberg.org/ebooks/3177
5. Twain, M. Life on the Mississippi. Project Gutenberg eBook No. 245. Original work published 1883. https://www.gutenberg.org/ebooks/245
6. Carkeet, D. The dialects in Huckleberry Finn. American Literature. 1979;51(3):315-332. https://doi.org/10.2307/2925388
7. Sewell, D. R. Mark Twain's Languages: Discourse, Dialogue, and Linguistic Variety. Berkeley: University of California Press; 1987. 201 p.
8. Fishkin, S. F. Was Huck Black? Mark Twain and African-American Voices. New York: Oxford University Press; 1993. 288 p.
9. Jones, G. Strange Talk: The Politics of Dialect Literature in Gilded Age America. Berkeley: University of California Press; 1999. 255 p.
10. Smith, H. N. Mark Twain: The Development of a Writer. Cambridge, MA: Belknap Press of Harvard University Press; 1962. 212 p.
11. Cox, J. M. Mark Twain: The Fate of Humor. Princeton, NJ: Princeton University Press; 1966. 351 p.
12. Blair, W. Mark Twain & Huck Finn. Berkeley: University of California Press; 1960. 436 p.
13. Budd, L. J. Our Mark Twain: The Making of His Public Personality. Philadelphia: University of Pennsylvania Press; 1983. 266 p.
14. Messent, P. The Cambridge Introduction to Mark Twain. Cambridge: Cambridge University Press; 2007. 140 p.
15. Camfield, G. The Oxford Companion to Mark Twain. New York: Oxford University Press; 2003. 800 p.
16. Leech, G., Short, M. Style in Fiction: A Linguistic Introduction to English Fictional Prose. 2nd ed. Harlow: Pearson Longman; 2007. 404 p.
17. McEnery, T., Hardie, A. Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press; 2012. 292 p.
18. Biber, D., Conrad, S., Reppen, R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press; 1998. 300 p.
19. Mahlberg, M. Corpus Stylistics and Dickens's Fiction. New York: Routledge; 2013. 232 p.
20. Hoover, D. L., Culpeper, J., O'Halloran, K. Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama. New York: Routledge; 2014. 216 p. https://doi.org/10.4324/9780203698914

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

21. Gerlach, M., Font-Clos, F. A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. Entropy. 2020;22(1):126. https://doi.org/10.3390/e22010126